

Estimation de la moyenne et de la variance de l'abondance de populations en écologie à partir d'échantillons de petite taille

Lise Vaudor

▶ **To cite this version:**

Lise Vaudor. Estimation de la moyenne et de la variance de l'abondance de populations en écologie à partir d'échantillons de petite taille. Sciences agricoles. Université Claude Bernard - Lyon I, 2011. Français. NNT : 2011LYO10013 . tel-00842873

HAL Id: tel-00842873

https://theses.hal.science/tel-00842873

Submitted on 9 Jul 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre: 13-2011

Thèse de doctorat de l'Université Lyon 1 - Claude Bernard

Spécialité Biostatistiques

Ecole Doctorale E2M2

ESTIMATION DE LA MOYENNE ET DE LA VARIANCE DE L'ABONDANCE DE POPULATIONS EN ÉCOLOGIE À PARTIR D'ÉCHANTILLONS DE PETITE TAILLE.

Vaudor Lise

Dirigée par: Nicolas Lamouroux

Soutenue le 25 janvier 2011

Jury:

M. René Ecochard	Président du jury
M. Nicolas Lamouroux	Directeur de thèse
M. Pascal Monestiez	Rapporteur
M. Franck Torre	
M. Christian Wolter	Rapporteur

Résumé

En écologie comme dans bien d'autres domaines, les échantillons de données de comptage comprennent souvent de nombreux zéros et quelques abondances fortes. Leur distribution est particulièrement surdispersée et asymétrique. Les méthodes les plus classiques d'inférence sont souvent mal adaptées à ces distributions, à moins de disposer d'échantillons de très grande taille. Il est donc nécessaire de s'interroger sur la validité des méthodes d'inférence, et de quantifier les erreurs d'estimation pour de telles données.

Ce travail de thèse a ainsi été motivé par un jeu de données d'abondance de poissons, correspondant à un échantillonnage ponctuel par pêche électrique. Ce jeu de données comprend plus de 2000 échantillons, dont chacun correspond aux abondances ponctuelles (considérées indépendantes et identiquement distribuées) d'une espèce pour une campagne de pêche donnée. Ces échantillons sont de petite taille (en général, $20 \leq n \leq 50$) et comprennent de nombreux zéros (en tout, 80% de zéros). Les ajustements de plusieurs modèles de distribution classiques pour les données de comptage ont été comparés sur ces échantillons, et la distribution binomiale négative a été sélectionnée.

Nous nous sommes donc intéressés à l'estimation des deux paramètres de cette distribution : le paramètre de moyenne μ, et le paramètre de dispersion, θ. Dans un premier temps, nous avons étudié les problèmes d'estimation de la dispersion. L'erreur d'estimation est d'autant plus importante que le nombre d'individus observés est faible, et l'on peut, pour une population donnée, quantifier le gain en précision résultant de l'exclusion d'échantillons comprenant très peu d'individus. Nous avons ensuite comparé plusieurs méthodes de calcul d'intervalles de confiance pour la moyenne. Les intervalles de confiance basés sur la vraisemblance du modèle binomial négatif sont, de loin, préférables à des méthodes plus classiques comme la méthode de Student. Par ailleurs, ces deux études ont révélé que certains problèmes d'estimation étaient prévisibles, à travers l'observation de statistiques simples des échantillons comme le nombre total d'individus, ou le nombre de comptages non-nuls. En conséquence, nous avons comparé la méthode d'échantillonnage à taille fixe, à une méthode séquentielle, où l'on échantillonne jusqu'à observer un nombre minimum d'individus ou un nombre minimum de comptages non-nuls. Nous avons ainsi montré que l'échantillonnage séquentiel améliore l'estimation du paramètre de dispersion mais induit un biais dans l'estimation de la moyenne ; néanmoins, il représente une amélioration des intervalles de confiance estimés pour la moyenne.

Ainsi, ce travail quantifie les erreurs d'estimation de la moyenne et de la dispersion dans le cas

de données de comptage surdispersées, compare certaines méthodes d'estimations, et aboutit à des recommandations pratiques en termes de méthodes d'échantillonnage et d'estimation.

Mots-clés

Binomiale négative ; Echantillonnage ; Estimation ; Maximum de vraisemblance ; Intervalle de confiance ; Surdispersion

Abstract

Estimating mean and variance of populations abundance in ecology with small-sized samples. In ecology as well as in other scientific areas, count samples often comprise many zeros, and few high abundances. Their distribution is particularly overdispersed, and skewed. The most classical methods of inference are often ill-adapted to these distributions, unless sample size is really large. It is thus necessary to question the validity of inference methods, and to quantify estimation errors for such data.

This work has been motivated by a fish abundance dataset, corresponding to punctual sampling by electrofishing. This dataset comprises more than 2000 samples : each sample corresponds to punctual abundances (considered to be independent and identically distributed) for one species and one fishing campaign. These samples are small-sized (generally, $20 \leq n \leq 50$) and comprise many zeros (overall, 80% of counts are zeros). The fits of various classical distribution models were compared on these samples, and the negative binomial distribution was selected.

Consequently, we dealt with the estimation of the parameters of this distribution : the parameter of mean μ and parameter of dispersion θ. First, we studied estimation problems for the dispersion. The estimation error is higher when few individuals are observed, and the gain in precision for a population, resulting from the exclusion of samples comprising very few individuals, can be quantified. We then compared several methods of interval estimation for the mean. Confidence intervals based on negative binomial likelihood are, by far, preferable to more classical ones such as Student's method. Besides, both studies showed that some estimation problems are predictable through simple statistics such as total number of individuals or number of non-null counts. Accordingly, we compared the fixed sample size sampling method, to a sequential method, where sampling goes on until a minimum number of individuals or positive counts have been observed. We showed that sequential sampling improves the estimation of dispersion but causes the estimation of mean to be biased ; still, it improves the estimation of confidence intervals for the mean.

Hence, this work quantifies errors in the estimation of mean and dispersion in the case of overdispersed count data, compares various estimation methods, and leads to practical recommendations as for sampling and estimation methods.

Keywords

Confidence interval ; Estimation ; Maximum likelihood ; Negative binomial ; Overdispersion ; Sampling

Remerciements

Avez-vous déjà fait l'un de ces rêves où, fuyant un monstre, un spectre, ou une ombre hostile, vous courez remarquablement vite et sautez remarquablement loin (un peu plus vite, et un peu plus loin, de fait, qu'il ne vous est humainement possible de le faire). Ainsi, le méchant, s'il est toujours sur vos talons, ne parvient jamais à vous attraper. Tant que vous continuez à croire à vos capacités, vous êtes non seulement sain et sauf, mais vous êtes de plus en mesure de tirer un certain plaisir de la cavalcade.

Ma thèse est un peu comparable à ce rêve, avec dans le rôle de l'ombre effrayante les doutes, les tâtonnements, les critiques et les tracasseries administratives. Grâce aux gens qui m'entouraient, qui m'ont encouragée à penser que je pouvais courir plus vite et sauter plus loin que je ne l'aurais pensé moi-même (au sens figuré, du moins, car pour ce qui est de mes capacités athlétiques réelles personne ne se berce d'illusions), j'ai pu mener cette thèse dans la sérénité. Je tiens donc à remercier tous ceux qui, en me faisant confiance, m'ont permis de vivre ma thèse comme une expérience passionnante, sans que j'aie, pour autant, à subir les affres de la passion.

En premier lieu, je souhaite adresser à mon directeur de thèse, Nicolas Lamouroux, un très grand merci pour la confiance et la liberté qu'il m'a accordées durant ces trois années de thèse. Tout en me guidant avec intelligence, rigueur et franc-parler (qualités essentielles qui ont fait, font, et feront encore trembler bien des générations de thésards après moi), il a su accepter avec philosophie et humour mon inexorable dérive de l'écologie vers les statistiques, ainsi que mon obsession personnelle pour la binomiale négative et les études de type Monte-Carlo. Si mon travail était issu d'une binomiale négative, alors je pourrais qualifier Nicolas de grand $\hat{\theta}$, tant il m'a accordé à moi, petit $\hat{\mu}$, un grand intervalle de confiance (cf article 3).

Je souhaite également remercier mon comité de thèse, René Ecochard, Bernard Hugueny, Jean-Michel Olivier, et Verena Trenkel pour leurs conseils, suggestions et encouragements. Bien que le nombre de paramètres de la binomiale négative ne me permette pas de l'exprimer métaphoriquement, qu'ils soient assurés de ma reconnaissance pour le regard avisé et bienveillant qu'ils ont posé sur mon travail. En particulier, je souhaite adresser mes remerciements à René Ecochard, qui m'a reçue plusieurs fois malgré un emploi du temps chargé, pour des discussions très enrichissantes et enthousiastes sur le thème des statistiques en général, et des statistiques bayésiennes en particulier.

La base de données d'abondances de poissons qui a motivé ce travail est le fruit d'un effort collectif colossal, sur plus de 20 ans. Merci donc à tous ceux qui ont participé à la collecte de ces données

(plusieurs équipes de l'Université Lyon 1, le Cemagref, le bureau d'études Aralep, et Henri Persat), ainsi qu'aux financeurs de la base de donnée (la Compagnie Nationale du Rhône, l'Agence de l'Eau Rhône-Méditerrannée-Corse, la Région Rhône-Alpes et diverses collectivités locales).

Merci également à Claire Bissery pour le travail qu'elle a effectué sur ce même sujet lors de son stage de M2.

Merci également à Pascal Monestiez et Christian Wolter pour avoir accepté d'être les rapporteurs de ce travail.

Je souhaite bien entendu remercier l'ensemble de mes collègues et amis de l'équipe Dynam, pour leur gentillesse, leur bonne humeur, leur langue bien pendue, et leur brin de folie. Tous ont contribué notamment à égayer mes pauses café/cantine pendant ces 3 années. Je pense notamment à Géraldine, Cynthia, Julien, Roland, Pascal, Jérôme, Virginie, Hervé P., Hervé C., Thibaut, et Raphaël.

Je tiens également à remercier ma famille et mes amis, entre autres Ginoute, les Jul*, Hub', Laetice... J'adresse aussi une pensée particulière à mes amis qui comme moi ont fait l'expérience de la thèse : Robin, Banty, Lucie, Marieke. A tous, merci pour les tribulations et autres moments sympathiques qui ont émaillé ces 3 années (et quand est-ce qu'on remet ça ?). Si j'étais X_i, ils seraient $X_1, X_2, X_3, \ldots, X_n$, un échantillon sans aucun zéro. Enfin, merci à Stan, car si j'étais X_i, nul doute qu'il serait un parfait Y_i.

Organisation du mémoire

Outre les quatre articles rédigés durant cette thèse (figurant en deuxième partie), ce mémoire comprend une synthèse par laquelle j'ai tenté d'expliciter la démarche adoptée pour étudier la précision de l'inférence sur de petits échantillons de données surdispersées.

En introduction, j'y explique rapidement en quoi consiste l'inférence statistique, et j'illustre la nécessité, dans le cas qui nous intéresse, de caractériser les données à travers un modèle de distribution.

Les quatre chapitres suivants correspondent aux quatre articles rédigés durant cette thèse. Ces quatre chapitres portent ainsi sur :

1. la comparaison de plusieurs modèles de distribution sur des données de comptage de poissons, ayant abouti à ce qu'on s'intéresse plus particulièrement à un modèle de distribution binomiale négative (de paramètres de moyenne μ, et de dispersion θ) ;

2. l'étude de l'estimation du paramètre θ, notamment dans des cas très problématiques allant de pair avec les faibles moyennes, fortes dispersions, et petites tailles d'échantillon considérées ;

3. la comparaison de plusieurs méthodes de calcul d'intervalles de confiance pour la moyenne μ ;

4. la comparaison des erreurs d'estimation de μ et θ en fonction de la méthode d'échantillonnage (échantillonnage aléatoire à taille fixe, ou échantillonnage séquentiel basé soit sur le nombre total d'individus observés, soit sur le nombre de comptages non-nuls).

Dans ces chapitres, j'expliciterai les points essentiels des articles, tout en les complétant soit par des exemples illustratifs simples, soit par des remarques sur des problèmes ou méthodes liés.

Enfin, le dernier chapitre revient sur les conclusions et perspectives de ce travail.

Par ailleurs, en annexe, je décris les données de comptage de poissons qui ont, en premier lieu, motivé cette étude. C'est sur ces données que nous nous sommes appuyés pour choisir un modèle de distribution (article 1), et pour déterminer les gammes de moyenne, de dispersion, et de valeurs de tailles d'échantillon sur lesquelles sont basées l'ensemble des études ayant abouti aux articles 2, 3 et 4.

Articles

Article 1 :

Comparing distribution models for small samples of overdispersed count data : the example of freshwater fish.

Vaudor, L., Lamouroux, N. Olivier, J.M.

Article accepté par Acta Oecologica, sous réserve de révisions mineures.

Article 2 :

Are small samples of overdispersed count data informative on the dispersion of abundance ?

Vaudor, L., Lamouroux, N.

Soumis à Environmental and Ecological Statistics.

Article 3 :

Confidence intervals for the mean abundance : which method is best suited to small samples of overdispersed count data ?

Vaudor, L., Ecochard, R.

Soumis à Biometrical Journal.

Article 4 :

Estimation of mean and dispersion with random samples of overdispersed count data : comparison of sampling with fixed size, and sequential sampling.

Vaudor, L., Lamouroux, N.

Notations

Dans ce mémoire de thèse, j'ai tenté, autant que possible, d'uniformiser les notations. J'ai regroupées les principales ici. Elles sont également définies au moment de leur première apparition dans le texte.

X_i : variable aléatoire correspondant à l'abondance au point i

x_i : réalisation de cette variable

y : échantillon de données, $y = (x_1, x_2, \ldots, x_n)$

n : taille d'échantillon

λ : caractéristique de la population

$\hat{\lambda}$: estimation de λ

$\bar{X}$: moyenne arithmétique de $X_1, X_2, \ldots, X_n$, $\bar{X} = \left(\sum_{i=1}^{n} X_i\right)/n$

$\bar{x}$: réalisation de $\bar{X}$ -i.e. moyenne arithmétique observée-

σ : écart-type

s : estimation de l'écart-type, $s = \left(\sum_{i=1}^{n} (X_i - \bar{x})^2\right)/(n-1)$

μ : paramètre de moyenne de la distribution binomiale négative

θ : paramètre de dispersion de la distribution binomiale négative

$\hat{\mu}$: estimation par maximum de vraisemblance de μ

$\hat{\theta}$: estimation par maximum de vraisemblance de θ

$I(A)$: variable indicatrice de l'événement A

T : le nombre total d'individus dans l'échantillon, $T = \sum_{i=1}^{n} X_i$

S : le nombre de comptages non-nuls dans l'échantillon, $S = \sum_{i=1}^{n} I(X_i \neq 0)$

N_{seq} : variable aléatoire correspondant à la taille d'échantillon dans le cas de l'échantillonnage séquentiel

Table des matières

I Synthèse **13**

Introduction **15**

I.0.1 Quelques généralités concernant l'inférence statistique 15

I.0.2 Premier exemple d'inférence non paramétrique : le Théorème Central Limite 17

I.0.3 Deuxième exemple d'inférence non paramétrique : le bootstrap 19

1 Sélection d'un modèle de distribution **23**

I.1.1 Quelques modèles classiques pour les données de comptage 23

I.1.2 Performance et généralité de ces modèles dans le cas des données d'abondance de poissons 24

I.1.3 Distribution de μ et θ, les paramètres du modèle de distribution binomiale négative . . 27

2 Estimation de la dispersion θ **33**

I.2.1 Quelques problèmes liés à l'estimation de la dispersion de la binomiale négative 33

I.2.2 Quelques problèmes d'estimation liés aux caractéristiques des échantillons 35

I.2.3 Etude de la précision en fonction du nombre d'individus : quelques conséquences pratiques 37

3 Estimation par intervalle de la moyenne μ **41**

I.3.1 Une grande diversité de méthodes de calcul 41

I.3.2 Résultats concernant les méthodes détaillées dans l'article 3 44

I.3.3 Résultats concernant les méthodes impliquant une log-transfomation des données . . . 46

4 Echantillonnage à taille fixe ou séquentiel **51**

I.4.1 Echantillonner pour obtenir une précision fixée 51
I.4.2 Echantillonner pour obtenir une meilleure précision 52
I.4.3 Echantillonner pour obtenir une meilleure précision quant à la moyenne 53

5 Conclusion et perspectives **57**
I.5.1 Conclusions : estimation du couple (μ,θ) 57
I.5.2 Perspectives : à la recherche de l'information 62

A Données de comptage de poissons **65**
I.A.1 Echantillonnage 65
I.A.2 Description des échantillons de données 66
I.A.3 Quatre exemples d'échantillons de données 67

II Articles **71**

1 Comparing distribution models for small samples of overdispersed counts of freshwater fish **73**
II.1.1 Introduction 75
II.1.2 Methods 77
II.1.3 Results 83
II.1.4 Discussion 86

2 Are small samples of overdispersed count data informative on the dispersion of abundance? **99**
II.2.1 Introduction 101
II.2.2 Methods 103
II.2.3 Results 106
II.2.4 Discussion 107

3 Confidence intervals for the mean abundance: which method is best suited to small samples of overdispersed count data? **117**
II.3.1 Introduction 119
II.3.2 Methods 122

II.3.3 Results 128
II.3.4 Discussion 130

4 Estimation of mean and dispersion with random samples of overdispersed count data: comparison of sampling with fixed size, and sequential sampling 139
II.4.1 Introduction 141
II.4.2 Methods 143
II.4.3 Results 147
II.4.4 Discussion 149

Première partie

Synthèse

Introduction

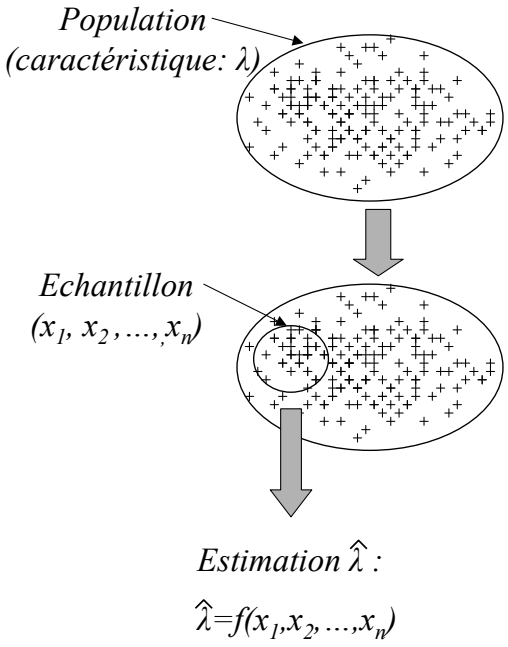
Tout jugement oscille sur la pointe de l'erreur.
Frank Herbert

Quelle information peut-on tirer d'un échantillon de données de comptage d'une population, quand cet échantillon est de petite taille ou ne comprend quasiment que des zéros ? Telle est la question à laquelle nous avons tenté d'apporter une réponse par ce travail de thèse.

Cette question a été en premier lieu motivée par des données d'abondance de poissons d'eau douce collectées dans le bassin du Rhône, présentées plus en détail dans l'annexe A. Les caractéristiques de ces données (moyennes faibles, variances fortes, petites tailles d'échantillons) ont déterminé en grande partie le cadre que nous avons fixé aux études réalisées. Néanmoins, la problématique et la démarche que nous avons adoptées peuvent s'appliquer à toutes les données de comptage dont les caractéristiques seraient similaires, i.e. des données de comptage pour lesquelles l'inférence statistique est particulièrement difficile ou imprécise. Nous avons ainsi choisi, notamment dans les articles rédigés durant cette thèse, de mettre l'accent sur la généralité de ce travail plutôt que sur le contexte hydrologique ou écologique des données qui l'ont motivé.

I.0.1 Quelques généralités concernant l'inférence statistique

L'inférence consiste à étudier et caractériser une population (par exemple une espèce de poissons présente sur un site donné) à partir d'un échantillon de cette population. Le recours à l'échantillonnage pour l'étude d'une population est nécessaire dans tous les cas où le recensement exhaustif de cette population est impossible.



Echantillonnage :

- méthode d'échantillonnage ?
- taille d'échantillon n?

Estimation:

- nature de l'estimateur *f* ?

FIGURE I.0.1 – Principe général de l'inférence et notations

Supposons que l'on s'intéresse à une population, et plus particulièrement à une de ses caractéristiques λ -par exemple la proportion de zéros dans la population, ou la moyenne de l'abondance- (fig. I.0.1). Pour obtenir des informations sur λ, on échantillonne la population, puis on calcule une estimation de λ, $\hat{\lambda}$, basée sur l'échantillon obtenu. Divers facteurs peuvent affecter la précision des informations que l'on tire d'un échantillon de données. Il s'agit, en particulier,

1. de la taille d'échantillon : naturellement, plus la part échantillonnée de la population ou taille d'échantillon *n* est grande, plus l'information fournie par l'échantillon a de chances d'être précise (i.e. plus il y a de chances que $\hat{\lambda}$ soit proche de λ),

2. de la méthode d'échantillonnage (c'est-à-dire de la manière dont on sélectionne la partie de la population examinée),

3. de la nature de l'estimateur de λ (c'est-à-dire selon la méthode utilisée pour calculer $\hat{\lambda}$).

Dans le cas des données de comptage de poissons qui nous intéressent, la taille d'échantillon est très souvent réduite ($n \leq 25$), de sorte que les estimations sont généralement entachées d'une erreur relativement importante. Il est ainsi particulièrement important d'assortir à toute estimation $\hat{\lambda}$, une évaluation de l'incertitude associée. Mais comment évaluer cette incertitude, ou imprécision, alors que par définition on ne connaît pas la valeur réelle de λ ? Dans la plupart des cas, on n'est en mesure d'évaluer la précision de l'estimation que sous conditions : sous l'hypothèse, par exemple, que l'échantillon est suffisamment grand pour que tel ou tel résultat mathématique soit juste, ou bien sous l'hypothèse que les données ont certaines propriétés. En particulier, faire l'hypothèse que les données suivent un certain modèle de distribution fournit des méthodes pour estimer l'imprécision des estimations.

Jusqu'où pourrait-on aller dans l'étude du problème qui nous intéresse ici, sans faire une telle hypothèse ? Autrement dit, dans quelle mesure peut-on faire de l'inférence non paramétrique sur de telles données ? Dans les sections I.0.2 et I.0.3, je teste deux méthodes à la fois simples et courantes d'inférence non paramétrique sur la moyenne. Ces exemples illustrent le fait que pour le type de données qui nous intéresse, l'inférence non paramétrique simple est souvent source d'erreur.

I.0.2 Premier exemple d'inférence non paramétrique : le Théorème Central Limite

Un résultat absolument fondamental en statistiques (Le Cam, 1986b), qui repose sur l'hypothèse selon laquelle « la taille d'échantillon n tend vers l'infini » (c'est-à-dire qu'il s'agit d'une loi « limite » ou « asymptotique »), est le Théorème Central Limite (TCL). Nous y faisons référence plusieurs fois dans ce travail de thèse, et allons nous appliquer ici à expliquer ses implications et limites pour les petits échantillons de données de comptage surdispersés.

Considérons un échantillon de données $y = (x_1, x_2, \ldots, x_n)$. Cet échantillon correspond au résultat d'une expérience aléatoire, c'est à dire à des valeurs observées de n variables aléatoire $(X_1, X_2, \ldots, X_n)$. Considérons également que ces variables $X_1, X_2, \ldots, X_n$ sont indépendantes et identiquement distribuées (i.i.d.), et que leur distribution a pour moyenne μ et pour écart-type σ. Soit $\bar{X}$ la moyenne arithmétique des variables :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Le Théorème Central Limite (TCL) stipule que

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathscr{N}(0,1) \text{ pour } n \to +\infty \tag{1}$$

Ce résultat peut également être formulé comme suit : pour une taille d'échantillon n suffisamment grande, la moyenne arithmétique $\bar{X}$ suit une loi normale $\mathscr{N}(\mu, \sigma/\sqrt{n})$.

Une des applications de ce résultat est, par exemple, le calcul d'un intervalle de confiance pour la moyenne μ pour des échantillons de taille suffisamment grande. En effet, asymptotiquement,

$$pr\left(\mu \in \left[\bar{X} + q_{\mathscr{N},\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + q_{\mathscr{N},1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]\right) = 1-\alpha \tag{2}$$

où $q_{\mathscr{N},\alpha/2}$ et $q_{\mathscr{N},1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi normale $\mathscr{N}(0,1)$ i.e. $-q_{\mathscr{N},\alpha/2} = q_{\mathscr{N},\alpha/2} = 1.96$ pour $\alpha = 0.05$.

En pratique, σ est inconnu, mais peut être estimé par s où

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Plus précisément alors, $(\bar{X}-\mu)/(s/\sqrt{n})$ ne suit pas une loi normale $\mathscr{N}(0,1)$, mais une loi de Student à $(n-1)$ degrés de liberté. Par conséquent, l'intervalle de confiance $(1-\alpha)$, dit de Student, est :

$$CI_S = \left[\bar{X} + q_{\mathscr{S},n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + q_{\mathscr{S},n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{3}$$

où $q_{\mathscr{S},n-1,\alpha/2}$ et $q_{\mathscr{S},n-1,1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi de Student à (n-1) degrés de liberté, $\mathscr{S}$(n-1).

Ce qu'on entend, concrètement, par taille d'échantillon « suffisante » doit être défini. Il s'agit de la taille d'échantillon n à partir de laquelle les distributions de $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ ou de $(\bar{X}-\mu)/(s/\sqrt{n})$ peuvent être considérées comme suffisamment proche des distributions $\mathscr{N}(0,1)$ ou $\mathscr{S}(n-1)$, respectivement. Un point de repère empirique, cité par exemple dans les manuels de statistiques destinés aux étudiants, est que le TCL est approximativement vérifié pour $n \geq 30$ (Boos and Hughes-Oliver, 2000). Néanmoins, nombre de travaux mettent en garde contre cette approximation et appellent à une évaluation plus nuancée de ce qui constitue une taille d'échantillon suffisante (par exemple : Pearson and Adyanthaya, 1929; Gayen, 1949; Wallace, 1958; Johnson, 1978; Boos and Hughes-Oliver, 2000). En effet, comme le souligne Le Cam (1986a), « les théorèmes limites "quand n tend vers l'infini" sont logiquement vides de sens quant à ce qui se produit pour une valeur particulière de n. Au mieux, il ne

peuvent que suggérer certaines approches, dont la performance doit être vérifiée pour le cas que l'on considère en pratique. » (traduction libre)[1]. En effet, si les variables $X_1, X_2, \ldots, X_n$ sont elles-mêmes distribuées selon une loi $\mathcal{N}(0,1)$ alors le résultat (1) est vérifié quelle que soit la taille d'échantillon n. Plus généralement, plus la distribution des variables $X_1, X_2, \ldots, X_n$ est proche d'une gaussienne, et plus la convergence en loi de $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ vers une gaussienne sera rapide (i.e. plus la taille d'échantillon « suffisante » sera faible) (Johnson, 1978; Boos and Hughes-Oliver, 2000).

En d'autres termes, bien que le résultat (1) soit valide asymptotiquement quel que soit la distribution des données, en pratique la taille d'échantillon n est limitée. Par conséquent il convient de s'interroger sur la rapidité de convergence de la distribution de $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ vers une gaussienne, notamment dans le contexte qui nous intéresse, où les tailles d'échantillon sont particulièrement faibles, et où les distributions sont particulièrement asymétriques. Par exemple, Shilane et al (2008) illustre que pour des données de comptage de distribution binomiale négative (i.e. une distribution surdispersée, de moyenne 5 et variance 255), l'approximation selon laquelle la distribution de $\bar{X}$ est gaussienne est mauvaise pour $n = 25$ et même $n = 100$. De la même manière, pour tester la validité de cette approximation dans notre cas, il est nécessaire de caractériser la distribution des variables $X_1, X_2, \ldots, X_n$.

I.0.3 Deuxième exemple d'inférence non paramétrique : le bootstrap

Le bootstrap est un exemple de méthode statistique visant à caractériser la distribution de la variable X sans recourir à un modèle, simplement en se basant sur un échantillon de données $y = (x_1, x_2, \ldots, x_n)$. Soit p la loi de probabilité de la variable X :

$$p(x) = pr(X = x)$$

On s'intéresse à une des caractéristiques de p, que l'on note λ. Soit Λ l'estimateur de λ. On considère que la distribution empirique de $X_1, X_2, \ldots, X_n$ fournit une estimation de p :

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x)$$

[1] « limit theorems "as n tends to infinity" are logically devoid of content about what happens at any particular n. All they can do is suggest certain approaches whose performance must then be checked on the case at hand. »

où $I(X_i = x)$ est la variable indicatrice de l'événement $\{X_i = x\}$. Si l'on suppose qu'on peut faire l'approximation $\{p = \hat{p}\}$, alors on est en mesure de calculer certaines propriétés de Λ.

En effet, le bootstrap consiste à,

(i) rééchantillonner y à de nombreuses reprises (par exemple, 1000 fois) : chaque rééchantillonnage correspond à un tirage au hasard, n fois et avec remise, dans l'échantillon y. Cela équivaut à tirer des échantillons $y'_1, y'_2, \ldots, y'_{1000}$ dans la loi $\hat{p}$

(ii) pour chaque échantillon y' ainsi obtenu, calculer $\hat{\lambda}'$

(iii) déduire de la distribution des $\hat{\lambda}'$ certaines propriétés de l'estimateur Λ (par exemple, l'écart-type de l'estimation, ou un intervalle de confiance pour λ).

Supposons par exemple qu'on s'intéresse à l'échantillon y_4 (de taille $n = 25$), issu de la bases de données poisson (cf Annexe A), et dont la distribution empirique est représentée dans la figure I.0.2.a. On s'intéresse à la distribution de l'estimateur de la moyenne $\bar{X} = \left(\sum_{i=1}^{n} X_i\right)/n$. En appliquant la méthode du bootstrap à l'échantillon y_4, on peut décrire la distribution de l'estimateur de moyenne (cf figure I.0.2.b).

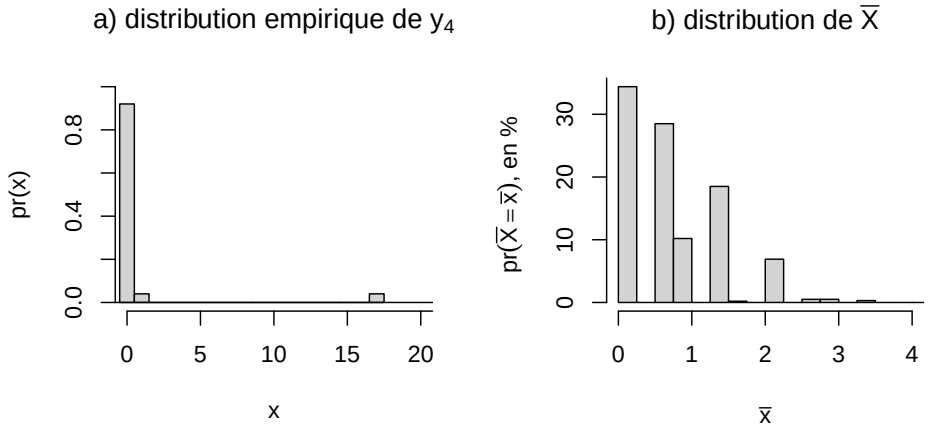


FIGURE I.0.2 – distribution empirique de l'échantillon y_4

On peut alors déduire de la distribution de l'estimateur de moyenne certaines informations quant à la précision de l'estimation obtenue : par exemple, on peut obtenir un intervalle de confiance à 95% pour μ, en considérant les percentiles 2.5 et 97.5 de cette distribution : $CI_{bootstrap} = [0, \approx 2.12]$.

Néanmoins, la méthode du bootstrap n'est valide que sous l'hypothèse $\{p = \hat{p}\}$, c'est-à-dire sous hypothèse que la distribution empirique est assez proche de la distribution réelle. Comme souligné par Efron and Tibshirani (1986) et Efron (1987), les méthodes les plus simples de bootstrap ne sont pas applicables telles quelles dans le cas de distributions asymétriques et de petits échantillons. Dans l'exemple donné en figure I.0.2, cela est particulièrement évident, car il parait peu raisonnable de faire l'hypothèse, en particulier, que la probabilité d'observer 17 individus est de 1/25, tandis que celle d'observer entre 2 et 16 individus, ou plus de 18 individus, est nulle. L'information apportée par l'échantillon est insuffisante, en tant que telle, pour décrire la distribution de la population. Cette information doit être « complétée » par un modèle de distribution.

Les deux exemples d'inférence non-paramétriques que j'ai présenté ici montrent que, s'il existe en théorie des solutions pour faire de l'inférence statistique sans avoir recours à un modèle de distribution, ces solutions sont mal adaptées au type de données auquel nous nous intéressons. Ainsi, pour être en mesure d'estimer la précision de l'inférence statistique pour ce type de données, nous avons choisi de nous intéresser plus particulièrement à l'inférence paramétrique. Nous avons donc tâché, en premier lieu, de choisir un modèle de distribution approprié pour nos données.

Chapitre 1

Sélection d'un modèle de distribution

Pour peindre les portraits,
observez les modèles.
Charles-Guillaume Etienne

I.1.1 Quelques modèles classiques pour les données de comptage

Nous nous sommes intéressés à des modèles assez classiques pour des données de comptage. En premier lieu, il s'agit de la distribution de Poisson, P (table I.1.1), qui suppose que la variance et la moyenne sont égales ($Var(X) = E(X) = \lambda$). La distribution binomiale négative, NB, est également devenue très classique pour les données de comptage surdispersées (i.e telles que $Var(X) > E(X)$) car elle correspond à $E(X) = \mu$ et $Var(X) = \mu + \mu^2\theta$ (Bliss and Fisher, 1953).

Les distributions de Poisson et binomiale négative enflées en zéro, ZIP et ZINB (« ZI » pour « zero-inflated »), sont des modèles relativement récents (Lambert, 1992) qui ont suscité beaucoup d'intérêt, notamment en écologie (e.g. Welsh et al, 1996; Ridout et al, 1998; Gray, 2005; Martin et al, 2005; Wenger and Freeman, 2008), notamment de par l'interprétation qui en est faite. En effet, ces modèles permettent une interprétation plus fine des processus à l'origine des zéros, par exemple en distinguant les zéros dits « stochastiques » (qui correspondent à des situations où la présence d'une espèce sur un site serait possible mais n'est pas effective) des zéros dits « structurels » (qui correspondent à des situations où la présence d'une espèce sur un site est impossible, par exemple du fait des conditions

TABLE I.1.1 – Loi de probabilité et paramétrisation des quatre modèles de distribution considérés

Modèle	Loi	Paramètres
P	$\forall x \in \mathbb{N} \;\; P(X=x) = \frac{\lambda^x}{x!}e^{-\lambda}$	$\lambda \in \mathbb{R}^{*+}$
NB	$\forall x \in \mathbb{N} \;\; P(X=x) = \frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})}$	$\mu \in \mathbb{R}^{*+}$ $\theta \in \mathbb{R}^{*+}$
ZIP	$P(X=0) = (1-\tau)+\tau e^{-\lambda}$ $\forall x \in \mathbb{N}^* \;\; P(X=x) = \tau\frac{\lambda^x}{x!}e^{-\lambda}$	$\tau \in [0,1]$ $\lambda \in \mathbb{R}^{*+}$
ZINB	$P(X=0) = (1-\tau)+\tau(1+\mu\theta)^{-\theta^{-1}}$ $\forall x \in \mathbb{N}^* \;\; P(X=x) = \tau\frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})}$	$\tau \in [0,1]$ $\mu \in \mathbb{R}^{*+}$ $\theta \in \mathbb{R}^{*+}$

environnementales).

La figure I.1.1 illustre les relations entre ces différents modèles : les modèles ZIP et ZINB sont équivalents aux modèles P et NB respectivement si l'inflation en zéro $(1-\tau)$ est égale à zéro. Les modèles NB et ZINB convergent en distribution vers P et ZIP quand le paramètre de dispersion, θ, tend vers zéro.

I.1.2 Performance et généralité de ces modèles dans le cas des données d'abondance de poissons

Dans l'article 1, nous avons ajusté ces quatre modèles par maximum de vraisemblance à chacun des 2258 échantillons de comptage de la base de données poissons (Annexe A). Nous avons ensuite comparé les performances des quatre modèles selon un critère simple (BIC pour « Bayesian Information Criterion »).

D'après ce critère, le modèle NB est le plus approprié pour 58% des échantillons, tandis que les modèles ZIP, P, et ZINB sont sélectionnés dans 24%, 17%, et 1% des cas respectivement. Comme l'illustre la figure I.1.1, ces différents modèles de distributions sont emboîtés deux-à-deux. Dans un contexte où les tailles d'échantillon sont faibles, les modèles les moins parcimonieux peuvent être assez sévèrement défavorisés par un critère tel que le BIC. Néanmoins, nous avons montré dans l'article 1

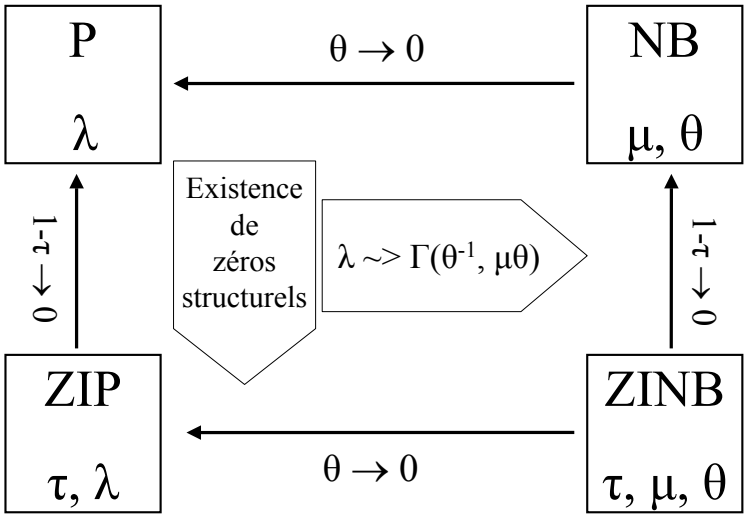


FIGURE I.1.1 – Quatre modèles de distribution et leurs liens

que le modèle ZINB (i.e. celui qui compte le plus de paramètres) est généralement écarté non du fait de son manque de parcimonie, mais parce que les proportions de zéros sont généralement assez bien justifiées par un modèle surdispersé tel que le modèle NB.

Comment, alors, interpréter ces résultats ? Doit-on en conclure que les populations ne présentent effectivement pas de surdispersion (modèle de Poisson) dans 17% des cas, ou que les populations présentent des zéros de plusieurs types (modèles ZIP et ZINB) dans 18% des cas ? Cette interprétation a-t-elle un sens écologique ?

Le hasard d'échantillonnage aboutit naturellement à des échantillons de données présentant des gammes de moyenne et de variance divers. Dans l'article 1, nous avons montré que la sélection d'un modèle de distribution par le critère BIC était très largement influencée -et ce sans surprise- par les moyennes et variances observées. En revanche la sélection d'un modèle, une fois que l'influence de la moyenne, notamment, est prise en compte, dépend très peu des espèces ou sites considérés. Il semble donc peu raisonnable d'interpréter la sélection d'un modèle de distribution en termes écologiques. Plus vraisemblablement, les gammes de moyenne et variance, et le hasard d'échantillonnage, aboutissent à des échantillons pour lesquels la qualité de l'ajustement des différents modèles est variable.

La simulation suivante corrobore cette hypothèse. Considérons en effet chacun des 2258 échantillons de données :

(i) A un échantillon y_j (de taille n_j), correspondent les estimations $\hat{\mu}_j$ et $\hat{\theta}_j$ des paramètres de la NB.

(ii) Simulons dans la NB($\hat{\mu}_j, \hat{\theta}_j$) un échantillon y'_j (de taille n_j). Comme pour l'étude correspondant à l'article 1, on ne conserve que les échantillons y'_j comprenant au moins 3 individus.

(iii) Ajustons à cet échantillon les modèles de distribution P, NB, ZIP et ZINB, et sélectionnons un modèle de distribution d'après le critère BIC.

TABLE I.1.2 – Sélection des différents modèles de distribution pour des échantillons simulés selon chacun des modèles.

		Sélection du modèle			
		P	NB	ZIP	ZINB
échant. de données poissons		17	58	24	1
échant. simulés selon modèle :	P	92	2	5	1
	NB	21	51	26	2
	ZIP	17	11	72	0.4
	ZINB	11	47	40	1

Les résultats de cette simulation figurent dans la table I.1.2. Ils montrent que si toutes les populations (toutes espèces, toutes campagnes confondues) étaient effectivement distribuées selon une NB, et avec des valeurs de paramètres égales aux estimations $(\hat{\mu}, \hat{\theta})$, on observerait néanmoins une importante variabilité dans le choix du modèle. De plus, la sélection des différents modèles se ferait dans des proportions (21%,51%,26% et 2%) proches de celles que l'on observe (17%, 58%, 24% et 1%). Ce n'est pas le cas si l'on fait l'hypothèse que toutes les populations sont distribuées selon un modèle P, ZIP ou ZINB.

Nous considérons donc, pour les études suivantes, que le modèle NB est approprié pour modéliser la distribution des données de comptage de poissons, et ce quelle que soit l'espèce ou le site considéré. Les études réalisées par la suite sont donc basées sur le modèle de distribution NB, avec la paramétrisation

suivante :

$$\forall x \in \mathbb{N}\ P(X = x) = \frac{\Gamma(x + \theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1 + \mu\theta)^{-(x+\theta^{-1})} \quad (\mu, \theta) \in \mathbb{R}^{*+} \times \mathbb{R}^{*+}$$

où μ est le paramètre de moyenne et θ le paramètre de dispersion.

I.1.3 Distribution de μ et θ, les paramètres du modèle de distribution binomiale négative

Si le modèle binomial négatif suffit à décrire l'ensemble des populations étudiées, il existe néanmoins une certaine variabilité entre les populations étudiées (différentes espèces, différents sites, etc.) qui s'exprime par la variabilité des valeurs de moyenne μ et de dispersion θ des populations.

Pour nos données poissons, à quelles gammes de valeurs de μ et θ a-t-on affaire ? Les distributions observées de $\hat{\mu}$ et $\hat{\theta}$ (voir fig. I.1.2) reflètent évidemment les distributions de μ et θ. Néanmoins, du fait des erreurs et biais d'estimation (notamment dans le cas du paramètre θ qui tend à être sous-estimé), il est vraisemblable que les distributions observées de $\hat{\mu}$ et $\hat{\theta}$ soient un peu « déformées » par rapport aux distributions réelles de μ et θ.

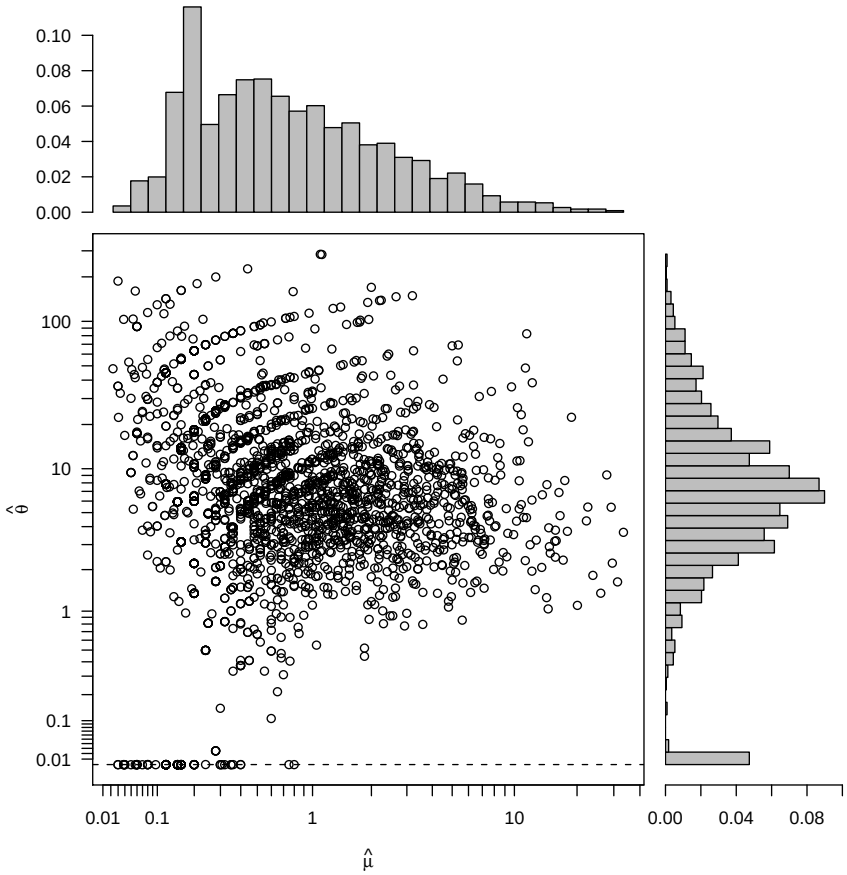


FIGURE I.1.2 – Distribution des estimations $(\hat{\mu},\hat{\theta})$ pour les 2258 échantillons de la base de données poissons. Les histogrammes correspondant aux distributions marginales de $\hat{\mu}$ et $\hat{\theta}$ figurent respectivement en haut et à droite.

En effet, supposons que les distributions de μ et θ sont identiques aux distributions empiriques de $\hat{\mu}$ et $\hat{\theta}$. Simulons alors des échantillons dans les lois NB($\mu = \hat{\mu}$,$\theta = \hat{\theta}$). Ajustons ensuite un modèle NB à chacun des échantillons obtenu, de sorte que nous obtenons la distribution de $\hat{\mu}'$ et $\hat{\theta}'$ (cette simulation correspond en fait à celle déjà réalisée dans la partie I.1.2, p. 26).

La figure I.1.3 révèle que la distribution de $\hat{\theta}'$ est assez différente de la distribution de θ. Autrement dit, si la distribution de (μ,θ) était identique à la distribution observée de $(\hat{\mu},\hat{\theta})$, alors la distribution de $\hat{\theta}$ serait différente de celle que l'on observe. Ce raisonnement par l'absurde montre que l'on ne peut

pas supposer que la distribution θ est exactement semblable à la distribution de $\hat{\theta}$.

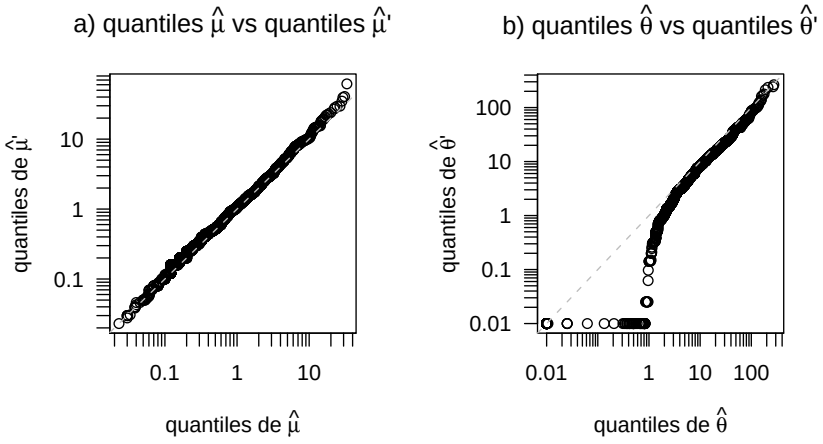


FIGURE I.1.3 – Quantile-quantile plots de a) $\hat{\mu}$ (distribution empirique) vs $\hat{\mu}'$ (distribution simulée) et b) $\hat{\theta}$ (distribution empirique) vs $\hat{\theta}'$ (distribution simulée). Les valeurs $\hat{\mu}'$ et $\hat{\theta}'$ sont simulées à partir de l'hypothèse selon laquelle μ et θ sont distribués identiquement à $\hat{\mu}$ et $\hat{\theta}$.

Les observations $\hat{\mu}$ et $\hat{\theta}$ suivent approximativement une distribution log-normale :

$$\begin{aligned} log(\hat{\mu}) &\sim \mathcal{N}(-0.43, 1.27) \\ log(\hat{\theta}) &\sim \mathcal{N}(1.59, 2.20) \end{aligned}$$

On teste un modèle de distribution de (μ,θ) tel que

1. les distributions de μ et θ sont indépendantes.

2. la distribution de μ est relativement proche de la distribution de $\hat{\mu}$

3. la distribution de θ est légèrement modifiée par rapport à la distribution de $\hat{\theta}$: l'espérance de θ est plus élevée que celle de $\hat{\theta}$ (pour compenser le fait que $E(\Theta) < \theta$), et la variance de θ est moins importante que celle de $\hat{\theta}$ (pour compenser la forte variabilité de l'estimateur)

Par « tâtonnements successifs » (l'ajustement précis du modèle requérant des calculs particulière-

ment intensifs) on retient le modèle suivant :

$$\begin{aligned} log(\mu) &\sim \mathcal{N}(-0.4, 1.2) \\ log(\theta) &\sim \mathcal{N}(2, 1) \end{aligned}$$

qui correspond à des distributions $\hat{\mu}$ et $\hat{\theta}$ assez proches de celles observées (cf fig. I.1.4). Cette distribution de (μ, θ) est représentée dans la figure I.1.5.

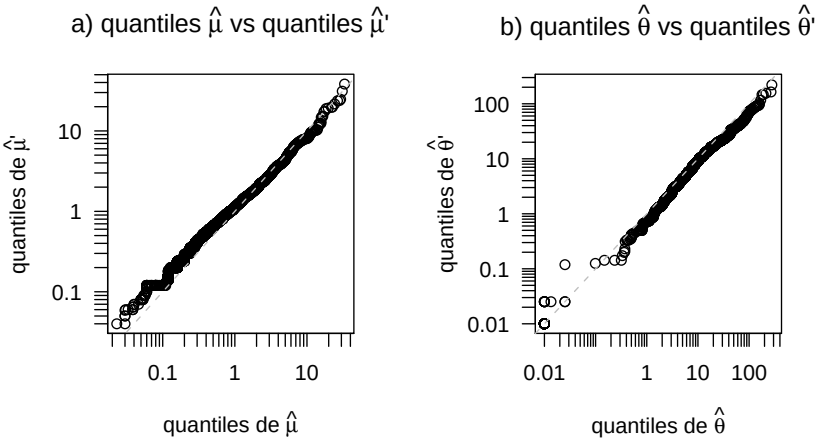


FIGURE I.1.4 – Quantile-quantile plots de a) $\hat{\mu}$ (distribution empirique) vs $\hat{\mu}'$ (distribution simulée) et b) $\hat{\theta}$ (distribution empirique) vs $\hat{\theta}'$ (distribution simulée). Les valeurs $\hat{\mu}'$ et $\hat{\theta}'$ sont simulées à partir de l'hypothèse selon laquelle μ et θ sont indépendants et distribués identiquement à $\hat{\mu}$ et $\hat{\theta}$.

On utilisera ce modèle de distribution afin d'aller plus loin dans les interprétations des résultats correspondant aux articles, dans la section I.2.3 (p.37), dans l'article 4 (p.139), et dans la section I.4.3 (p.53), et dans la section I.5.2 (p.62).

Remarquons que l'on fait l'hypothèse que μ et θ sont indépendants, ce qui peut sembler contradictoire avec la corrélation observée entre $\hat{\mu}$ et $\hat{\theta}$ (et, plus particulièrement, entre la moyenne et la variance observées). La relation entre moyenne et variance, en effet, est bien connue des écologistes, et a été l'objet de beaucoup d'attention, par exemple à travers l'étude la loi de Taylor qui correspond à une corrélation linéaire entre le logarithme de la moyenne et le logarithme de la variance (Taylor, 1984). Néanmoins, cette corrélation est forte bien que μ et θ soient supposés indépendants, tout sim-

plement parce que le modèle de distribution binomiale négative correspond à une variance telle que $Var(X) = E(X) + \theta * E(X)^2$. Pour la simulation décrite ci-dessus, le coefficient de corrélation entre log(moyenne observée) et log(variance observée) est ainsi de 0.94.

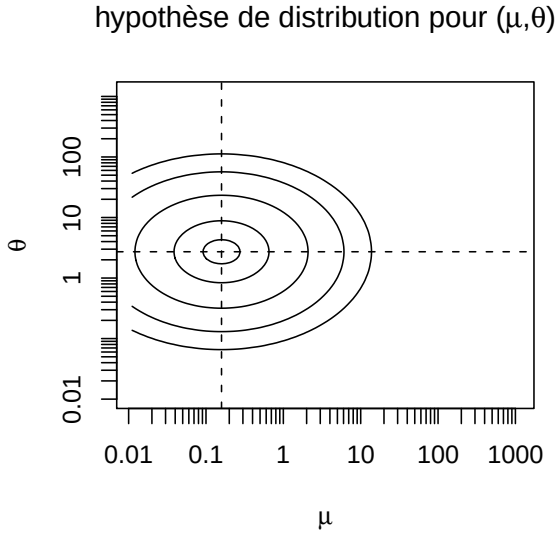


FIGURE I.1.5 – Hypothèse de distribution pour (μ, θ). Les contours indiquent les niveaux de probabilité égaux à 0.001, 0.01, 0.1, 0.5 et 0.9 fois la probabilité maximale

Chapitre 2

Estimation de la dispersion θ

La valeur d'un hasard est égale
à son degré d'improbabilité.
Milan Kundera

Sous l'hypothèse que la population étudiée suit une NB(μ,θ), il suffit d'estimer μ et θ pour décrire la distribution de la population dans son ensemble. Nous nous sommes tout d'abord intéressés à l'estimation du paramètre de dispersion θ.

I.2.1 Quelques problèmes liés à l'estimation de la dispersion de la binomiale négative

L'estimation du paramètre de dispersion θ de la NB a fait l'objet d'une littérature extensive, car elle présente plusieurs problèmes (voir par exemple Pieters et al, 1977; Willson et al, 1984; Clark and Perry, 1989; Anraku and Yanagimoto, 1990; Piegorsch, 1990; Al-Saleh and Al-Batainah, 2003; Saha and Paul, 2005; Lloyd-Smith, 2007; Lord and Miranda-Moreno, 2008).

(a) Tout d'abord, il n'est pas toujours possible de calculer une estimation de ce paramètre. C'est notamment le cas quand les échantillons sont sous-dispersés (i.e. quand la variance observée s^2 est inférieure à la moyenne observée $\bar{x}$). Dans ce cas, l'estimateur de θ par la méthode des moments par exemple, i.e. $\left(s^2-\bar{x}\right)/\bar{x}^2$, renvoie une estimation négative alors que par définition $\theta \in]0,+\infty]$.

(b) Ensuite, la variabilité des estimations $\hat{\theta}$ est particulièrement forte. La RMSE (pour Root Mean Square Error), $RMSE = \sqrt{E\left((\hat{\theta}-\theta)^2\right)}$, est importante.

(c) Enfin, les estimateurs Θ de la dispersion sont biaisés, c'est-à-dire $E(\Theta) \neq \theta$ (Wang, 1996).

Il existe de nombreuses manières possibles de définir un estimateur de la dispersion (les plus classiques étant certainement l'estimateur par la méthode des moments, et l'estimateur par maximum de vraisemblance). Certaines études s'emploient à proposer et à comparer de nombreux estimateurs notamment par des méthodes de type Monte-Carlo (par exemple, Pieters et al, 1977; Clark and Perry, 1989; Saha and Paul, 2005; Robinson and Smyth, 2008).

Nous avons constaté que, dans notre cas, les problèmes d'estimation étaient la plupart du temps associés à des échantillons d'un certain type. Par conséquent, plutôt que de nous consacrer à la recherche d'estimateurs de la dispersion meilleurs ou nouveaux, nous avons tenté de comprendre la façon dont les propriétés des échantillons influencent la qualité de l'estimation.

Nous nous sommes intéressés, en particulier, à l'estimateur par maximum de vraisemblance du paramètre de dispersion. Ce choix repose d'une part sur le fait qu'il s'agit d'un estimateur relativement classique, et d'autre part sur son utilisation par certaines méthodes de calcul d'intervalles de confiance pour μ (en particulier la méthode du profil de vraisemblance à laquelle nous nous intéressons dans l'article 3).

La figure I.2.1 illustre quelques problèmes d'estimation des paramètres de la binomiale négative, et en particulier de la dispersion, θ. Cette figure correspond à

(i) la simulation de 10 000 échantillons de taille $n = 20$ dans une loi NB($\mu = 0.25$,$\theta = 20$)

(ii) le calcul de $\hat{\mu}$ et $\hat{\theta}$ (estimations de μ et θ par maximum de vraisemblance) pour chacun de ces 10 000 échantillons. Dans le cas où l'estimation de θ tendait vers 0, nous avons fixé $\hat{\theta} = 0.0001$.

Cette figure illustre la variabilité des estimations $\hat{\mu}$ et $\hat{\theta}$, dans un cas particulièrement problématique (la moyenne μ est faible, la dispersion θ est forte, la taille d'échantillon n est petite). On observe effectivement que :

(a) Dans plus de 20% des cas, les échantillons sont sous-dispersés (alors que la population est largement surdispersée : $Var(X) = 1.5$ et $E(X) = 0.25$).

(b) Les valeurs de $\hat{\mu}$ et de $\hat{\theta}$ sont très variables. Ainsi, $\hat{\mu}$ varie entre 0.05 et 3.05, et $\hat{\theta}$ varie entre ≈ 0 et 98. De plus $RMSE(\Theta) \approx 17.6$.

(c) On observe un biais dans l'estimation de θ ; en effet $E(\Theta) \approx 15.1$ alors que $\theta = 20$.

I.2.2 Quelques problèmes d'estimation liés aux caractéristiques des échantillons

Là où de nombreuses études s'attachent à proposer et comparer divers estimateurs du paramètre θ, nous nous sommes attachés à décrire la façon dont certaines propriétés des échantillons influencent la qualité des estimations. En effet, le fait que certains échantillons soient peu informatifs quant aux caractéristiques des populations est relativement intuitif. C'est le cas notamment des échantillons qui comprennent peu d'individus ou peu de comptages non-nuls.

La figure I.2.1 montre ainsi l'effet de deux statistiques des échantillons sur l'estimation de θ. Il s'agit du nombre de comptages non-nuls $S = \sum_{i=1}^{n} I(X_i \neq 0)$, (où I est la variable indicative de l'événement), et du nombre total d'individus $T = \sum_{i=1}^{n} X_i$. Les faibles valeurs de S correspondent généralement à des valeurs hautes de $\hat{\theta}$, tandis que les faibles valeurs de T correspondent généralement à des valeurs basses de $\hat{\theta}$. Dans la mesure où le biais d'estimation de θ correspond à une sous-estimation ($E(\Theta) < \theta$), et où l'occurence des valeurs faibles de T est beaucoup plus fréquente que celle des valeurs faibles de S (cf histogrammes en marge de la figure I.2.1), nous nous sommes intéressés en particulier à l'effet de T sur la précision d'estimation de θ.

En pratique, un chercheur confronté à des échantillons comprenant très peu d'individus, considérant que de tels échantillons ne sont pas à même de lui fournir des estimations suffisamment fiables, peut décider de les exclure purement et simplement des analyses. Bien que relevant du bon sens, cette démarche pose question à différents égards.

Tout d'abord, il est difficile de fixer un seuil pour T en dessous duquel on exclut les échantillons. En particulier, est-il raisonnable d'estimer les paramètres des populations sur les échantillons comprenant au moins 5, au moins 10, ou au moins 20 individus ?

Par ailleurs, l'exclusion d'échantillons représente un « gaspillage » (inévitable si la taille d'échantillon est fixée) de l'effort d'échantillonnage, qui peut être particulièrement important dans certains contextes. Considérons par exemple la base de données poissons. L'échantillonnage est plurispécifique, c'est à dire qu'à une campagne (et une taille d'échantillon fixée) correspondent des échantillons de diverses espèces, plus ou moins abondantes. Cela suppose que si la taille d'échantillon est fixée de ma-

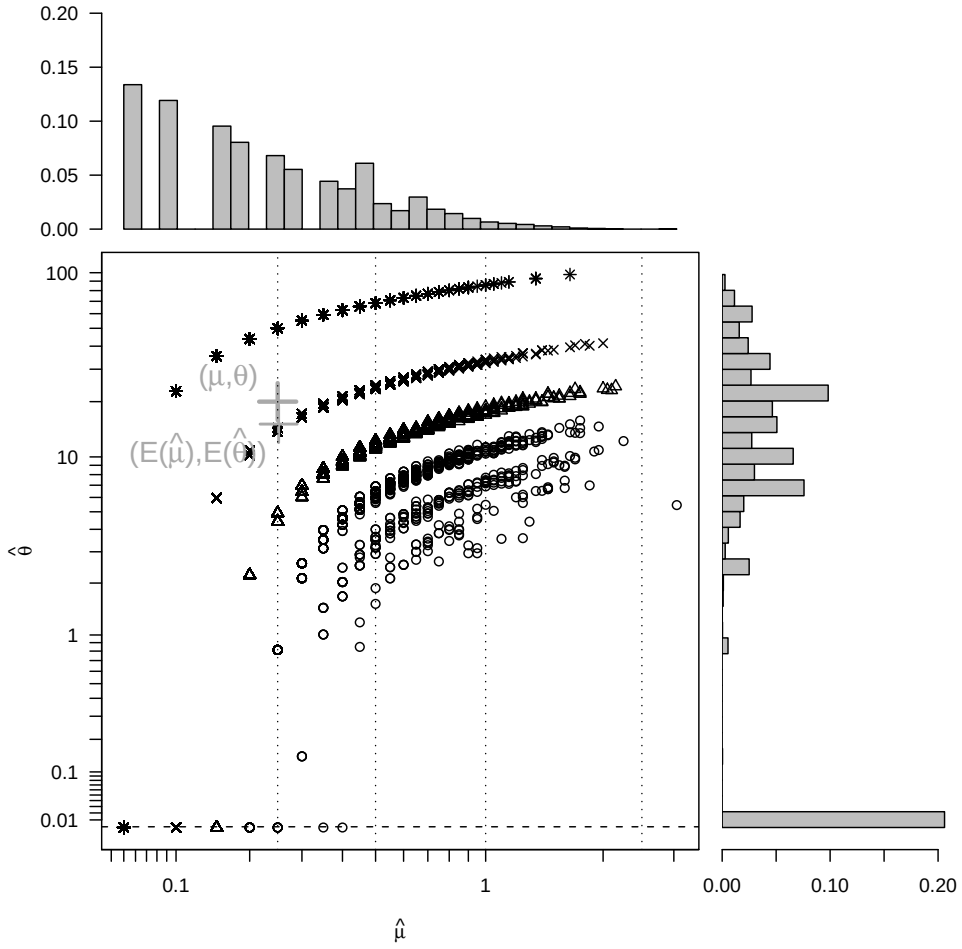


FIGURE I.2.1 – Valeurs de $\hat{\mu}$ et $\hat{\theta}$ (estimations par maximum de vraisemblance de μ et θ), pour 10 000 échantillons de taille $n = 20$ simulés dans une loi NB($\mu = 0.25, \theta = 20$). La droite horizontale discontinue correspond à la valeur fixée de $\hat{\theta}$ dans le cas où les échantillons sont sous-dispersés. Les droites verticales discontinues correspondent aux valeurs de moyennes pour lesquelles le nombre d'individus T est égal à 5, 10, 20 et 50 respectivement. Les échantillons pour lesquels $(\hat{\mu}, \hat{\theta})$ est indiqué par une astérisque, une croix et un triangle présentent un nombre de comptages non nuls S égal à 1, 2, et 3 respectivement. Un seul point peut représenter, par superposition, plusieurs échantillons : les histogrammes permettent d'évaluer la fréquence d'occurence des couples $(\hat{\mu}, \hat{\theta})$.

nière à ce qu'on échantillonne un nombre assez important d'individus des espèces les plus communes, il ne fournira néanmoins qu'un nombre restreint d'individus des espèces les plus rares. Ainsi, si l'on considère l'ensemble des échantillons considérés, $T \leq 5$ pour 19% des échantillons, $T \leq 10$ pour 36% des échantillons, et $T \leq 20$ pour 55% des échantillons. Ainsi, si l'on exclut une partie des échantillons des analyses, on a tout intérêt à le faire de manière « raisonnée ».

I.2.3 Etude de la précision en fonction du nombre d'individus : quelques conséquences pratiques

Dans l'article 2, nous avons étudié certaines propriétés de l'estimateur par maximum de vraisemblance de Θ (biais, écart-type et RMSE) en fonction de la valeur de T. Nous nous sommes placés dans des cas a priori problématique : faibles valeurs de T ($T \leq 40$), et petites tailles d'échantillon ($n = 20$ et $n = 50$). Outre la valeur de T, et la valeur de n, les propriétés de l'estimateur dépendent de la valeur de θ : schématiquement, plus la valeur de θ est importante, et plus l'estimation est problématique.

Les résultats obtenus dépendent donc en grande partie de la valeur réelle de θ, qui par définition est inconnue. En d'autres termes, les résultats de cette étude ne sont interprétables, par exemple dans le but de faire un tri rationnel des échantillons basé sur T, que sous réserve d'avoir un a priori quant à la gamme de valeurs dans laquelle se situe θ.

Considérons par exemple la gamme de valeurs de θ plausible pour nos données poisson (cf section I.1.3, p.30). Si l'on s'intéresse, par exemple, à la RMSE de (Θ), on peut pondérer les résultats obtenus concernant $RMSE_{n,\theta}(\Theta \mid T = t)$ par la distribution a priori de θ. On obtient ainsi la valeur de $RMSE_n(\Theta \mid T = t)$ pour tout $t \in [\![2, 40]\!]$:

$$RMSE_n(\Theta \mid T = t) \quad \approx \quad \sum_{k=1}^{k_{max}} RMSE_{n,\theta \in [\theta_k, \theta_{k+1}]}(\Theta \mid T = t) pr(\theta \in [\theta_k, \theta_{k+1}]$$

où l'on fait l'approximation suivante

$$RMSE_{n,\theta \in [\theta_k, \theta_{k+1}]}(\Theta \mid T = t) \approx \frac{1}{2} \left[RMSE_{\theta = \theta_k}(\Theta \mid T = t) + RMSE_{\theta = \theta_{k+1}}(\Theta \mid T = t) \right]$$

en choisissant 371 valeurs discrètes θ_k, régulièrement espacées sur une échelle logarigthmique, et allant de 0.01 à 200.

On obtient les résultats correspondant à la figure I.2.2. Dans le cas $n = 20$, exlure tous les échantillons tels que $T \leq 10$ (c'est à dire environ la moitié des échantillons) correspond à une situation où

$RMSE \leq 10.6$ (c'est à dire une baisse de la RMSE d'au moins 19% par rapport aux cas où $T = 2$). Dans le cas $n = 50$, exlure tous les échantillons tels que $T \leq 10$ (c'est à dire environ 22% des échantillons) correspond à une situation où $RMSE \leq 12.0$ (c'est à dire une baisse de la RMSE d'au moins 47% par rapport aux cas où $T = 2$).

Si l'on vérifie l'influence du nombre d'individus sur la fiabilité de l'estimation de dispersion, l'erreur reste néanmoins importante y compris pour des nombres d'individus T relativement élevés. Dans le cas $n = 20$, notamment, en écartant près de 50% des données on atteint une RMSE qui demeure supérieure à 10.

Bien évidemment, la précision des estimations pourrait être améliorée, moyennant l'augmentation des tailles d'échantillon. Néanmoins, la taille d'échantillon est souvent limitée du fait de contraintes pratiques ou budgétaires. En ce sens, l'amélioration de l'estimation « à effort d'échantillonnage égal » est un problème essentiel en biostatistiques. En particulier, nous avons montré ici que la qualité de l'estimation de θ était meilleure quand le nombre d'individus observés T était important. Ce nombre T étant lui-même une variable aléatoire, on pourrait adopter une méthode d'échantillonnage « séquentiel » , dans lequel la taille d'échantillon est adaptée séquentiellement aux observations de manière à ne pas observer un nombre d'individus T trop faible. L'influence de telles méthodes d'échantillonnage sur la qualité des estimations a été étudiée dans l'article 4. J'y reviendrai donc dans le chapitre 4 de cette synthèse.

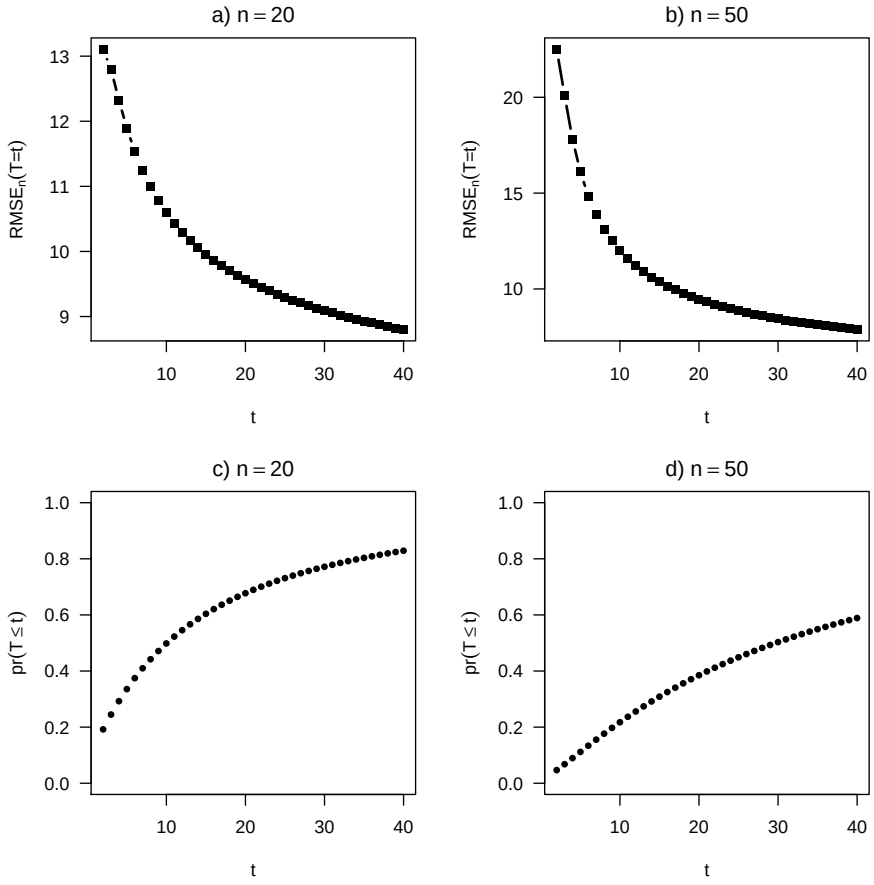


FIGURE I.2.2 – RMSE en fonction du nombre d'individus t observés dans l'échantillon (figures a et b), et fonction de répartition de la variable T (figures c et d), pour une taille d'échantillon n donnée (figures a et c : $n = 20$; figures b et d : $n = 50$)

Chapitre 3

Estimation par intervalle de la moyenne μ

Il n'y a pas de cause d'erreur plus fréquente
que la recherche de la vérité absolue.
Samuel Butler

I.3.1 Une grande diversité de méthodes de calcul

Il existe une littérature extensive sur le sujet de l'estimation d'intervalles pour la moyenne. En particulier, comme déjà évoqué dans l'introduction (cf section I.0.2, p. 17) de nombreuses études ont souligné le besoin, pour de petits échantillons surdispersés, de vérifier la validité des intervalles de confiance. Ici, nous n'avons pas tenté de faire un inventaire complet des méthodes de calcul d'intervalles de confiance, mais plutôt de tester certaines méthodes assez classiques ou éprouvées, et d'autres plus récentes et plus spécifiquement adaptées au type de données qui nous intéresse.

Ainsi, nous avons testé les quatre méthodes suivantes, détaillées dans l'article 3 :

S : méthode de Student

B : méthode de Bernstein

L : méthode du profil de vraisemblance (Profile Likelihood)

J : méthode de Jeffreys

Nous avons utilisé la méthode de Student (cf p. 17) comme exemple de méthode « classique » de calcul d'intervalle de confiance.

Néanmoins, pour des données surdispersées, les scientifiques, conscients de la non-normalité des données, utilisent très souvent cette méthode sur des données préalablement log-transformées. Plus précisément, pour des données de comptage, la transformation est de la forme $ln(x+a)$, où $a > 0$. Bien que l'article 3 ne traite pas de ce type de méthode, nous avons souhaité illustrer dans cette synthèse quelques problèmes allant de pair avec l'estimation d'intervalles de confiance pour la moyenne sur données log-transformées. Nous avons ainsi testé les méthodes *l1*, *l2*, *C1* et *C2* :

*l*1 : méthode de Student sur données log-transformées : $z = ln(x+1)$

l2 : méthode de Student sur données log-transformées : $z = ln(x+0.1)$

C1 : méthode de Cox corrigée avec transformation : $z = ln(x+1)$

C2 : méthode de Cox corrigée avec transformation : $z = ln(x+0.1)$

Par ailleurs, la table I.3.1 détaille quelques caractéristiques des huit différentes méthodes d'estimations testées ici.

La figure I.3.1 illustre la variabilité des intervalles de confiance estimés selon diverses méthodes pour les quatre échantillons y_1, y_2, y_3, et y_4, décrits dans l'annexe (p.67). Dans ces quatre exemples, on observe que CI_L et CI_J sont, dans certains cas (par exemple pour l'échantillon y_4), beaucoup plus larges que CI_S et CI_B. Les intervalles de confiance calculés sur des données log-transformées sont généralement beaucoup moins larges que CI_S, CI_B, CI_L et CI_J. De plus, comme indiqué dans la table I.3.1, ils correspondent à une estimation de la moyenne différente de la moyenne arithmétique.

Si la figure I.3.1 illustre l'ampleur des différences pouvant exister entre les intervalles de confiance estimés selon diverses méthodes, elle ne permet pas de trancher sur la méthode la plus adaptée, dans la mesure où l'on ne connaît pas la valeur réelle de μ. Une étude basée sur des simulations (avec μ connu) comme celle réalisée pour l'article 3, permet de trancher.

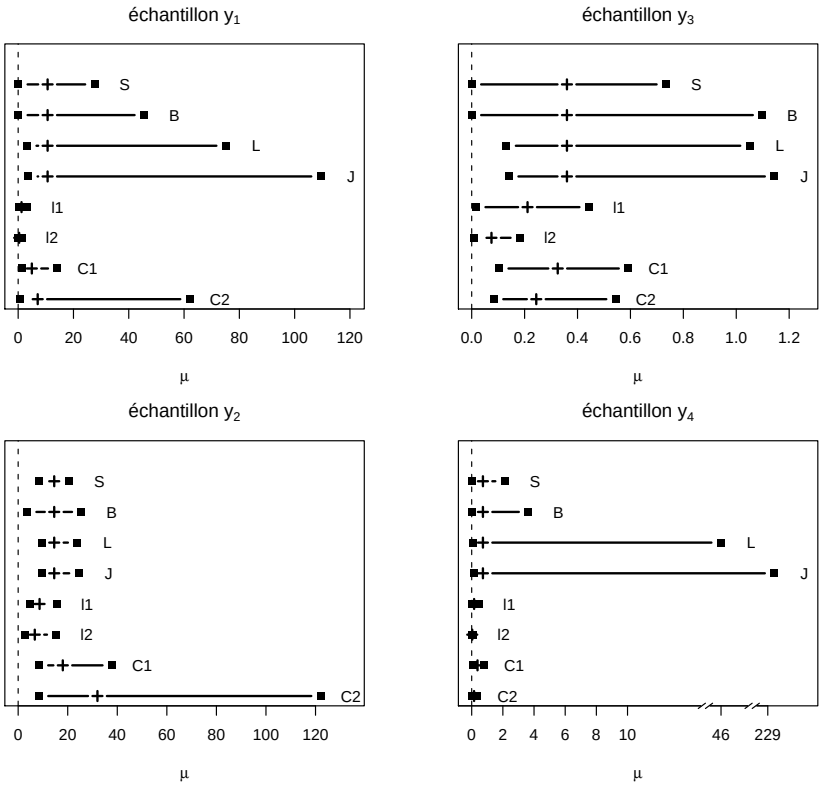


FIGURE I.3.1 – Quelques intervalles de confiance pour la moyenne des quatre exemples d'échantillons. Les quatre intervalles de confiance CI_S, CI_B, CI_L et CI_J correspondent à ceux qui sont étudiés plus en détail dans l'article 3. Les quatre intervalles de confiance CI_{l1}, CI_{l2}, CI_{C1} et CI_{C2} correspondent à des intervalles de confiance calculés sur les données log-transformées. Les croix (+) représentent les moyennes estimées.

TABLE I.3.1 – Quelques caractéristiques des différentes méthodes d'intervalles de confiance

	Symétrie	Basée sur TCL	Basée sur vraisemblance (NB)	Méthode bayésienne	Transformation z=f(x)	Estimation de la moyenne
S	oui	oui	non	non	non	$\bar{x}$
B	oui	non	non	non	non	$\bar{x}$
L	non	non	oui	non	non	$\bar{x}$
J	non	oui	oui	oui	non	$\frac{(n\bar{x}-0.5)}{(n+\hat{\theta})}$
l1	non	oui	non	non	$z = ln(x+1)$	$e^{\bar{z}} - 1$
l2	non	oui	non	non	$z = ln(x+0.1)$	$e^{\bar{z}} - 0.1$
C1	non	oui	non	non	$z = ln(x+1)$	$e^{\bar{z}+s_z^2} - 1$
C2	non	oui	non	non	$z = ln(x+0.1)$	$e^{\bar{z}+s_z^2} - 0.1$

I.3.2 Résultats concernant les méthodes détaillées dans l'article 3

Les résultats de l'article 3 démontrent la supériorité des intervalles de confiance basés sur la vraisemblance (méthodes *L* et *J*), par rapport aux méthodes *S* et *B*. Les méthodes *S* et *B* correspondent à des intervalles de confiance symétriques autour de la moyenne estimée. Ils tendent à fréquemment sous-estimer la moyenne (c'est à dire que la moyenne réelle est au dessus de la borne supérieure de l'intervalle). De plus, leur bonne inférieure est très souvent non-informative car inférieure à 0. Cette fréquente sous-estimation aboutit à une probabilité de couverture généralement inférieure à la valeur nominale (ici, 95%), en particulier pour l'intervalle de Student, *S*. Cette tendance est particulièrement forte quand la dispersion θ est importante.

En effet, en utilisant une méthode basée sur la vraisemblance (i.e. sur un modèle NB), on obtient des intervalles de confiance asymétriques (de côté droit plus long que le côté gauche, cf fig. I.3.1), d'autant plus larges, et d'autant plus asymétriques que la dispersion estimée $\hat{\theta}$ est forte. Les intervalles de confiance CI_L et CI_J tendent ainsi à être plus larges que les intervalles CI_S et CI_B, en particulier quand $\hat{\theta}$ est élevé.

En effet, si les données sont distribuées selon une NB de dispersion θ relativement forte, alors

l'occurence d' « amas » d'individus (ou de bancs, dans le cas des poissons) est assez rare, mais possible. Plusieurs cas, illustrés par la figure I.3.2, peuvent ainsi se présenter lors de l'échantillonnage.

(a) On n'observe aucun amas (échantillon y_3). On estime alors que la dispersion est faible ($\hat{\theta} = 1.0$). Le profil de vraisemblance diminue rapidement lorsqu'on s'éloigne de $\hat{\mu}$.

(b) On observe beaucoup d'individus et d'abondances non-nulles (échantillon y_2). L'estimation de θ est moyenne ($\hat{\theta} = 4.0$).

(c) On observe, entre autres, un de ces amas (échantillons y_1 et y_4). On estime donc que la dispersion est forte ($\hat{\theta}$=11.4 et 41.1, respectivement). Le profil de vraisemblance est relativement « plat » à droite $\hat{\mu}$.

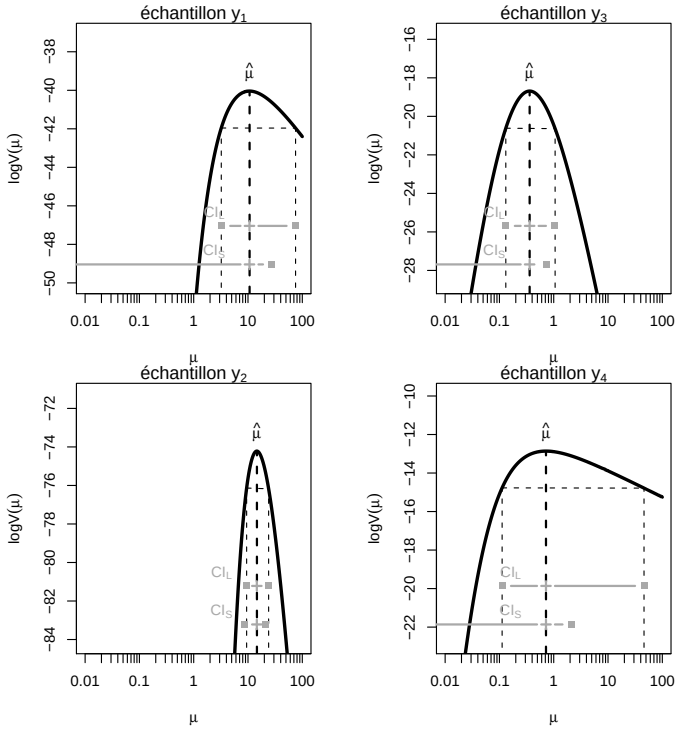


FIGURE I.3.2 – Profil de vraisemblance pour la moyenne μ, en considérant $\theta = \hat{\theta}$, pour les quatre exemples d'échantillons. Les intervalles CI_S et CI_L sont représentés (en gris).

I.3.3 Résultats concernant les méthodes impliquant une log-transfomation des données

Comme évoqué dans la partie I.0.2 (p. 17), l'intervalle de Student, CI_S, repose sur le fait que, asymptotiquement, $(\bar{X}-\mu)/(s/\sqrt{n}) \sim \mathscr{S}(n-1)$. Or, la valeur de n suffisante pour faire cette approximation est d'autant plus forte que la distribution de X est différente d'une gaussienne. Ici, les tailles d'échantillon n sont particulièrement faibles et les distributions particulièrement asymétriques. Il est donc peu raisonnable de faire cette approximation, et en effet la probabilité de recouvrement de la méthode S est particulièrement faible.

Par conséquent, le calcul d'intervalles de confiance pour des données dont la distribution est asy-

métrique est fréquemment réalisé sur des données transformées $z = f(x)$. On choisit la transformation de sorte que la distribution des données transformées z soit proche d'une loi normale. Ainsi, l'approximation $(\bar{Z} - \mu_z)/(s_z/\sqrt{n}) \sim \mathscr{S}(n-1)$ est raisonnable pour une valeur de n relativement faible.

La log-transformation est très souvent utilisée dans ce but, y compris pour des données de comptage (Bland and Altman, 1996; O'Hara and Kotze, 2010). Elle est même recommandée dans certains ouvrages de références en biostatistiques (voir par exemple Sokal and Rohlf, 1995, pp. 413-415). Elle revient, en quelque sorte, à faire l'hypothèse que la distribution des données x est approximativement log-normale.

Pour les méthodes *l1* et *l2*, l'intervalle de confiance pour la moyenne est calculé comme suit :

$$CI_{l1/l2} = f^{-1}\left[\bar{z} - q_{\mathscr{S},n-1,\alpha/2}\frac{s_z^2}{\sqrt{n}}, \bar{z} + q_{\mathscr{S},n-1,1-\alpha/2}\frac{s_z^2}{\sqrt{n}}\right]$$

où

$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$

$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z-\bar{z})^2$

f^{-1} est la transformation inverse de f, de la forme $x = e^z - a$.

En outre, on suppose que $\bar{z} = f(\mu)$, i.e. $E(f(X)) \approx f(E(X))$. On estime donc μ par $\hat{\mu} = f^{-1}(\bar{z})$.

Or cette estimation de μ est biaisée. En effet, comme le soulignent Zhou and Gao (1997) et O'Hara and Kotze (2010), un des problèmes des méthodes *l1* et *l2* est qu'elles ne fournissent pas, à proprement parler, un intervalle de confiance pour μ, mais pour la moyenne géométrique des valeurs $(x+a)$. La moyenne géométrique de $(x+a)$ étant moins influencée par les fortes valeurs x que la moyenne arithmétique, elle tend à être plus faible. Or, les résultats correspondant à la méthode *S* (méthode de Student sur données brutes) indiquent déjà un problème de sous-estimation de la moyenne. Les méthodes *l1* et *l2* correspondront donc vraisemblablement à une sous-estimation extrêmement fréquente de la moyenne. En ce sens, les méthodes *l1* et *l2*, quoiqu'extrêmement classiques pour des données de comptage (O'Hara and Kotze, 2010), correspondent à des méthodes assez naïves et discutables (Zhou and Gao, 1997; Olsson, 2005; O'Hara and Kotze, 2010).

Zhou and Gao (1997) ont ainsi testé la méthode de Student sur des données log-transformées de la forme $z = ln(x)$ (i.e. une méthode semblable à *l1* et *l2* à ceci près que a=0). En supposant que les données x sont distribuées selon une loi log-normale de paramètres ($m_l = -1, \sigma_l = 1.42$) (i.e. $E(X) = 1$ et $Var(X) = 6.4$), ils montrent que la probabilité de recouvrement des intervalles de Student à 90 %

sur données log-transormées est de 24% seulement. En revanche, en testant dans les mêmes conditions une méthode dite « de Cox », ils montrent que cette méthode a une probabilité de recouvrement bien meilleure, de 86%.

Nous testons donc également la méthode de Cox corrigée (Olsson, 2005), a priori mieux appropriée que les méthodes *l1* et *l2* au calcul d'intervalles de confiance pour la moyenne sur données log-transformées. Notons néanmoins que dans notre cas, le calcul de tels intervalles de confiance implique que l'on fait une hypothèse supplémentaire par rapport à Zhou and Gao (1997), qui est que loi log-normale constitue une approximation satisfaisante de la distribution des données (i.e., d'une binomiale négative).

L' intervalle de Cox corrigé est de la forme :

$$CI_{C1/C2} = \left[E(f(\mu)) - q_{\mathscr{S},n-1,\alpha/2} se(f(\mu)) \ , \ E(f(\mu)) + q_{\mathscr{S},n-1,1-\alpha/2} se(f(\mu)) \right]$$

où $q_{\mathscr{S},n-1,\alpha/2}$ et $q_{\mathscr{S},n-1,1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi de Student à $(n-1)$ degrés de liberté. La méthode de Cox non-corrigée (Land, 1972; Zhou and Gao, 1997) correspond à ce même intervalle de confiance où les quantiles de la loi de Student $\mathscr{S}(n-1)$ sont remplacés par les quantiles de la loi normale $\mathscr{N}(0,1)$.

On estime $E(f(\mu))$ et $se(f(\mu))$ de la manière suivante :

$$\begin{aligned} \widehat{E(f(\mu))} &= \bar{z} + \frac{s_z^2}{2} \\ \widehat{se(f(\mu))} &= \frac{s_z^2}{n} + \frac{s_z^4}{2(n-1)} \end{aligned}$$

Procédons à une étude par simulation du même type que celle réalisée dans le cadre de l'article 3, mais en nous intéressant cette fois-ci aux intervalles de confiance *l1*,*l2*, *C1* et *C2*.

Les résultats correspondant aux figures I.3.3 et I.3.4 montrent que les intervalles de confiance de Student calculés sur données log-transformées, *l1* et *l2*, ont une probabilité de recouvrement très faible. Comme prévu, les méthodes *l1* et *l2* correspondent à une très fréquente sous-estimation de la moyenne (cf fig. I.3.3). La probabilité de recouvrement de *l1* -respectivement *l2*- est ainsi au mieux de 75% -respectivement 18%- (fig. I.3.4). Ces méthodes sont ainsi encore plus problématiques que la méthode de Student appliquée à des données non transformées. Les probabilités de recouvrement des intervalles de Cox *C1* et *C2* sont meilleures que celles de *l1* et *l2*. Néanmoins elles demeurent très faibles pour les fortes valeurs de dispersion θ. Ces résultats montrent très clairement que log-transformer les données pour calculer un intervalle de confiance pour la moyenne est une pratique à proscrire dans le cas de

données de comptage surdispersées. Notons que ce résultat ne qualifie en rien la pertinence de la log-transformation, par exemple pour l'utilisation d'un modèle linéaire (McArdle and Anderson, 2004).

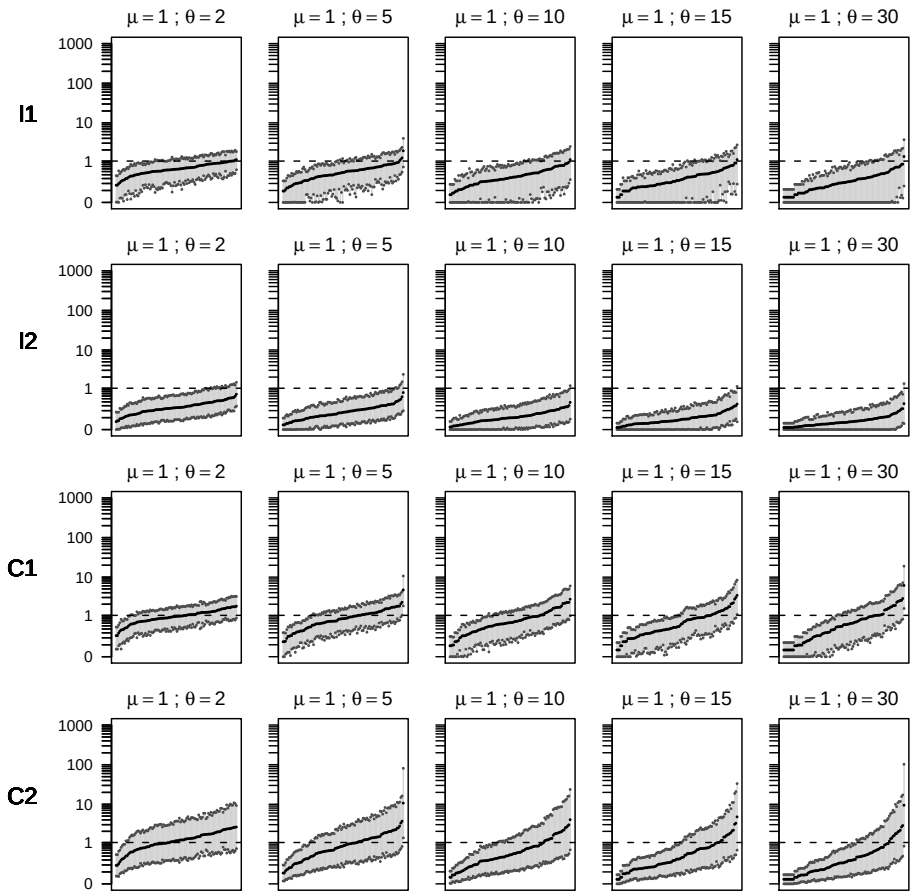


FIGURE I.3.3 – Intervalles de confiance pour la moyenne des échantillons simulés, pour les quatre méthodes testées, dans le cas $\mu = 1$ and $n = 20$. Les points gris foncés sont les bornes des intervalles de confiance, et les points noirs représentent l'estimation de la moyenne. La droite horizontale discontinue représente la valeur réelle de μ.

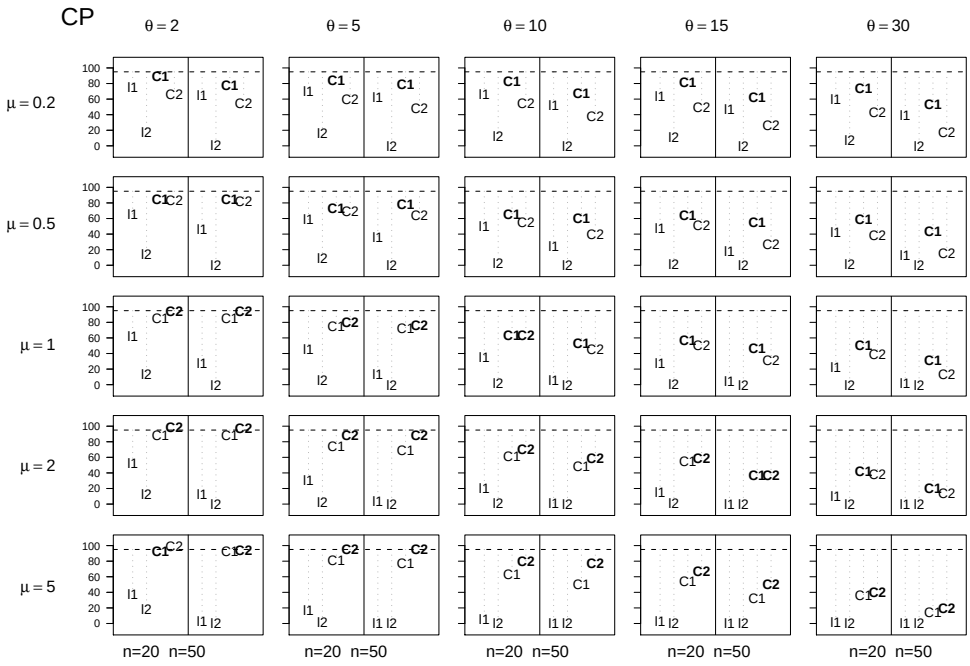


FIGURE I.3.4 – Probabilité de couverture (CP, en %) des quatre méthodes de construction d'intervalle de confiance, en fonction des valeurs de paramètres(μ, θ). La droite horizontale discontinue représente la valeur nominale des intervalles de confiance (i.e. 95%). La méthode en caractères gras est celle pour laquelle la CP est la plus proche du niveau nominal.

Chapitre 4

Echantillonnage à taille fixe ou séquentiel

Chacun vaut ce que valent
les objectifs de son effort.
Marc-Aurèle

Contrairement à l'échantillonnage à taille fixe pour lequel la taille d'échantillon est fixée à l'avance et ne dépend pas des observations, l'échantillonnage séquentiel consiste à adapter la taille d'échantillon aux observations.

I.4.1 Echantillonner pour obtenir une précision fixée

Dans le cas de l'échantillonnage à taille fixe, Young and Young (1998) soulignent que « la difficulté première dans la détermination d'une taille d'échantillon optimale n [...] est que, dans chaque cas, la taille d'échantillon optimale dépend d'un paramètre inconnu ou plus. On peut utiliser des valeurs « raisonnables » pour ces paramètres inconnus pour déterminer approximativement la taille d'échantillon. Cependant, on n'est jamais sûr que d'une taille d'échantillon prédéterminée résulteront des estimations de la densité de population μ ayant le niveau désiré de précision » (traduction libre) [1].

[1] The primary difficulty in determining the optimal sample size n [...] is that, in each case, the optimal sample size depends upon one or more unknown parameters. A worker may put in some « reasonable » values for these unknown parameters to

Dans le cas de l'échantillonnage séquentiel, certaines règles d'arrêt d'échantillonnage peuvent être définies de manière à ce que l'estimation de μ atteigne un degré de précision spécifié à l'avance (Mukhopadhyay et al, 1992; Johnson et al, 1996; Mukhopadhyay et al, 1997; Young and Young, 1998). Cependant, comme le remarquent Willson et al (1984), Johnson et al (1996), et Young and Young (1998), ces méthodes reposent sur une connaissance, a priori, de la dispersion des données. Or, dans notre cas, la dispersion θ est inconnue et doit être estimée sur le même échantillon de données que μ. Ainsi, si adapter la taille d'échantillon de manière à obtenir la précision voulue est un particularité remarquable de ces méthodes d'échantillonnage, elles peuvent difficilement être appliquées dans notre cas.

I.4.2 Echantillonner pour obtenir une meilleure précision

Les études précédentes (articles 2 et 3) ont mis en évidence le fait qu'à certains échantillons, correspondent des estimations de θ peu précises, et ce de manière « intuitive ». En particulier, l'estimations de θ peut être particulièrement erronée quand il y a peu d'individus dans l' échantillon (cf article 2). Parallèlement, les intervalles de confiance ont tendance à être extrêmement larges dans le cas où l'échantillon comprend peu de valeurs non-nulles (cf article 3), en lien avec une très forte estimation $\hat{\theta}$. Ainsi, comme le soulignent Willson et al (1984), « En étudiant les estimations de $[1/\theta]$ basées sur des échantillons de taille fixe, nous avons remarqué que quelques estimations étaient extrêmement mauvaises. Il apparaissait, intuitivement, que de telles estimations seraient radicalement différentes si quelques observations étaient ajoutées à l'échantillon (traduction libre) » [2].

Dans le cas de la base de données poissons, ces cas problématiques sont d'autant plus inévitables que les tailles d'échantillon fixées sont faibles, et que les moyennes d'abondance sont plutôt faibles et les dispersions fortes.

Par conséquent, nous avons choisi de nous intéresser aux propriétés des estimateurs dans le cas de l'échantillonnage séquentiel, dans le cas où les règles d'arrêt d'échantillonnage dépendent du nombre d'individus observés ou du nombre d'abondances non-nulles. Ainsi, en reprenant une notation déjà

determine approximate sample sizes. However, one can never be sure whether a predetermined fixed sample size will result in estimates of the population density μ with the desired level of precision.

[2] Upon studying the estimates of $[1/\theta]$ based on samples of fixed size, we noticed that a few of the estimates were extremely poor. It was intuitively apparent that such estimates would change dramatically if a few more observations were added to the sample

utilisée précédemment, nous avons considéré les variables T et S correspondant au nombre total d'individus et d'éléments non-nuls des échantillons :

$$\begin{aligned} T &= \sum_{i=1}^{n} X_i \\ S &= \sum_{i=1}^{n} I(X_i \neq 0) \end{aligned}$$

où $I(X_i \neq 0)$ est la variable indicatrice de l'événement $\{X_i \neq 0\}$. Nous avons testé des méthodes d'échantillonnage séquentiel dont la règle d'arrêt repose sur T et S. Nous avons fixé les critères d'arrêt d'échantillonnage, de la forme $T \geq t_{min}$ ou $S \geq s_{min}$, en lien avec les caractéristiques de la base de données poissons. Dans cette base de données données, $T \leq 15$ pour 47% des échantillons, $T \leq 30$ pour 64% des échantillons, $S \leq 5$ pour 50% des échantillons, et $S \leq 10$ pour 81% des échantillons.

Comme pour l'article 2, nous nous sommes intéressés à la précision des estimations par maximum de vraisemblance, non seulement de θ, mais aussi de μ. Nous avons évalué cette précision à travers la RMSE (Root Mean Square Error), et nous avons comparé les valeurs de RMSE d'une part dans le cas où la taille d'échantillon n est fixée, et d'autre part dans le cas où l'échantillonnage est séquentiel, i.e. la taille d'échantillon N_{seq} est une variable aléatoire. Nous avons réalisé cette comparaison « à taille d'échantillon égale, en moyenne », i.e. en fixant n tel que $n = E(N_{seq})$.

Cette étude a révélé que l'échantillonnage séquentiel correspond généralement, comme on s'y attendait, à une estimation plus précise de θ. Néanmoins, elle correspond à une estimation moins précise de μ.

I.4.3 Echantillonner pour obtenir une meilleure précision quant à la moyenne

Le fait que la qualité de l'estimation de la moyenne par maximum de vraisemblance tend à être moins bonne dans le cas d'un échantillonnage séquentiel correspond notamment à l'existence d'un biais. En effet, avec ces règles d'arrêt basées sur T ou S, la dernière observation correspond forcément à une abondance non nulle. L'échantillonnage correspond donc à une légère tendance à la surestimation de la moyenne (Whitehead, 1986). La recherche d'estimateurs non biaisés dans le cas de l'échantillonnage séquentiel est ainsi une piste importante pour l'amélioration de l'inférence (de Cristofaro, 2004; Bunouf, 2006).

Par ailleurs, dans l'article 4 nous nous sommes intéressés à l'estimation de μ non par intervalle, mais par point. Si l'échantillonnage séquentiel correspond à une estimation de μ par point moins précise, elle pourrait néanmoins correspondre à une meilleure estimation par intervalle, dans la mesure où elle fournit de meilleures estimations $\hat{\theta}$. Nous avons donc étendu ici les simulations de l'article 4 au calcul d'intervalles de confiance. Nous nous sommes intéressés, en particulier, à l'intervalle de Jeffreys (CI_J), dont nous avons montré dans l'article 3 que les propriétés étaient particulièrement intéressantes.

La table I.4.1 rassemble ainsi les probabilités de couverture de CI_J, pour les 4 distributions de (μ, θ) et les quatre règles d'arrêt d'échantillonnages considérés, dans le cas $n_{min} = 2$ (les résultats correspondant à l'estimation par point -par opposition à l'estimation par intervalle- figurent donc dans la table II.4.1 de l'article 4). Cette table montre que généralement, l'estimation de μ par l'intervalle CI_J est légèrement meilleur dans le cas de l'échantillonnage séquentiel, notamment quand la règle d'arrêt d'échantillonnage repose sur S.

TABLE I.4.1 – Probabilité de couverture de CI_J, l'intervalle de confiance de Jeffreys pour la moyenne, pour les quatre règles d'arrêt et les quatre distributions de (μ, θ) considérées.

Stopping rule	Param. range		N_{seq}		CP(CI_J)	
	m_μ	m_θ	$E(N_{seq})$	$sd(N_{seq})$	fix (%)	seq (%)
$T \geq 30$	-0.4	2	97	146	88.1	90.7
	-0.4	1	93	144	91.4	91.8
	1	2	31	45	84.9	87.7
	1	1	28	43	89.7	88.3
$T \geq 15$	-0.4	2	52	76	89.2	89.5
	-0.4	1	49	74	91.5	90.4
	1	2	20	24	84.0	86.9
	1	1	17	22	88.5	86.6
$S \geq 10$	-0.4	2	72	70	89.3	90.8
	-0.4	1	52	56	91.9	92.4
	1	2	42	37	82.6	84.3
	1	1	26	22	89.3	90.0
$S \geq 5$	-0.4	2	36	38	88.3	89.9
	-0.4	1	26	29	89.4	90.5
	1	2	22	20	85.3	86.3
	1	1	15	11	87.7	87.9

Chapitre 5

Conclusion et perspectives

Dans un couple, l'un au moins doit être fidèle,
de préférence l'autre.
Marcel Achard

I.5.1 Conclusions : estimation du couple (μ,θ)

Dans les articles 2, 3, et 4, nous nous sommes intéressés à l'estimation par maximum de vraisemblance du couple (μ,θ). Nous avons montré, dans l'article 3, l'influence de $\hat{\theta}$ sur l'incertitude entourant μ. En effet, les bornes des intervalles de confiance L et J dépendent de la valeur de $\hat{\theta}$. Dans la plupart des cas, $\hat{\theta} < \theta$, ce qui aboutit à une probabilité de couverture des intervalles de confiance inférieure au niveau de confiance nominal. Dans le cas contraire, quand $\hat{\theta}$ est particulièrement élevé (par exemple quand il y a, outre des zéros, une ou deux abondances fortes dans l'échantillon), l'intervalle de confiance pour de la moyenne est tellement large qu'il en est non-informatif. Réciproquement, nous avons montré, dans l'article 2, l'influence de $\hat{\mu} = T/n$ sur l'incertitude entourant θ. Quand $\hat{\mu} < \mu$, θ a tendance a être sévèrement sous-estimé. En d'autres termes, les erreurs d'estimation d'un paramètre sont d'autant plus importantes que l'on traite l'autre comme un paramètre de « nuisance », (i.e. que l'on fait l'hypothèse que son estimation est fidèle à sa valeur réelle).

Le constat général selon lequel l'estimation de la dispersion (ou de la variance) est rendue plus difficile par l'estimation de la moyenne sur le même jeu de données est, somme toute, assez classique

en statistiques (voir par exemple Harville, 1977). Dans un autre contexte (modèles linéaires mixtes notamment), l'étude de ce problème a pu, par exemple, aboutir aux techniques d'estimations par REML (« Restricted maximum likelihood », ou maximum de vraisemblance restreint).

Il serait ainsi intéressant d'estimer un paramètre, non pas en considérant l'autre comme connu, mais au contraire en prenant en compte la double incertitude dans laquelle on se trouve. L'examen de la vraisemblance « en deux dimensions » (contrairement au cas du profil de vraisemblance, cf fig. I.3.2, p.46) pourrait-il permettre de comprendre les problèmes liés à cette double incertitude ?

La figure I.5.1 montre la vraisemblance du couple (μ,θ) pour les quatre exemples d'échantillon y_1, y_2, y_3 et y_4 (décrits p.69). On pourrait interpréter les vraisemblances de (μ,θ) de la manière suivante :

1. Pour l'échantillon y_1, comprenant seulement quelques abondances non-nulles, l'incertitude quant à (μ,θ), et plus particulièrement μ, est forte. En effet, la dispersion estimée est particulièrement forte, entraînant une incertitude quant à μ particulièrement importante.

2. Pour l'échantillon y_2, comprenant beaucoup d'individus et d'abondances non-nulles, l'incertitude quant à (μ,θ) est faible. Non seulement la dispersion estimée est faible, mais il y a également une faible incertitude quant à cette estimation.

3. Pour l'échantillon y_3, comprenant peu d'individus, l'incertitude quand à (μ,θ), et surtout quant à θ est forte. En effet, la dispersion estimée est faible, mais il y a une forte incertitude quand à cette estimation. Par ailleurs, l'incertitude quant à θ a une forte influence sur l'incertitude quant à μ.

4. Pour l'échantillon y_4, comprenant très peu de points non-nuls, l'incertitude quant à (μ,θ) est très forte.

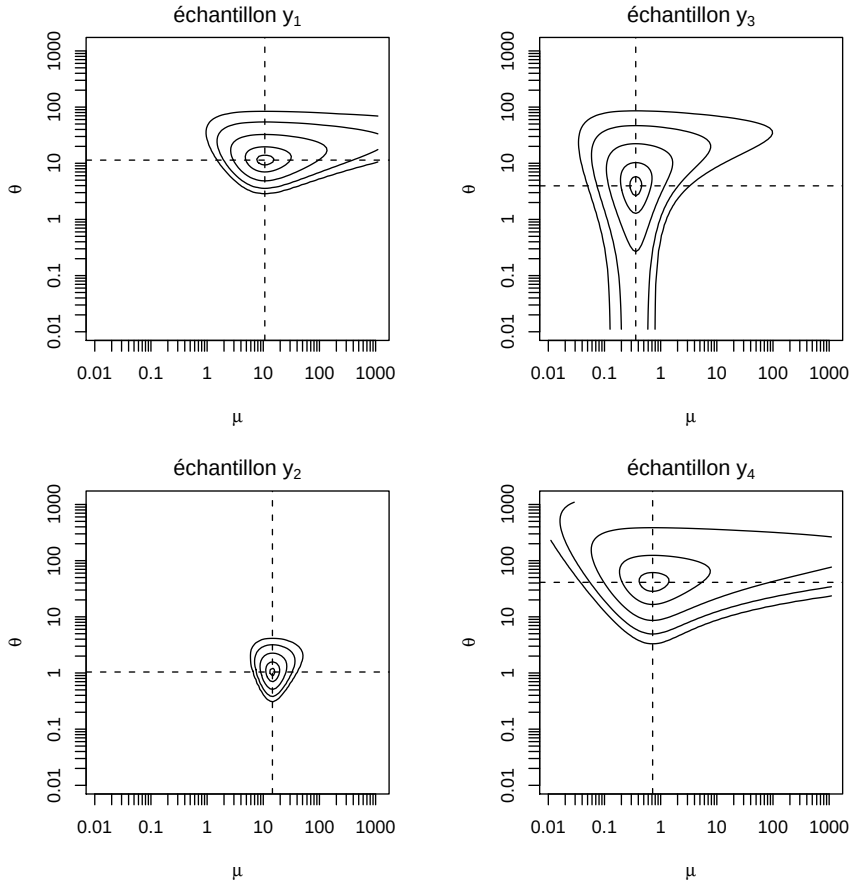


FIGURE I.5.1 – Vraisemblance des exemples d'échantillon. Les contours indiquent les vraisemblances égales à 0.001, 0.01, 0.1, 0.5 et 0.9 fois la vraisemblance maximale

La vraisemblance de (μ,θ) illustre donc en partie la double incertitude affectant l'estimation de ces paramètres. On pourrait tenter d'aller plus loin dans l'interprétation de cette vraisemblance, par exemple en quantifiant la double incertitude affectant les estimations de μ et θ. En particulier, un moyen de la quantifier serait de fournir une région de confiance (généralisation de la notion d'intervalle de confiance à plusieurs -ici,2- dimensions) basée sur la vraisemblance de (μ,θ).

On cherche ainsi à généraliser la méthode du profil de vraisemblance (cf fig. I.3.2, p.46), qui permet de calculer des intervalles de confiance basés sur la vraisemblance d'un paramètre en considérant l'autre comme fixé.

La méthode du profil de vraisemblance consiste à déduire les quantiles de la loi de μ de $pr_{\mu,\theta=\hat{\theta}}(y)$ (vraisemblance de μ à θ fixé, d'où le nom de « profil »). En effet, dans la méthode du profil de vraisemblance, le test de ratio de vraisemblance fournit $[a,b]$ tels que $pr_{\theta=\theta_t}(\mu \in [a,b]) = p\%)$. On peut ainsi connaître approximativement la distribution de μ en supposant que $\{\theta = \theta_t\}$, en calculant par exemple les couples $(a_p,b_p)_{p\in[\![1,99]\!]}$ tels que

$$\forall p \in [\![1,99]\!] \;\; pr_{\theta=\theta_t}(\mu \in [a_p,b_p]) = p/100$$

De la même manière, on peut connaître approximativement la distribution de θ en supposant que $\{\mu = \mu_t\}$.

Ainsi, partant des lois conditionnelles de $(\mu \mid \theta)$ et $(\theta \mid \mu)$, l'échantillonnage de Gibbs permet d'estimer la loi de (μ,θ) ainsi que les lois marginales de μ et θ :

1. On calcule la distribution de θ en supposant que $\{\mu = \mu_t\}$.

2. On tire au hasard une valeur θ_{t+1} dans cette distribution.

3. On calcule la distribution de μ en supposant que $\{\theta = \theta_{t+1}\}$.

4. On tire au hasard une valeur μ_{t+1} dans cette distribution.

En prenant $\mu_0 = \hat{\mu}, \theta_0 = \hat{\theta}$ et par itération (1500 fois) de la séquence ci-dessus, on obtient $(\mu_1,\mu_2,\ldots,\mu_{1500})$ et $(\theta_1,\theta_2,\ldots,\theta_{1500})$. On considère que la distribution de $(\mu_{500},\mu_{501},\ldots,\mu_{1500})$ et $(\theta_{500},\theta_{501},\ldots,\theta_{1500})$

La figure I.5.2 montre la distribution conjointe estimée de (μ,θ), ainsi que les intervalles de confiance CI_{2D} correspondants (calculés à l'aide des distributions marginales) de μ et de θ. Au vu de ces quatre exemples, il semblerait que prendre en compte la variabilité conjointe des deux paramètres ne modifie que légèrement les estimations d'incertitude. En ce sens, les trop faibles probabilités de couverture des intervalles L (profil de vraisemblance) ne s'expliqueraient pas uniquement par le paramètre de nuisance, mais également, sans doute, par le fait que les hypothèses du test de ratio de vraisemblance ne soient pas vérifiées. En particulier, tout comme le Théorème Central Limite, dont nous avons souligné qu'il était inadapté à nos données dans la section I.0.2 (p.17) ce test est valable dans un cadre asymptotique (voir par exemple Sprott, 1975, pour une discussion de ce problème).

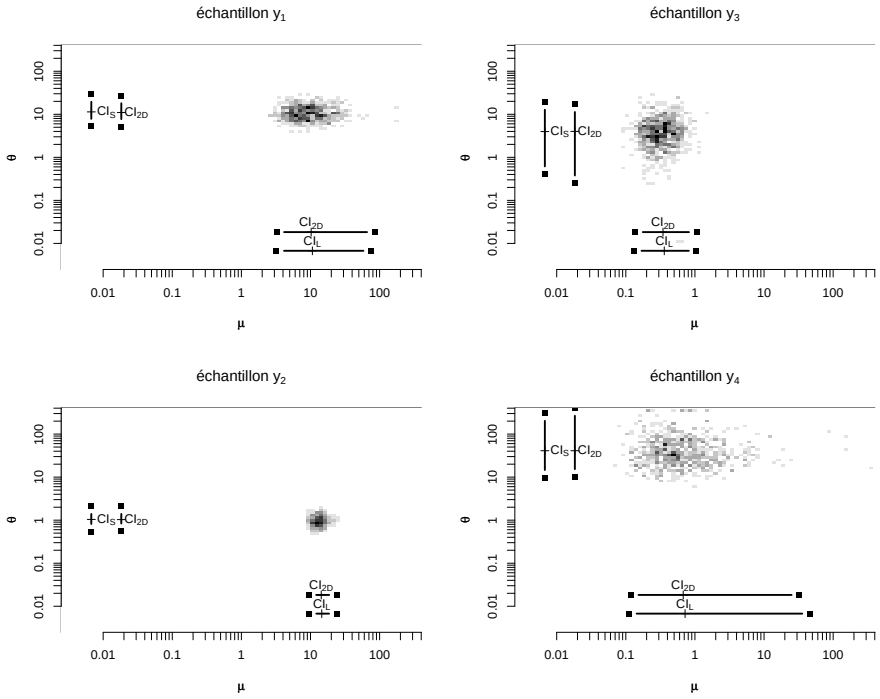


FIGURE I.5.2 – Réalisations de (μ, θ) issues d'un échantillonnage de Gibbs, et intervalles de confiance à 95% CI_L (calculés par la méthode du profil de vraisemblance) et CI_{2D} (calculés à partir des quantiles des distributions marginales de μ et θ.

De manière générale, on a accordé dans ce travail de thèse une place importante à l'estimation par maximum de vraisemblance, notamment du paramètre de dispersion θ. Il s'agissait là d'un parti pris correspondant à notre volonté d'étudier les problèmes d'estimation de ce paramètre en nous focalisant non pas sur les estimateurs, mais sur les caractéristiques des échantillons (article 2) ou sur la méthode d'échantillonnage (article 4). En effet, en dépit de ses éventuels défauts, la vraisemblance est un outil extrêmement classique en statistiques, de par son interprétabilité et la multitude des ses applications. Néanmoins, la recherche de meilleurs estimateurs, et la réflexion sur les propriétés de la vraisemblance demeurent bien évidemment un aspect essentiel de l'étude des problèmes d'estimation. En particulier, la proposition de nouveaux estimateurs, et l'étude analytique de leurs propriétés (voir par exemple Saha and Paul, 2005, en ce qui concerne les estimateurs de θ) constituent une piste essentielle à l'améliora-

tion de l'inférence, dans le cas où les échantillons sont de taille limitée.

I.5.2 Perspectives : à la recherche de l'information

Les résultats obtenus lors de ce travail de thèse suggèrent que les problèmes d'estimation sont en partie liés non pas aux estimateurs, mais simplement au fait que les échantillons ne sont pas assez « informatifs », par exemple lorsqu'ils ne comprennent que très peu d'individus ou de comptages non-nuls. Si ce manque d'information ne peut pas être compensé par un effort d'échantillonnage accru, alors il peut être judicieux d'intégrer un maximum d'informations dans les inférences réalisées sur ces échantillons. En particulier, il pourrait être intéressant d'avoir recours à des méthodes d'estimation bayésiennes, et ainsi d'intégrer l'information correspondant à des distributions a priori de (μ,θ).

Dans ce travail de thèse, nous avons adopté des méthodes d'inférence fréquentiste, à l'exception des intervalles de confiance de Jeffreys. Ces intervalles reposent sur un prior de Jeffreys, dit « non-informatif », dans le sens où il n'apporte aucune information qui ne soit déjà contenue dans l'échantillon et le modèle de distribution. En effet, le prior de Jeffreys repose uniquement sur la vraisemblance de l'échantillon (plus exactement, sur l'information de Fisher, elle-même dérivée de la vraisemblance). En ce sens, le recours à un prior de Jeffreys se justifie aisément, y compris d'un point de vue fréquentiste, et correspond à une forme d'« estimation par maximum de vraisemblance améliorée ». Néanmoins, dans la mesure où l'on constate que les résultats basés sur la vraisemblance peuvent parfois être erronés du fait des tailles d'échantillon limitées, il est peu plausible que les résultats dérivés de la vraisemblance comme l'information de Fisher soient totalements satisfaisants.

Dans l'article 3, on a montré qu'aux intervalles de confiance pour la moyenne reposant sur un prior de Jeffreys (CI_J), correspondaient de meilleures probabilités de couverture que pour les méthodes de Student, de Bernstein, et, dans une moindre mesure, du profil de vraisemblance (cette dernière méthode donnant, logiquement, des résultats très proches de CI_J). Néanmoins, deux problèmes majeurs demeurent, que CI_J ne parvient pas à corriger totalement : d'une part, les probabilités de couverture sont généralement plus faibles que le niveau de confiance nominal, d'autre part, certains intervalles de confiance sont si larges que l'intervalle de confiance fournit une information quasi-nulle quant au paramètre de moyenne.

Un prior informatif quant à μ et θ constitue sans aucun doute une piste intéressante pour améliorer l'inférence. Dans la section I.1.3, page 27, on fournit des priors pour μ et θ, qui pourraient être utilisés

pour obtenir de meilleurs estimations. Naturellement, ces priors ont pour effet de « resserrer » les estimations $\hat{\mu}$ et $\hat{\theta}$ autour de la valeur du mode des distributions de μ et θ, i.e. 0.16 et 2.72, respectivement (voir figure I.5.2). Cette figure illustre que l'utilisation d'un prior diminue l'incertitude des estimations, et conduit à des estimations moins extrêmes que dans le cas fréquentiste.

Ce travail de thèse permet de prendre la mesure des incertitudes entourant les estimations de moyenne et dispersion, pour des données de comptage surdispersées. Selon les méthodes d'estimations ou d'échantillonnage adoptées, les estimations peuvent-être plus ou moins précises. Néanmoins, un choix d'estimateur ou de méthode d'échantillonnage judicieux ne peuvent pallier le fait que l'information fournie par les échantillons est, de fait, limitée. Il est ainsi souhaitable de tempérer toute estimation fournie par de tels échantillons, par la connaissance, a priori, que l'on a des populations.

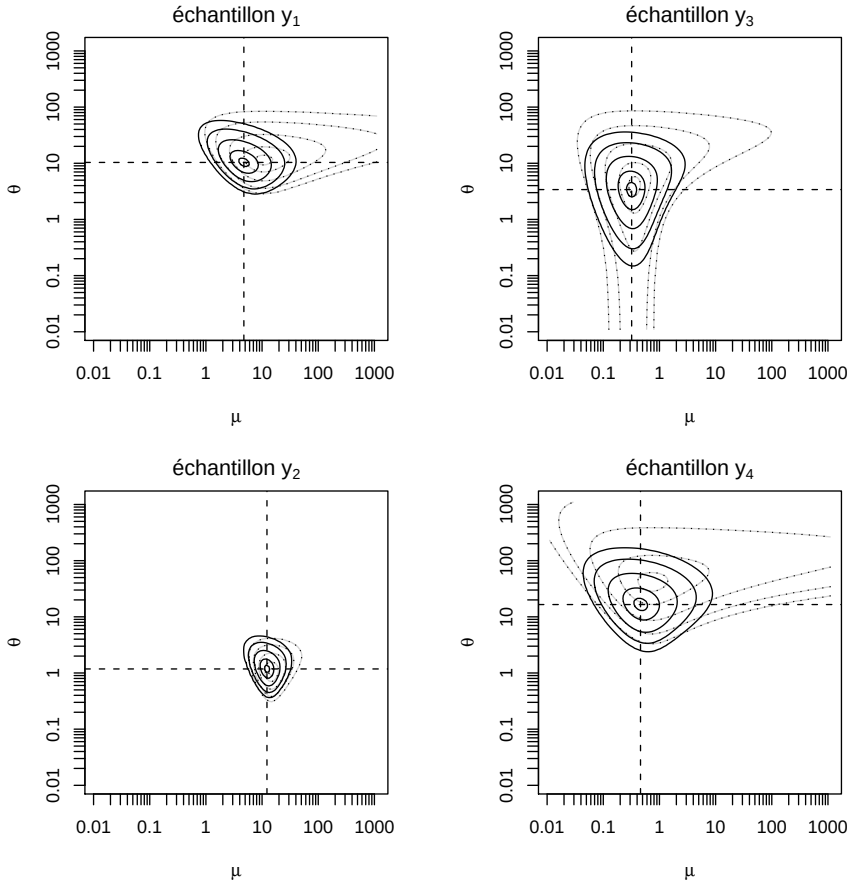


FIGURE I.5.3 – Distribution a posteriori de (μ, θ) (en noir) pour les quatre échantillons y_1, y_2, y_3 et y_4. La vraisemblance apparaît en gris. Les contours indiquent les niveaux de probabilité égaux à 0.001, 0.01, 0.1, 0.5 et 0.9 fois la probabilité maximale.

Annexe A

Données de comptage de poissons

La base de données d'abondances de poissons sur laquelle nous avons travaillé correspond à un effort collectif, auquel ont participé plusieurs équipes de l'Université Lyon 1, le Cemagref, le bureau d'études Aralep, et Henri Persat. La collecte de ces données a été financée pour partie par la Compagnie Nationale du Rhône, l'Agence de l'Eau Rhône-Méditerrannée-Corse, la Région Rhône-Alpes ainsi que d'autres collectivités locales.

I.A.1 Echantillonnage

La base de données inclut 350 campagnes de pêche électrique réalisées entre 1985 et 2007 (voir Persat, 1988; Lamouroux et al, 1999, pour un exemple d'utilisation antérieure de ces données). Les sites d'échantillonnage, au nombre de 25, sont situés sur de larges cours d'eau (plusieurs dizaines de mètre de large, environ 1m de profondeur à bas débit) : ils correspondent à 18 tronçons du Rhône et 7 tronçons de l'Ain (Olivier et al, 2009). Tous les sites incluent plusieurs alternances « mouille-radier », et reflètent ainsi l'hétérogénéité de l'habitat des communautés de poissons (Lamouroux et al, 1999). Parmi les sites situés sur le Rhône, 15 se situent sur une portion du fleuve court-circuitée pour la production d'hydroélectricité, en aval d'un barrage (barrages de Chautagne, Belley, Brégnier-Cordon, Péage-de-Roussillon, Montélimar). Les trois autres sites du Rhône, et les sept sites de l'Ain sont également sujets à des variations artificielles des débits.

Chacune des 350 campagnes de pêche correspondent à une durée de un à quatre jours consécutifs,

et ont généralement lieu à débit faible ou moyen. Durant chaque campagne, 20 à 180 points ont été échantillonnés selon une stratégie « d'échantillonnage ponctuel d'abondance » (Persat and Copp, 1990; Copp, 1992; Lamouroux et al, 1999). On pêche chaque point en lançant ou mettant l'électrode dans l'eau et en récupérant tous les poissons attirés à l'aide d'une épuisette. On évalue l'aire d'attraction de l'électrode à environ 7m^2 (Regis et al, 1981). La pêche est réalisée en bateau, ou à pied quand cela est possible, et les points sont aléatoirement ou régulièrement espacés dans le cours d'eau dans les zones où la pêche électrique est efficace (profondeur < 2m, courant $< 1\text{m.s}^{-1}$). La distance entre les points est d'au moins 30m, de sorte que nous avons considéré que les abondances $(x_1, x_2, \ldots, x_n)$ pour une campagne étaient indépendantes.

Nous nous sommes intéressés aux abondances de chaque espèce considérée individuellement, indépendamment de la présence ou de l'abondance d'autres espèces sur la même campagne ou sur le même point d'échantillonnage. Chaque campagne représente ainsi à un certain nombre d'échantillons de données, correspondant au nombre d'espèces observées lors de la campagne. Au total, en ne considérant que les espèces pour lesquelles on a observé au moins 3 individus sur une campagne donnée, on dispose ainsi de 2823 échantillons (104 446 individus de 31 espèces).

I.A.2 Description des échantillons de données

Un total de 2258 échantillons (96 665 individus de 12 espèces) a finalement été analysé. Ces échantillons correspondent aux 12 espèces les plus fréquentes (i.e. aux 12 espèces pour lesquelles on dispose de plus de 100 échantillons, cf table I.A.1).

La plupart des échantillons sont de taille réduite : $n \in [20, 30]$ dans 78.3% des cas, $n \in [31, 50]$ dans 9.8% des cas, $n \in [51, 100]$ dans 9.3% des cas, et $n \in [101, 180]$ dans les 2.5% de cas restants.

La table I.A.2 rassemble quelques statistiques descriptives de ces échantillons, qui sont généralement largement surdispersés (i.e. variance observée > moyenne observée), à l'exception des échantillons pour lesquels la moyenne d'abondance est particulièrement faible. Tous les échantillons ont une distribution asymétrique à droite : plus de 50% des échantillons comprennent au moins 80% de zéros. Les moyennes d'abondance sont généralement assez faibles : plus de 50% des échantillons ont une moyenne ≤ 0.6.

TABLE I.A.1 – Nombre d'échantillons et d'individus des 12 espèces les plus fréquentes (nombre d'échantillons)

Espèce	Code	Nom commun	Nombre d'échantillons	Nombre d'individus
Squalius cephalus (Linnaeus, 1758)	CHE	Chevaine	327	16424
Gobio gobio gobio (L., 1758)	GOU	Goujon	230	7589
Rutilus rutilus (L., 1758)	GAR	Gardon	214	19754
Chondrostoma nasus (L., 1758)	HOT	Hotu	211	6178
Barbus barbus (L., 1758)	BAF	Barbeau	203	5529
Barbatula barbatula (L.,1758)	LOF	Loche	193	5953
Alburnus alburnus (L., 1758)	ABL	Ablette	179	10336
Phoxinus phoxinus (L.,1758)	VAI	Vairon	173	14141
Leuciscus leuciscus (L.,1758)	VAN	Vandoise	156	4565
Alburnoides bipunctatus (Bloch, 1782)	SPI	Spirlin	143	3370
Perca fluviatilis (L.,1758)	PER	Perche	118	921
Lepomis gibbosus (L., 1758)	PES	Perche soleil	112	1909

I.A.3 Quatre exemples d'échantillons de données

La figure I.A.1 présente la distribution empirique des données de comptage correspondants à quatre échantillons, y_1, y_2, y_3 et y_4, issus de la base de données poissons. Il s'agit des données pour une même espèce (gardon) sur deux sites :

$$
\begin{aligned}
y_1 &= (0,0,1,0,3,0,3,4,162,0,0,0,0,0,31,0,0,0,9,0)\\
y_2 &= (11,11,2,1,26,0,16,8,3,23,14,42,38,15,36,14,20,9,2,0)\\
y_3 &= (0,1,0,0,0,0,0,0,1,0,0,1,0,0,2,0,0,0,0,0,0,0,0,4,0)\\
y_4 &= (1,0,0,0,0,0,0,0,0,0,0,0,0,0,17,0,0,0,0,0,0,0,0,0,0)
\end{aligned}
$$

TABLE I.A.2 – Statistiques descriptives des échantillons : taille, moyenne, variance, proportion de zéros, et coefficient de dissymétrie

	Taille d'échantillon	Moyenne	Variance	Proportion de zéros	Coefficient de dissymétrie
Minimum	20	0.02	0.02	0.05	0.46
1^{er} quartile	25	0.24	0.56	0.68	2.26
Médiane	25	0.60	2.59	0.80	3.05
3^{me} quartile	25	1.56	16.41	0.88	3.96
Maximum	180	33.40	14644.21	0.99	13.07
Moyenne	34.9	1.52	62.95	0.77	3.30

J'ai choisi ces exemples d'échantillons comme « représentatifs »de divers cas pouvant se présenter dans la base de données poissons. Ainsi, pour certains échantillons la moyenne observée est assez forte ($\bar{y_1}$=10.65, $\bar{y_2}$=14.55), et pour d'autres elle est au contraire plutôt faible ($\bar{y_3}$=0.36, $\bar{y_4}$=0.72). De même, pour certains échantillons la variance observée est plutôt forte ($Var(y_1)$=1319, $Var(y_4)$=11.54), et pour d'autres elle est au contraire plutôt faible ($Var(y_2)$=166, $Var(y_3)$=0.80) par rapport à la moyenne. J'utilise ces échantillons à plusieurs reprises dans la synthèse pour illustrer certains problèmes d'estimations (intervalles de confiance pour la moyenne, vraisemblance, etc.).

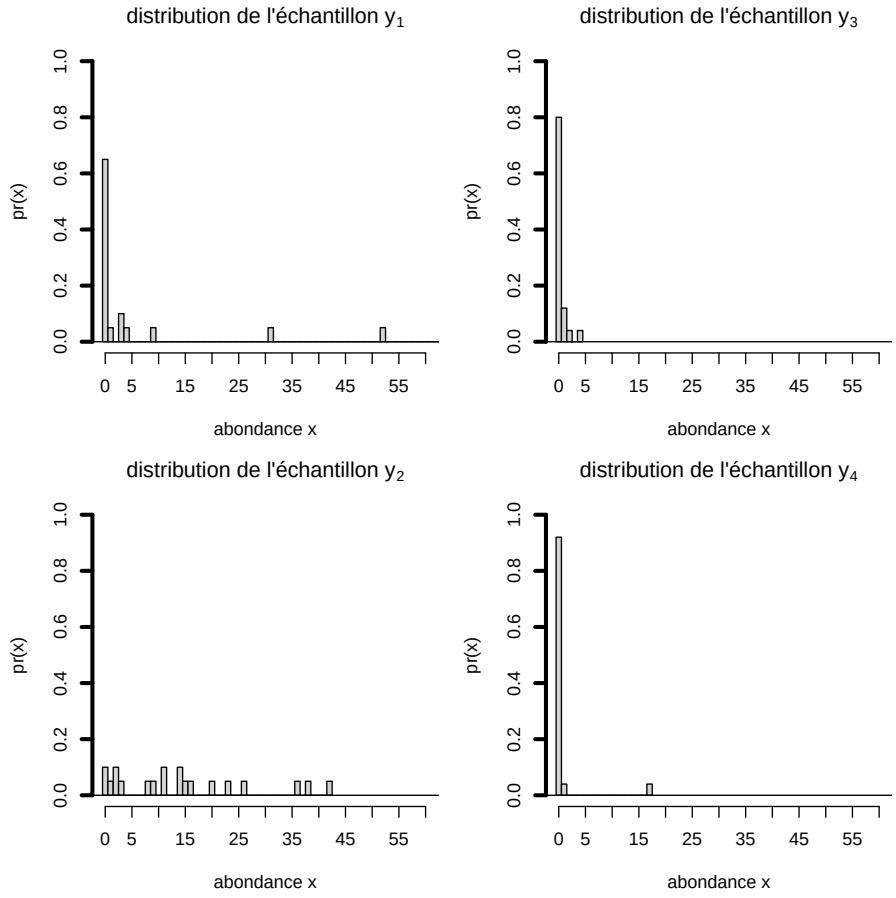


FIGURE I.A.1 – Quatre exemples de distribution empirique des échantillons

Part II

Articles

Chapter 1

Comparing distribution models for small samples of overdispersed counts of freshwater fish

Vaudor, L., Lamouroux, N., Olivier, J.M.

Abstract

The study of species abundance often relies on repeated abundance counts whose number is limited by logistic or financial constraints. The distribution of abundance counts is generally right-skewed (i.e. with many zeros and few high values) and needs to be modelled for statistical inference. We used an extensive dataset involving about 100 000 fish individuals of 12 freshwater fish species collected in electrofishing points ($7m^2$) during 350 field surveys made in 25 stream sites, in order to compare the performance and the generality of four distribution models of counts (Poisson, negative binomial and their zero-inflated counterparts). The negative binomial distribution was the best model (Bayesian Information Criterion) for 58% of the samples (species - survey combinations) and was suitable for a variety of life histories, habitat, and sample characteristics. The performance of the models was closely related to samples' statistics such as total abundance and variance. Finally, we illustrated the consequences of a distribution assumption by calculating confidence intervals around the mean abundance, either based on the most suitable distribution assumption or on an asymptotical, distribution-free (Student's) method. Student's method generally corresponded to narrower confidence intervals, especially when there were few ($\leq$3) non-null counts in the samples.

Keywords: Abundance; Negative binomial distribution; Poisson distribution; Zero-inflation; Profile likelihood; Confidence intervals

II.1.1 Introduction

A common issue in ecology consists of modelling the abundance of organisms and the environmental determinants of this abundance, using repeated counts. In many cases, authors pay minor attention to the general shape of the distribution of counts and focus on summary properties such as the variance-to-mean relationship. Nevertheless, the need to characterize this distribution is increasingly acknowledged in the literature for ecological and statistical reasons.

Ecological interest lies in the determination of the biotic and abiotic factors that can influence the distribution of abundance counts. For example, abundance heterogeneity might be linked to the heterogeneity of the environment, and the high frequency of zero counts can be due to unsuitable habitats (Martin et al, 2005). More generally, variations in the distribution shape might reflect differences in species behaviour, due to e.g. population density, season, presence of predators or competitors (Taylor, 1984). Abundance samples repeated in time and space are particularly important for identifying the generality of a given distribution model and inferring the limits of its relevance.

Concerning statistical reasons, inference and estimation of abundance often require a distributional assumption for abundance counts. An inappropriate assumption has many implications, including misleading confidence intervals around the estimate of mean abundance (Rosenblum and Van der Laan, 2008), or unreliable error rates with traditional statistical analyses such as linear models (Power and Moser, 1999; McArdle and Anderson, 2004).

Poisson-mixture distributions are widely recognised as potentially appropriate distribution models for count data. In particular, the negative binomial has a long history as a suitable model for clumped count data (Anscombe, 1949; Bliss and Fisher, 1953; Evans, 1953) and is increasingly used in abundance count studies. Recent developments have discussed extending the Poisson or negative binomial distributions into models that account for extra zeros (Martin et al, 2005; Wenger and Freeman, 2008). Some studies have used extensive ecological data sets to compare the performance of these distributions under a variety of environmental and biological conditions (Welsh et al, 1996; Ridout et al, 1998; Gray, 2005; Martin et al, 2005; Warton, 2005; Potts and Elith, 2006; Wenger and Freeman, 2008). Nevertheless, few studies of this kind concern freshwater fish (Wenger and Freeman, 2008; Lewin et al, 2009).

Freshwater systems are often heavily impacted by human activities, and the study of fish abundance and of its environmental determinants is of particular importance to develop objective diagnostic

and management tools for these systems (Lamouroux et al, 1999; Lewin et al, 2009). Estimating fish abundance is particularly challenging in medium and large rivers due to the overdispersion of counts, that can result from schooling behaviour and/or strong habitat preferences (Lamouroux et al, 1999). The rarity of many species and the limitations in sampling effort due to budget constraints adds further difficulties (Lewin et al, 2009).

Studies comparing distribution models for count data generally focus on modelling the distribution of one or a few large size samples (i.e. set of counts) rather than the distribution of many small size samples, repeated in space and time (Warton, 2005). Still, repeated small size samples are extremely common in ecological studies (for instance for monitoring species). When sample size is limited, due to logistic or financial constraints, it is of particular importance that inference is carried out based on a sound, parsimonious, and general distribution model. It is therefore important to identify which sample characteristics affect the performance and generality of candidate distribution models.

In this paper, we compared the performance of four distributions (Poisson, negative binomial and their zero-inflated counterparts) for modelling observed abundance counts of freshwater fish. Models were tested on a large existing data set involving about 100 000 fish individuals of 12 species collected during 350 field surveys made in 25 stream sites of the Rhône basin (France) between 1985 and 2007. On each of these repeated surveys, a limited number (20 to 180) of independent abundance counts (electrofishing of a fixed area) were made. After selecting the best model for each survey-species combination according to the Bayesian Information Criterion (BIC), we used multinomial logistic regressions to identify which factors (e.g., total abundance, belonging to a species or a site) influenced this selection. Finally, we illustrated the consequences of a distributional assumption for estimating confidence intervals around the mean abundance estimate.

II.1.2 Methods

Sites and fish sampling

The initial fish dataset considered included 104 446 individuals belonging to 31 species collected during 350 field surveys made in 25 stream sites of the Rhône basin, between 1985 and 2007 (see Persat, 1988; Lamouroux et al, 1999, for previous use of part of the data). The sampling sites were 18 stream reaches of the Rhône river (the largest river in France) and seven reaches of its tributary the Ain river (Olivier et al, 2009). All sites were long enough to include several "pool-riffle" sequences, i.e. they reflected the habitat heterogeneity in which the fish communities mostly live. Fifteen of the Rhône sites were situated in a section of the river by-passed for hydropower generation, downstream from a dam (dams of Chautagne, Belley, Brégnier-Cordon, Péage-de-Roussillon, Montélimar). The three other Rhône sites and the seven Ain sites were not in by-passed sections but were also subject to artificial flow regimes. All sites were situated on large regulated rivers (several dozen meters wide, around 1m deep at low flow) and sites had varied habitat characteristics (Lamouroux et al, 1999).

Each of the 350 field surveys was an electrofishing survey made during one to four consecutive days, generally at low to medium flow rates. During each field survey, 20 to 180 points were sampled using a "Point Abundance Sampling" strategy (Persat and Copp, 1990; Copp, 1992; Lamouroux et al, 1999). Points were electrofished by dropping or putting the electrode in the river and net-catching all fishes attracted. The attraction area of the electrode was approximately 7m^2 (Regis et al, 1981). Electrofishing was made from a boat or wading in the stream where possible, and points were randomly or regularly spaced within the reach in areas where electrofishing was efficient (water depth $<$ 2m, water velocity $<1\text{m.s}^{-1}$). The distance between points was $>$30m, and we considered that the abundance counts were independent.

Statistical modelling

We modelled the abundance data separately for the different species-survey combinations (called samples hereafter). In other words, a sample x_i with i in $1,\ldots,n$, corresponded to n abundance counts (n is between 20 and 180, according to the survey) of a single species in a single stream site. Having few samples of abundance for the rarest species could make the analyses of the effect of species on model selection particularly unbalanced. To limit this problem we excluded from our study the species that

were present in less than 100 surveys. Having really few individuals in a sample could make the use of a distribution model irrelevant. To limit this problem we excluded from our study the samples that had less than three individuals. This limit of three individuals is quite low, and the relevance of modelling a distribution where the total number of individuals is this low will be discussed.

We considered several candidate Poisson-mixture statistical distributions of abundance counts, that are routinely used to model count data. The Poisson (P) distribution accounts for discrete and positive counts but cannot account for overdispersion (when variance is higher than mean), which is a frequent feature of abundance counts. The negative binomial (NB) distribution is a generalization of the P model, that allows for overdispersion (Johnson et al, 1992). This overdispersion may for instance be interpreted as resulting from the heterogeneity of habitat, or from a gregarious behaviour of the species under study. The P and NB distribution can also be extended to account for an additional proportion of extra-zeros: they are then called zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), respectively (Lambert, 1992). This kind of model has lately been much studied in the ecological literature (Welsh et al, 1996; Ridout et al, 1998; Gray, 2005; Martin et al, 2005; Potts and Elith, 2006; Wenger and Freeman, 2008). Zero-inflation is mainly interpreted as reflecting the existence of two kinds of zeros: structural zeros, which occur because the sampling is carried out in conditions such that the presence of an individual is impossible, and stochastic zeros, which occur simply by chance.

Many other candidate distribution models, and in particular hurdle models, were not included in our modelling exercise. Hurdle models (also called two-stage models) are another type of zero-modified counts models (Mullahy, 1986; Ridout et al, 1998; O'Neill and Faddy, 2003; Cunningham and Lindenmayer, 2005; Gray, 2005). Contrary to zero-inflated models that reflect a latent zero process increasing the proportion of zeros (Gray, 2005), they treat positive counts separately from zeros. With hurdle models, the occurrence of a positive count is treated as a binomial process, and the positive counts are generally modelled by a truncated-at-zero count data model. The underlying idea is that distinct ecological processes govern the presence or absence of individuals and the number of individuals given presence (Mullahy, 1986). We did not include these models in the present study because we considered that zero-inflated models were more realistic from an ecological point of view. Indeed, it has been demonstrated that freshwater fish species often have null counts in their suitable habitat: some of the zeros we observe are probably stochastic zeros (Lamouroux et al, 1999; Martin et al, 2005).

We thus considered four potential models for our data: P, NB, ZIP, and ZINB, whose probability

mass functions are given in Table II.1.1. P and NB are special cases of ZIP and ZINB respectively, with the zero-inflation parameter $1-\tau$ equal to zero. P and ZIP are special cases of NB and ZINB respectively, with the dispersion parameter θ tending towards 0. This implies that the model that best fits the data will necessarily be the ZINB and the one that fits it worst will be the P model. On the other hand, the P model is the most parsimonious (only one parameter) while the ZINB model is the least parsimonious (three parameters).

We fitted the four models to each sample with the maximum likelihood method, using the function "nlminb" of the statistical software package R (R Development Core Team, 2010), which carries out constrained minimization using a Newton-type algorithm. This function maximizes the log-likelihood within the one to three-dimensional (depending on the distribution model considered) parameter space. We set the constraints of lower and upper bounds for the parameter sets accordingly to the limits indicated in Table II.1.1.

Comparison of models

For each sample, we selected the best distribution model, using a criterion balancing goodness of fit and parsimony. Preliminary comparisons of different criteria among which Akaike Information Criterion (AIC), AIC corrected for small samples (AICc) and Bayesian Information Criterion (BIC) provided comparable results. We finally used BIC:

$$BIC = -2ln(L) + kln(n)$$

where L is the maximum likelihood of the sample, k the number of parameters of the model and n the sample size (i.e. the number of electrofished points). The differences in performance of models reflect their adequacy for describing the observed distribution of individuals. These differences are influenced by both the ecological processes that generate distribution patterns, and sampling fluctuations, which might cause the observed sample to have a distribution quite different from the underlying distribution. Residual plots were used to understand which distribution patterns were more or less captured by the different models.

We analysed the influence of six potential explanatory variables on the selection of a distribution model using a series of multinomial logistic regressions (McCullagh and Nelder, 1989). Specifically, we modelled $log(\pi_j/\pi_{ref})$ (where π_j is the probability that the model selected by the BIC criterion is the distribution j -i.e. either P, ZIP, or ZINB- and π_{ref} is the probability that the model selected is the

NB distribution) as a function of some explanatory variables.

The first four explanatory variables considered were continuous and were: (1) the log-transformed total abundance, $tot.ab$, (2) the log-transformed sample size, *ssize*, (3) an index of relative variance defined as the residuals of the linear regression between sample log-variance and sample log-mean (i.e. Taylor's power law; Taylor, 1984), $rel.var$, and (4) the proportion of zeros, $p.zero$. The variable $rel.var$ was used instead of the sample variance itself to reduce correlation between our explanatory variables. The two remaining variables were ecological factors coded as categories: belonging to a species, and belonging to a reach. All regression models were described by (1) their general performance (the proportion of samples for which the model the most likely according to the regression matches the model selected according to the BIC criterion), (2) their Cohen's kappa statistic (a correction of performance that takes into account the proportion of matches between predicted and observed distribution that are simply due to chance) and (3) their deviance. We also compared nested regression models two by two, carried out likelihood ratio tests to assess the significance of the explanatory variables, and used a discriminant analysis plot to further illustrate the combined influence of explanatory variables.

The influence of our quantitative variables was largely expected. The BIC criterion might be more favourable to parsimonious models when models do not fit the data well. In particular, for small samples comprising few individuals, the fit is generally quite weak. We thus expected small values of the variables $tot.ab$ and *ssize* to favour the selection by the BIC criterion of the P model against the others. In addition, for each distributional hypothesis, expected values of statistics such as mean, variance, and proportion of zeros could be expressed as functions of the distribution parameters. As a consequence, each statistic's expected value could theoretically be expressed as a function of the others, this function depending on the distributional hypothesis considered. In practice, these are generally complex functions that are hard to interpret except through a few general principles: a P model corresponds to situations where the variance is equal to the mean whereas a NB model corresponds to the variance being higher than the mean. In the same way, zero-inflated models apply better when the observed proportion of zeros is higher than predicted by the non-inflated model. Because the influence of quantitative sample statistics on model selection was partly expected, the aim of our multinomial regressions was essentially to quantify its magnitude. For example, a weak explanation of model selection by sample statistics would suggest that the examination of the considered sample statistics is not enough to be able to predict the selection of a distributional model. In particular, we wanted to test if the belonging to a species or a reach could influence model selection once sample statistics are taken into account. Such

an effect would suggest that describing the distribution of abundance for a species or a reach based on sample statistics only, rather than considering the whole shape of the distribution, implies a significant loss of information.

Consequences of a distributional assumption on mean abundance estimation

We illustrated the practical consequences of the choice of a distributional assumption, by analysing its consequences for interval estimation of mean abundance. We thus considered two methods for estimating confidence intervals for the mean: one does not rely on any distributional assumption, while the other one does.

The first method is Student's method, which relies on the fact that asymptotically (i.e. for sufficiently large sample sizes)

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim \mathscr{S}(n-1)$$

where μ is the population mean, $\bar{X}$ is the arithmetic (observed) mean, n is the sample size, s is the estimate of standard deviation of X $\left(s^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)\right)$, and $\mathscr{S}(n-1)$ is Student's distribution with $n-1$ degrees of freedom. This method corresponds to one of the most classical method for constructing confidence intervals (whether the assumption that sample size is sufficiently large is met or not), probably because it is simple, ready-to-use in many statistical software packages. It is also well-known by biologists because it is often the main (if not only) method for building confidence intervals around the mean that is taught to biology students (Meeker and Escobar, 1995).

The second method we used is the profile likelihood method (Venzon and Moolgavkar, 1988; Fletcher and Faddy, 2007), which is based on a distributional assumption (here, the distribution model selected by the BIC). Profile likelihood is the likelihood function (based on the distribution considered) of the parameter of interest, with all other parameters held fixed at the values that maximize the likelihood. A confidence interval can be approximated using the asymptotic properties of the likelihood ratio (Venzon and Moolgavkar, 1988). Cases where zero-inflated models were selected were excluded from this comparison, because their inclusion would strongly complicate our illustrative exercise. Indeed, calculating a confidence interval around the mean through the profile-likelihood method is easy when the mean is a parameter of the model, which is the case for the P and NB distribution. In contrast, it is more complicated for zero-inflated models, for which two sources of uncertainty should be taken into account because the mean abundance is a combination of two parameters (zero-inflation and underly-

ing mean). Hence, although a method to calculate confidence around the mean for zero-inflated models would be of high interest, it is not in the scope of the present study to provide one.

II.1.3 Results

Samples

A total of 2258 of the 2823 available samples, involving 12 species out of 31, were finally analysed based on our selection criteria. This corresponds to 96,665 individuals sampled in the 25 sites during the 350 surveys. The species sampled and the number of samples per species are indicated in Table II.1.2. The size of most samples (78.3%) was between 20 and 30. A proportion of 9.8% of samples comprised between 31 and 50 sampling points, and 9.3% of samples comprised between 51 and 100 sampling points. The 2.5% remaining samples comprised up to 180 sampling points. Table II.1.3 summarizes some other features of our samples. The samples were generally greatly overdispersed, except those with the smallest means. More than 50% of the samples had at least 80% zero counts. All samples were right-skewed (skewness ≥ 0).

Statistical models

For some samples the maximum-likelihood estimate of the parameter of dispersion (θ) of the NB or ZINB model was not numerically convergent: the log-likelihood function continuously increased as θ decreased, and therefore had no maximum. This occurred for 107 samples for the NB model, which all corresponded to underdispersed samples (variance $\leq$ mean). This occurred more frequently (709 samples out of 2258) for the ZINB model, mainly for samples with low mean abundance (the average of mean abundance for these 709 samples was 0.28 individuals/sampling point whereas the general average was 1.12 individuals/sampling point). In these cases, NB and ZINB fits reduced to P and ZIP fits respectively, such that NB and ZINB could never be selected by BIC due to their higher number of parameters.

For some samples the maximum-likelihood estimates of zero inflation ($1-\tau$) of the ZIP or ZINB models were equal to 0 (respectively 105 and 876 cases out of 2258). In this case, ZIP or ZINB reduced to P or NB respectively, such that ZIP or ZINB could never be selected by BIC due to their higher number of parameters.

Comparison of models

The BIC selected the NB as the best model for 58% of the samples. The second best-performing model according to the BIC was the ZIP model (24% of the samples). The P model was selected for 17% and the ZINB model for only 1% of the samples.

The P model tended to underestimate the proportion of zeros and to overestimate low abundances, even for samples for which it was selected (Fig. II.1.1). Overall, the NB model was always better adjusted than (or at least as well as) the P model because the P model is nested into the NB model. The NB residuals were generally centred around zero for all values of observed abundance. However, when the ZIP and ZINB models were selected, the NB model tended to underestimate zeros and overestimate low abundances, while ZIP and ZINB adjusted better to this range of abundances: in particular, the ZIP and ZINB models adjusted the proportion of ones better than P and NB. When the ZINB model was selected, it fitted the proportion of high abundances better than other models.

We excluded samples where the ZINB was selected from our multinomial logistic regressions explaining model selection, because these samples were not numerous enough. When calculating the sample statistic *rel.var*, the log-log regression of sample variance on sample mean had a slope > 1 (slope = 1.65, intercept = 1.90, R^2 = 0.90), confirming the overdispersion of the data. A 'full' multinomial regression model based on our six explanatory variables explained model selection with a performance (kappa statistic) of 79% (Table II.1.4). Excluding each explanatory variable in turn from the full model suggested that the variable with the most important marginal (i.e. unique) effect was *rel.var*, followed by *tot.ab*, *p.zero*, *ssize*, *reach* and finally *species* (Table II.1.4). For example, excluding *rel.var* from the model caused the performance (kappa statistic) to decrease from 79% to 65%, while it decreased by only 1% if *species* or *reach* were the excluded variable, despite the large number of parameters they added to the model as categorical variables. Nevertheless, according to likelihood ratio tests, all six variables had a significant marginal effect (all p-values $<$ 10e-6) on the selection of a distribution model.

The interpretation of the effects of the explanatory variables was complicated by their intercorrelation. Therefore, we used a linear discriminant analysis to illustrate how the four continuous explanatory variables influenced model choice (Fig. II.1.2). Variables *tot.ab* and *p.zero* were quite strongly negatively correlated (correlation coefficient: -0.67) whereas *tot.ab* and *rel.var* were weakly correlated (correlation coefficient: 0.13). The P, ZIP, NB and ZINB distribution groups corresponded to increas-

ing average values of $tot.ab$ (respectively: 6, 12, 65, and 112 individuals/sample). Expectedly, the P distribution group corresponded to the lowest average values of $tot.ab$ and $rel.var$. Situations in which the selected model was NB were more variable than those in which other models were selected (the bagplot corresponding to these situations is wider in Fig. II.1.2).

Figure II.1.3 provides, by species, the observed proportions of samples belonging to the three distribution classes (among P, NB and ZIP, according to the BIC criterion). These proportions clearly changed according to the species : in particular, species such as the perch (Pfl) and pumpkinseed (Lgi) displayed a quite high proportion of the distribution class P, whereas the minnow (Pph) exhibited a quite high proportion of the distribution class NB. Figure II.1.3 also illustrates that the two explanatory variables of model selection with highest marginal effect ($tot.ab$ and $rel.var$) largely explain these differences between species. It provides the average value of exp($tot.ab$) (i.e. the average number of individuals per sample) and $rel.var$ by species. The average $rel.var$ value is much higher for species such as the Schneider (Abi), the nase (Cna), the bleak (Aal), or the minnow (Pph) than for the stone loach (Nba), the pumpkinseed (Lgi) or the chub (Lce).

Consequences of a distributional assumption on mean abundance estimation

The 95% confidence intervals around the mean abundance based on the selected distribution model (CI_P and CI_{NB} for, respectively, P and NB cases) were asymmetric in all cases, with a right side on average 6.8 times wider than the left side. On the contrary, the 95% confidence intervals based on Student's method CI_S were symmetric by construction. The widths of CI_P and CI_{NB} were both positively related to the width of CI_S (Fig. II.1.4). However, CI_P had a width close to but narrower than that of CI_S in 84% of the cases (Fig. II.1.4a). On the contrary, CI_{NB} was wider than CI_S in 92.5% of cases (Fig. II.1.4b), although the left side of CI_{NB} was always narrower than the left side of CI_S. Figure II.1.4b shows that the width of CI_{NB} differed from the width of CI_S in series of cases. Plots of potential explanatory variables on this graph clearly revealed that the number of non-null counts was the main factor of deviation: on average, when non-null counts were ≤ 3, CI_{NB} was 8 times wider than CI_S. The most extreme difference between CI_{NB} and CI_S corresponded to a sample comprising 23 zeros, and only two points with non-null abundance : one with only one individual, and one with 60 individuals. For this sample, CI_{NB} was more than 68 times wider than CI_S.

II.1.4 Discussion

The model most often selected to describe our samples was the NB model. This result is consistent with many other studies that concluded that the negative binomial distribution is suitable for modelling abundance count data of other biological groups (Welsh et al, 1996; Power and Moser, 1999; Gray, 2005). The NB model was selected for a wide range of sample characteristics (Fig. II.1.2), while the other models seemed more relevant for particular cases (e.g. low to very low abundance mean and relative variance for ZIP and P model, or high abundance mean for ZINB). The frequent rejection of ZINB by the BIC criterion, as the least parsimonious model, could partly be expected since we dealt with quite small samples and low total abundance values. Nevertheless, the proportion of zeros predicted by a ZINB fit was generally very close to the proportion predicted by a NB fit, with a maximum difference of 7%. This suggests that the ZINB was not only rejected because of lack of power, but also because the observed proportion of zeros could generally be well explained by the NB model.

Although they also studied freshwater fish species count data sampled with comparable methods as ours, Lewin et al (2009) concluded that zero-modified models such as the ZINB were preferable to the NB distribution model in three German reaches. This apparent contradiction with our conclusions may be due to larger sample sizes in Germany, that favour statistical selection of the ZINB over the NB despite comparable performance (Lewin et al, 2009). In our case, the observed proportion of zeros, even if it was large, was generally quite well explained by models without zero-inflation. This is consistent with the conclusions of Warton (2005) on other count data sets. The strong influence of $tot.ab$ on the selection of a distribution model (Fig. II.1.2) could reflect a certain density-dependence of the patterns of overdispersion and/or zero-inflation. More likely, in our context, it could simply result from seemingly random (Poisson) patterns occurring when mean abundance is particularly low (Taylor et al, 1978; Mante et al, 2005). Samples with few individuals are particularly likely to occur in a context where samples are small and abundance is overdispersed. In such a context, sampling fluctuation can explain a large part of the variability in the selection of a distribution model across samples.

The variability in the selection of a distribution model across samples could also be partly explained by variations in the processes underlying the abundance distribution of populations (e.g. the influence on abundance of habitat characteristics or gregarism). Such an effect was likely to occur because the four distribution models correspond to varying abilities to describe overdispersion or zero-inflation. Consistently, the selection of a distribution model was clearly influenced by the observed

relative variance of counts and proportion of zeros. The explanatory variables with the largest marginal (i.e. unique) effects on model choice were *rel.var* and *tot.ab* (Table II.1.4). However, Figure II.1.2 suggests that among these two factors *tot.ab* was the most influential on the selection of a distribution group: the apparent weak unique effect of *tot.ab* on model choice in Table II.1.4, in comparison to that of *rel.var*, was probably due to the quite strong correlation between *tot.ab* and *p.zero*.

Our study showed that the *reach* factor (which reflects different habitat characteristics) and the *species* factor (which reflects different behaviours) had a weak influence on model selection compared to *tot.ab* and *rel.var*. Still, varying species and reaches corresponded to varying ranges of these factors, and thus to varying proportions of distribution groups (Fig. II.1.3). For example, species such as the Schneider (Abi), the bleak (Aal), the minnow (Pph), and the nase (Can), which are known as particularly gregarious (Muus and Dahlström, 1991) have higher *rel.var* values than others, a property that favours the rejection of P model. In our case, the examination and ecological interpretation of selected distributions by species essentially reduce to the examination of total abundance and relative variance by species. An important implication of this result is that inference based on small samples of abundance counts can reasonably be made with a common NB distribution assumption among species and sites, because the NB distribution can account for varying ranges of mean and variance.

The consequences of selecting an inappropriate distributional assumption for abundance data has been subject to much attention in the ecological literature, in particular for clumped count data. An example of the consequences of an incorrect distributional assumption is the inflated error rates that result from studying abundance data using ordinary linear models, which rely on the hypothesis of a normal distribution (Bartlett, 1947; Venables and Ripley, 1999). Generalized linear models (GLM) enable to carry out analyses of linear models for abundance data, relying on more appropriate distributional assumptions such as the NB distribution (McCullagh and Nelder, 1989). To assess the impact of using a GLM with NB distributional assumption, Power and Moser (1999) carried out a simulation study. Testing the differences in sample means either using a GLM or using t-tests, they showed that the GLM had higher power but greater type I errors. In other words, the GLM was better at detecting differences between means, but could more frequently indicate a difference when there was none. This seems counter-intuitive considering the fact that, in our case, the NB distributional assumption generates larger confidence intervals. However, it could be due to the left side of NB-based confidence intervals being narrower than the left side of Student's confidence intervals.

We calculated confidence intervals around mean abundance as a simpler and more immediate illus-

tration of the consequence of the distributional assumption than the inflated error rates discussed above. We showed that the distributional assumption has profound effects on the confidence interval around the estimate of mean abundance of populations. In practice, selecting either P or NB as a distributional assumption for our data rather than Student's method had severe impacts on confidence intervals for the mean abundance. There was a general correlation between CIs based on P, NB and Student's CIs, because they have a common dependency on the variance and size of the samples and because of their asymptotic convergence towards the same interval bounds. However, our results showed that there was considerable change in the claimed precision of the estimation according to the distributional assumption used to calculate confidence intervals. The NB distributional hypothesis resulted in generally larger confidence intervals than Student's method. On the contrary, the P model, which relies on the hypothesis that variance and mean are equal, often resulted in a lower estimation of uncertainty. In some cases this difference was extreme, with the width of CI_{NB} up to 68 times greater than the width of CI estimated using Student's method. Our results indicated that when relying on the assumption of a NB distribution, the number of non-null counts was the key driver of the uncertainty around the estimate of mean. In practice, samples with less than 3 non-null counts correspond to such a level of uncertainty that estimates should either not be calculated, or at the very least considered with extreme caution. Moreover, our results have been calculated through the use of dispersion parameters uncorrected for bias. Because bias corresponds to an expected underestimation of variance (conditional on the NB model), it is likely that a correction for bias would result in even wider confidence intervals.

The fact that the P and NB model-based confidence intervals are asymmetric reflects that the main uncertainty as for the mean abundance lies in the fact that it could be a a lot higher than its estimate. Indeed, due to the skewness of these distributions and the limited sample size, the probability that no point with high abundance is sampled is high (although the event has non-null probability and would have weighted a lot in the estimation of the mean if it had occurred). This is particularly true for cases where there are very few non-null counts in a sample. More generally, all the methods we used to construct confidence intervals were approximate and, as such, are questionable. To assess the respective accuracy of the different methods, the coverage of both types of CI could for instance be compared to the nominal value (here, 95%) (Kvanli et al, 1998). This study would not be straightforward and will be considered by subsequent work.

Acknowledgements

We would like to thank all the people who helped collecting the point abundance data used here: various teams of the university of Lyon 1, Cemagref, the Aralep engineering office and Henri Persat. Their collection was partly funded by the Compagnie Nationale du Rhône, the Agence de l'Eau Rhône-Méditerrannée-Corse, the Région Rhône-Alpes and other local administrations. We would like to thank René Ecochard, Ton Snelder, Verena Trenkel, Bernard Hugueny and anonymous referees for their advice and corrections, that greatly helped to improve the manuscript.

Table II.1.1: Probability distributions and parameterization of the four models of distribution considered.

Model	Formula	Parameters
P	$\forall x \in \mathbb{N}\ P(X=x) = \frac{\lambda^x}{x!}e^{-\lambda}$	$\lambda \in \mathbb{R}^{*+}$
NB	$\forall x \in \mathbb{N}\ P(X=x) = \frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})}$	$(\mu,\theta) \in \mathbb{R}^{*+} \times \mathbb{R}^{*+}$
ZIP	$P(X=0) = (1-\tau)+\tau e^{-\lambda}$ $\forall x \in \mathbb{N}^*\ P(X=x) = \tau\frac{\lambda^x}{x!}e^{-\lambda}$	$\tau \in [0,1]$ $\lambda \in \mathbb{R}^{*+}$
ZINB	$P(X=0) = (1-\tau)+\tau(1+\mu\theta)^{-\theta^{-1}}$ $\forall x \in \mathbb{N}^*\ P(X=x) = \tau\frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})}$	$\tau \in [0,1]$ $(\mu,\theta) \in \mathbb{R}^{*+} \times \mathbb{R}^{*+}$

Table II.1.2: Number of samples and number of individuals of species involved.

Species scientific name	Code	Common name	Number of samples	Number of individuals
Alburnus alburnus (Linnaeus, 1758)	Aal	Bleak	179	10336
Alburnoides bipunctatus (Bloch, 1782)	Abi	Schneider	143	3370
Barbus barbus (L., 1758)	Bba	Barbel	203	5529
Chondrostoma nasus (L., 1758)	Cna	Nase	211	6178
Gobio gobio gobio (L., 1758)	Ggo	Gudgeon	230	7589
Squalius cephalus (L., 1758)	Lce	Chub	327	16424
Lepomis gibbosus (L., 1758)	Lgi	Pumpkinseed	112	1909
Leuciscus leuciscus (L.,1758)	Lle	Dace	156	4565
Barbatula barbatula (L.,1758)	Nba	Stone loach	193	5953
Perca fluviatilis (L.,1758)	Pfl	Perch	118	921
Phoxinus phoxinus (L.,1758)	Pph	Minnow	173	14141
Rutilus rutilus (L., 1758)	Rru	Roach	214	19754

Table II.1.3: Summary statistics of samples: mean abundance, variance of abundance, proportion of zeros, and skewness.

	Mean	Variance	Proportion of zeros	Skewness
Minimum	0.02	0.02	0.05	0.46
1st quartile	0.24	0.56	0.68	2.26
Median	0.60	2.59	0.80	3.05
3rd quartile	1.56	16.41	0.88	3.96
Maximum	33.40	14644.21	0.99	13.07
Mean	1.52	62.95	0.77	3.30

Table II.1.4: Multinomial logistic regressions, where the probability of selecting a distribution is modeled as a function of six explanatory variables. Regressions were carried using the six variables (full model), or using only five of them (partial models). Goodness of fit statistics (general performance, Cohen's kappa statistic, residual deviance) for each model, and comparisons between each partial model and the full model are displayed. The notation "full model - *variable*" for partial models indicates which variable was excluded.

Model		Goodness of fit				Comparison of full model vs partial model		
		perf.	kappa	dev.	df	Δdev.	Δdf	pvalue
full model		0.88	0.79	1290	4382			
full model	- *tot.ab*	0.81	0.65	1942	4384	652	2	$<$ 1e-15
full model	- *ssize*	0.88	0.78	1377	4384	87	2	$<$ 1e-15
full model	- *rel.var*	0.81	0.65	2034	4384	744	2	$<$ 1e-15
full model	- *p.zero*	0.83	0.70	1561	4384	271	2	$<$ 1e-15
full model	- *species*	0.88	0.78	1326	4404	36	22	3.19e-07
full model	- *reach*	0.87	0.78	1345	4430	55	48	9.04e-07

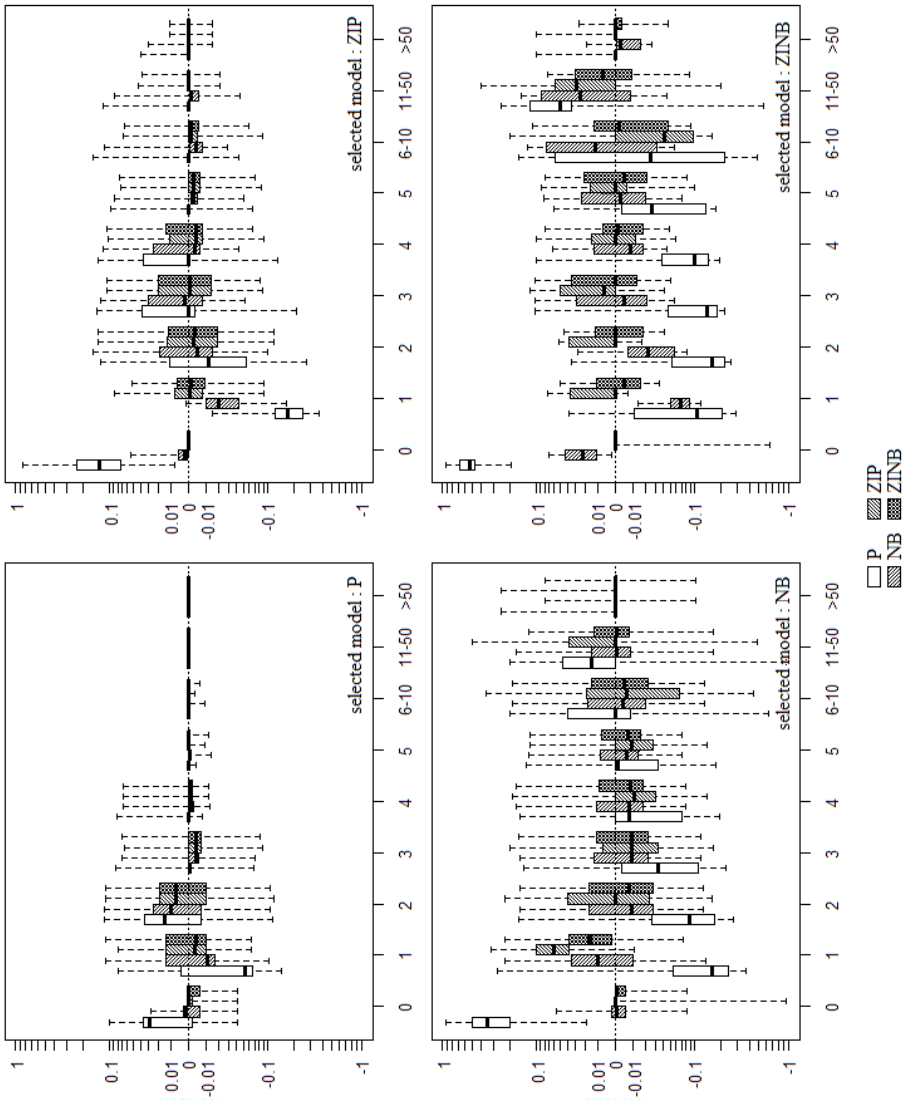


Figure II.1.1: Residuals of the four abundance distribution models, shown by groups of selected distributions: (a): P (b): NB, (c): ZIP and (d): ZINB. Boxes represent 50% of the data, and whiskers extend to the most extreme points. For each selected distribution group, and each abundance, the four boxes represent residuals of the four fitted distributions. For a better readability the results are summarized, for abundances greater than five individuals, by classes of abundance.

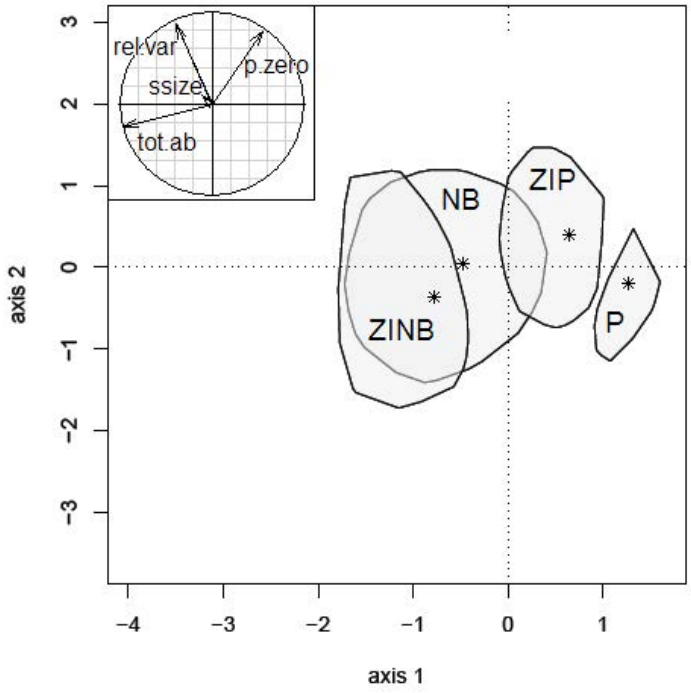


Figure II.1.2: LDA factorial map showing the discrimination of distribution groups as a function of sample characteristics. For each distribution group, bagplots contain 50% of the samples closest to the bivariate median (R software package "aplpack", Rousseeuw et al, 1999).

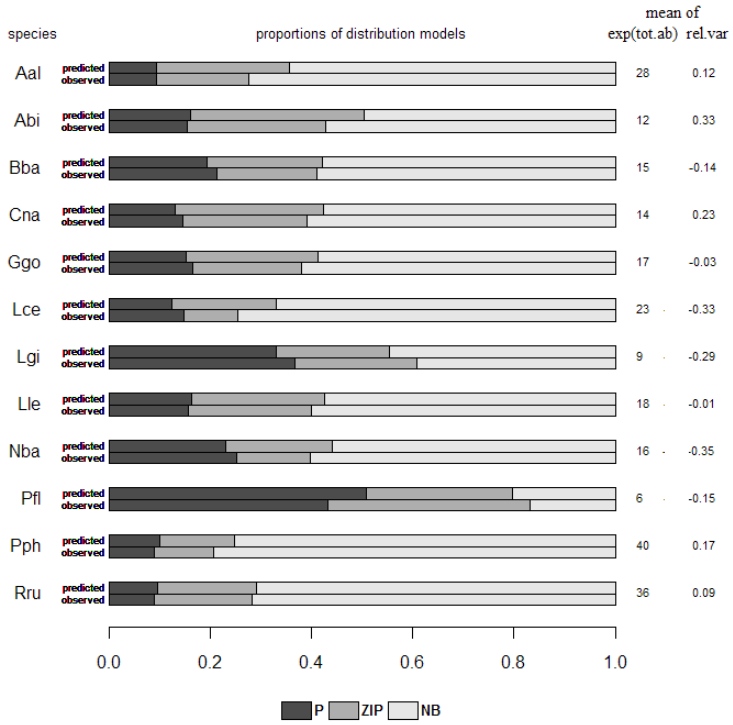


Figure II.1.3: Observed and predicted proportions of samples belonging to a distribution group, by species. The predictions are based on a multinomial regression analysis with $rel.var$ and $tot.ab$ as explanatory variables. For each species the average value of exp($tot.ab$) and $rel.var$ are displayed. Species codes are from Table II.1.2.

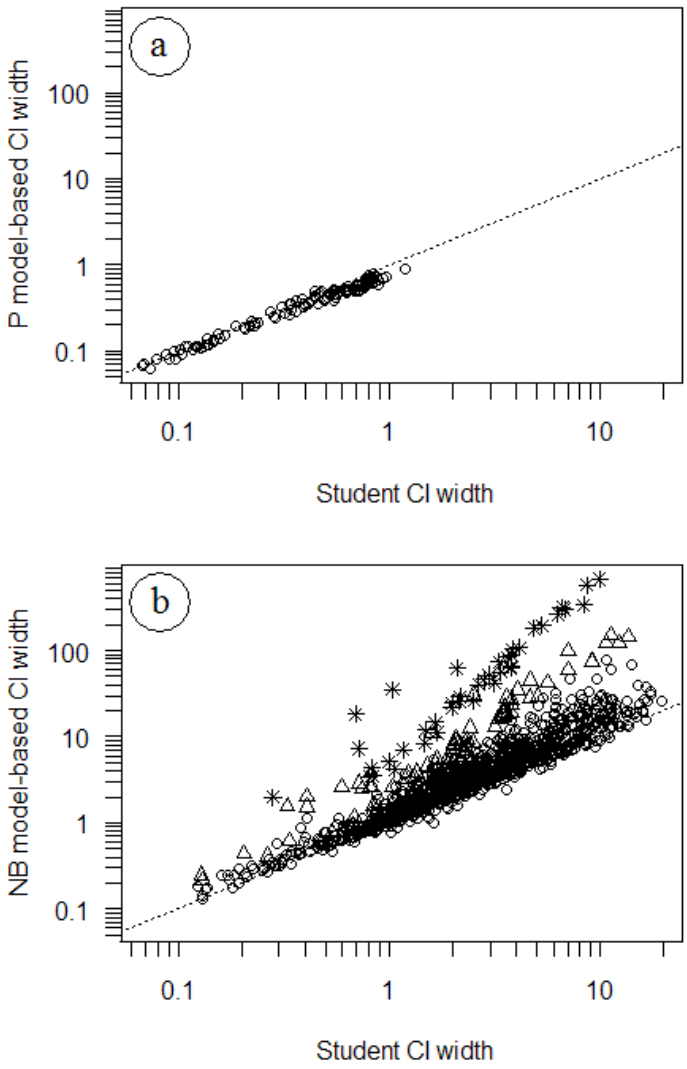


Figure II.1.4: Confidence interval widths around the mean abundance based on a P (a) and NB (b) distributional hypotheses (Y-axis) and on Student's method (X-axis). The dotted line is the x=y equation line. In (b) symbols vary according to the number of non-null counts observed in a sample: stars stand for 2 non-null counts, triangles for 3 and circles for strictly more than 3 (for all samples with only 1 non-null count, ZIP was the model selected).

Chapter 2

Are small samples of overdispersed count data informative on the dispersion of abundance?

Vaudor, L., Lamouroux, N.

Abstract

The negative binomial distribution is frequently used to model overdispersed count data. Adjusting the negative binomial and estimating its parameters is thus a major concern in various disciplinary fields dealing with this type of data such as ecology, epidemiology and actuaries. However, the estimation of the parameters of the negative binomial and especially the estimation of its dispersion parameter is often problematic because available estimators are biased and have high standard deviation. This is particularly true when sample size (i.e. number of counts) is small, when mean abundance is low and variance of abundance is high. In this study, we tried to better specify how and when these estimation problems occur. Specifically, we calculated the bias and standard deviation of the maximum likelihood estimator of dispersion of the negative binomial distribution, conditional on total abundance. This provides practical quantitative guidelines concerning the ability of the observed samples to provide valuable information about the dispersion of abundance.

Keyword: Bayesian inference; Bias; Clumped data; Efficiency; Estimation; Maximum likelihood; Sample size; Sampling effort

II.2.1 Introduction

In all scientific fields dealing with abundance data including ecology, epidemiology, actuaries, sociology and psychology, scientists seek to infer characteristics of the populations they study using count samples of limited size. Count samples are often overdispersed (i.e. sample variance $>$ sample mean), a situation that lead to the common use of the negative binomial (NB) distribution (Bliss and Fisher, 1953). The NB is frequently parameterized as a function of mean abundance and dispersion of abundance, two characteristics of interest when studying populations. The quality of the estimation of both parameters is thus important in a variety of scientific areas.

The quality of an estimator can be assessed through various criteria. An estimator Θ is said to be biased if its mean value $E(\Theta)$ is not equal to the value θ of the parameter being estimated, i.e. $E(\Theta) \neq \theta$. Along with low bias, a low variance or low standard deviation is another desirable property for an estimator Θ. Root mean square error (RMSE) summarizes both bias and standard deviation and is a frequently-used measure of the imprecision of estimators.

The estimation of the parameters of the NB distribution and especially of dispersion is widely recognised as problematic (Pieters et al, 1977; Willson et al, 1984). Existing estimators of dispersion, e.g. based on moments or on maximum-likelihood methods, are biased and have high standard deviation, hence high RMSE. Problems with the estimation of the dispersion of the NB have been studied by many authors who compared the properties of various existing estimators and proposed improved ones (e.g. Pieters et al, 1977; Willson et al, 1984; Clark and Perry, 1989; Anraku and Yanagimoto, 1990; Piegorsch, 1990; Al-Saleh and Al-Batainah, 2003; Saha and Paul, 2005; Lord and Miranda-Moreno, 2008). Many estimators have thus been tested, for varying ranges of parameters values, and for varying sample sizes. Still, the conclusions of these studies hardly lead to general practical recommendations as to which estimator should be used. This is partly due to the fact that the relative properties of the estimators depend on the true values of mean and dispersion and on sample size, and that comparisons have not been carried out for all estimators, in all ranges of values of parameters, and for all sample sizes.

More importantly, in some situations, the problems of bias and high standard deviation remain important whatever estimator is used (Wang, 1996). This suggests that estimation difficulties depend less on the nature of the estimator used than on the characteristics of data themselves. So far, studies on the estimators of the dispersion of count data have mainly focused on providing better estimators

and on assessing these estimators' properties, but did not provide detailed explanation of how sample characteristics were responsible for these estimation problems. A better understanding of how and when estimation problems occur would contribute to identify uninformative data and improve estimations. In particular, it seems quite intuitive that samples with very few individuals are quite uninformative or even misleading as to the dispersion of abundance: in practice, scientists discard such samples from inference. The empirical threshold, in terms of total abundance, between unusable and usable samples is thus a difficult compromise to make. Quantifying the estimation problems conditional on observed total abundance could provide guidelines to decide on such a threshold.

Simulation studies have shown that some particular situations lead to poor estimates of dispersion. In particular, Clark and Perry (1989) and Lloyd-Smith (2007) illustrated the dependence of bias and standard deviation on true parameter values and sample size. They showed that bias and standard deviation are particularly strong when true mean and sample size are low and when true dispersion is high, and that dispersion is generally underestimated. The present paper builds on Clark and Perry (1989) and Lloyd-Smith (2007) but addresses the study of NB estimators by analysing the bias and standard deviation of the dispersion parameter conditional on total abundance (T, which is proportional to the estimated mean for a given sample size). Our use of T as an entry for analysing dispersion estimation issues has several advantages. First, as a simple statistic, T provides a practical clue of the potential quality of the estimation. Second, the set of observable samples for a given value of T is finite, thereby enabling exact computations of bias and standard deviation of the estimator of dispersion. Third, the use of T as condition simplifies the study of the estimator of dispersion because T is a sufficient statistic for the parameter of mean.

II.2.2 Methods

Parameterization of the negative binomial

We considered the extended definition of the NB distribution used by Clark and Perry (1989) and Piegorsch (1990) following Binet (1986). The probability of observing x individuals in a sampling point under a NB(μ,θ) distributional hypothesis is:

$$\begin{aligned} pr_{\mu,\theta}(x) = &\quad \frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})} && \text{if } \theta > 0 \\ pr_{\mu,\theta}(x) = &\quad e^{-\mu}\frac{\mu^x}{x!} && \text{if } \theta = 0 \\ pr_{\mu,\theta}(x) = &\quad \frac{(-\theta^{-1})!}{x!(-\theta^{-1}-x)!}(-\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})} && \text{if } \theta < 0 \end{aligned}$$

where μ is the parameter of mean, θ is the parameter of dispersion, and Γ is the gamma function. In this paper we consider that samples $(x_1,\ldots,x_n)$ are realizations of independent and identically distributed -as a NB(μ,θ)- variables $(X_1,\ldots,X_n)$. The extended definition of the NB allows for null and negative values of θ, which correspond to underdispersed samples $(x_1,\ldots,x_n)$, i.e. samples such that variance $\leq$ mean, that can occur even though the underlying distribution is overdispersed. The extended distribution is based on the continuous transition from the negative binomial distribution for $\theta > 0$, to the Poisson distribution for $\theta = 0$, and to the binomial distribution for $\theta < 0$ (Binet, 1986). However, note that the probability of a sample $(x_1,\ldots,x_n)$ with $\theta < 0$ is defined only for discrete values of θ of the form $-1/n$ where n is an integer and $n \geq max(x_1,\ldots,x_n)$. We will note Θ the maximum likelihood estimator of θ.

Exhaustive list of observable samples and parameter estimates

Usually, to study the properties of Θ, authors carry out simulation studies (e.g. Clark and Perry, 1989; Piegorsch, 1990; Lloyd-Smith, 2007; Park and Lord, 2008). Such methods are time-consuming when involving many random samples, several samples sizes, and many values of (μ,θ). The number of samples to simulate to measure accurately the properties of Θ is high, because Θ itself has a skewed distribution, i.e. it takes extreme values (which weigh a lot in the calculation of its properties) with rare probability. Considering the properties of Θ conditional on T enabled us to reduce computational effort, while providing exact values for these properties. Indeed, the number of observable samples of size n and total abundance $T = t$ is finite and is particularly low for low values t. Listing all observable samples with total abundance t allows taking into account all possible realizations of Θ, even the rare

ones, to calculate the properties of Θ conditional on T. Moreover, such a list allows to calculate the estimate of dispersion for each observable sample only once, rather than calculating an estimate of dispersion for each simulated sample in the case of a simulation study.

We generated an exhaustive list of observable sets of counts S_i (unordered) of size $n = 20$ or $n = 50$, having a total abundance between 0 and 20 ($t = 20$). The choice of sample sizes and maximum total abundance was guided by our experience of a typical large dataset of small abundance counts (freshwater fish; Vaudor et al., submitted) and numerical constraints. We wrote a recursive algorithm to obtain this list of observable sets of counts. Each observable set of counts S_i corresponds to several equiprobable samples $(x_1,\ldots,x_n)$ obtained by permutation. For instance the set of counts $(1,0,0,0,\ldots,0)$ corresponds to 20 samples, each with the '1' in a different position. The probability of each set of counts S_i conditional on T, $pr_{\mu,\theta}(S_i \mid T = t)$, will be calculated as the probability of any of the corresponding ordered samples $(x_1,\ldots,x_n)$ multiplied by the number of possible permutations. A property of the NB distribution is that T is a sufficient statistic for μ, which implies that the probability of a set of counts conditional on T is independent of μ: $pr_{\mu,\theta}(S_i \mid T = t) = pr_{\theta}(S_i \mid T = t)$ (Anraku and Yanagimoto, 1990).

We associated the maximum likelihood estimates $\hat{\mu}_i$ and $\hat{\theta}_i$ of μ and θ to each observable set of counts S_i. In the case of a NB distribution, $\hat{\mu}_i$ actually corresponds to the arithmetic mean t/n (Anraku and Yanagimoto, 1990). For overdispersed sets of counts S_i (i.e. sets such that observed variance $>$ observed mean), we determined $\hat{\theta}_i$ using the function "nlminb" of the statistical software package R (R Development Core Team, 2010), which carries out the maximization of the log-likelihood over the parameter space $[0,+\infty[$. For underdispersed sets of counts S_i, we determined $\hat{\theta}_i$ through the calculation of the likelihood for each value of $\theta > 0$ for which the NB distribution was defined.

Bias, standard deviation and RMSE of the dispersion parameter conditional on total abundance T

To illustrate the properties of the maximum likelihood estimator Θ as functions of the observed value t of T, we first calculated the bias and standard deviation of Θ conditional on T. For a realization t of T, Θ can take n_s values corresponding to the n_s set of counts S_i with total abundance t. Bias and standard

deviation (SD) conditional on T are thus calculated as:

$$\begin{aligned} Bias_{n,\theta}(\Theta \mid T=t) &= E_{n,\theta}(\Theta \mid T=t) - \theta \\ &= \sum_{i=1}^{n_s} \hat{\theta}_i pr_\theta(S_i \mid T=t) \\ SD_{n,\theta}(\Theta \mid T=t) &= \left[var_{n,\theta}(\Theta \mid T=t)\right]^{1/2} \\ &= \left[\sum_{i=1}^{n_s} [\hat{\theta}_i - E_{n,\theta}(\Theta \mid T=t)] pr_\theta(S_i \mid T=t) - \theta\right]^{1/2} \end{aligned}$$

Root mean square error (RMSE) conditional on T is easily derived from the calculated bias and standard deviation through:

$$\begin{aligned} RMSE_{n,\theta}(\Theta \mid T=t) &= \{E_{n,\theta}((\Theta-\theta)^2 \mid T=t)\} \\ &= \left\{[Bias_{n,\theta}(\Theta \mid T=t)]^2 + [SD_{n,\theta}(\Theta \mid T=t)]^2\right\}^{1/2} \end{aligned}$$

We numerically derived and plotted bias and standard deviation conditional on T, for θ values covering the whole range of observable $\theta > 0$, for our 2 values of n, and for a variety of values $t \leq 20$. These calculations were limited to values of $\theta \geq 0$ because the likelihood is not defined for all samples when $\theta < 0$.

Distribution of total abundance T as a function of μ, θ, and n

We calculated the frequency distribution of T as a function of μ, θ and n, because this is an important information for appreciating the frequencies of situations where the bias and standard deviation of Θ conditional on $T = t$ are encountered within the considered range of values for t. This information also allows to appreciate how the estimation of μ interacts with the estimation of θ. Anraku and Yanagimoto (1990) provide an analytical formula for the calculation of $pr_{n,\mu,\theta}(T = t)$. This calculation was made for discrete values of μ consistent with our choice of sample sizes and modelled total abundance ($\mu = 0.25, 0.5, 1, 2$), and for discrete values of θ suggested by the range of observable $\hat{\theta}$ obtained above.

II.2.3 Results

The estimator Θ can take a limited range of values for low values t (Fig. II.2.1). For $t = 20$, Θ can not be more than 86 for $n = 20$, and 222 for $n = 50$. Moreover, for any $t \leq 20$, there are only few possible values of θ in the range of medium to high values of dispersion ($a \geq 5$).

For increasing values of θ, and all considered values t, Θ tends to become more and more biased, first upwards -it overestimates dispersion-, then, to a much greater extent, downwards -it underestimates dispersion- (Fig. II.2.2). The worst upwards bias is for highest values t: in the case $n = 50$ and $t = 20$, the bias reaches a value of ≈ 4 for $\theta \approx 50$ i.e. $E_{n=50,\theta\approx 50}(\Theta \mid T = 20) \approx 54$. The worst downwards bias is for maximum values of θ and lowest values t: in the case $n = 20$ and $t = 2$, the bias reaches a value of ≈ -180 for $\theta \approx 200$, i.e. $E_{n=20,\theta\approx 200}(\Theta \mid T = 2) \approx 20$. The shift between weak overestimation and strong underestimation occurs for increasing threshold values of θ as the value t increases. For a given t, the shift occurs for a higher threshold value of θ for $n = 50$ than for $n = 20$.

Standard deviation increases for increasing values of θ, but only for values of θ below a certain threshold whose value depends on n and t (Fig. II.2.3), ranging from $\theta \approx 15$ to $\theta \approx 70$ for $n = 20$, and $\theta \approx 30$ to $\theta \approx 200$ for $n = 50$. For instance, for $n = 20$ and $t = 20$, this threshold corresponds to very high values of θ ($\theta > 70$), while it is much lower for $n = 20$ and $t = 2$ ($\theta > 15$). Correspondingly, below the thresholds, the higher the value of t, the lower the standard deviation of Θ, whereas the trend is inversed above these thresholds. Above these thresholds, standard deviation starts decreasing again for increasing values of θ, although bias is still really important: this reflects the fact that estimates of dispersion lie more and more narrowly around a highly biased, low value of $E(\Theta)$.

RMSE increases for increasing values of θ (Fig. II.2.4). In general, higher values t correspond to lower RMSE of Θ, except for some ranges of values θ: between 15 and 60 ($n = 20$) and between 20 and 200 ($n = 50$), the trend can be inversed, with higher values t corresponding to slightly higher values of the RMSE. The probability of observing weak values t is particularly strong for high values of θ, even when μ is quite high (Fig. II.2.5). This trend is particularly strong for $n = 20$. For instance if $\mu = 0.25$, $\theta = 20$ and $n = 20$ (i.e. $E(T) = 5$), $pr(T = 0) = 17\%$ and $pr(T \leq 5) = 67\%$. If $\mu = 2$, $\theta = 20$ and $n = 20$ (i.e. $E(T) = 40$), $pr(T = 0) = 2.5\%$ and $pr(T \leq 5) = 14\%$.

II.2.4 Discussion

Previous studies have highlighted the fact that some conditions such as small-sized samples, low mean and high dispersion were responsible for poor estimation of θ (see for instance: Clark and Perry, 1989; Park and Lord, 2008). However, the true values of mean and dispersion are unknown, and thus it is difficult to assess the quality of the estimate provided by the data at hand. One classical method for assessing this would be the calculation of confidence intervals. Still, as there are no satisfactory estimators of θ for such samples, there are a fortiori no easy and unquestionable ways to calculate confidence intervals for θ.

In practice though, some samples may be discarded from inference because they are thought to be obviously uninformative. For instance, one does not expect to be able to estimate correctly the dispersion of a population, unless the data comprises at least a certain number of individuals. Hence, discarding samples with values of T below a certain threshold is a rule of thumb that is probably beneficial in terms of quality estimation, but might be costly in terms of data loss. In the large fish dataset that motivated this study (2258 repeated samples of maximum 180 abundance counts -between 20 and 50 counts for 88% of the samples- for 12 species; Vaudor et al., submitted), 19% of the samples have values of $t \leq 5$, 36% have values of $t \leq 10$, and 55% have values of $t \leq 20$. For this dataset, the frequent occurrence of samples with very few individuals is to a large extent linked to multispecific sampling, which yields data about rarer species, although sampling design (and, in particular, sample size) is mainly designed for more common species.

Our study measures the effect of the statistic T on estimation quality, quantifying the bias, standard deviation, and RMSE of the maximum likelihood estimator of θ conditional on T. It thus provides a measure of the precision of the estimation of a according to the data at hand. It also provides guidelines to define the threshold below which samples should be discarded according to the level of precision required for the estimation of θ.

If θ lies in very high ranges, it might actually be quite irrelevant to estimate θ at all, considering the very bad properties of Θ in these ranges (whichever the value t) (Fig. II.2.2, II.2.3, II.2.4). This is actually equivalent to stating that it is unreasonable to try to estimate the dispersion of rare populations (precisely, with rare occurrence) with too small samples. For $n = 20$ for instance, if $\theta > 10$ the bias and standard deviation are particularly important. Besides, higher values t do not necessarily correspond to an improvement in estimation quality; higher values t might even correspond to less precise estimation

(e.g., higher RMSE). In this range of values θ, the poor quality of estimation can not be improved through discarding samples with low values t. For $n = 50$, this problem occurs for higher ranges of θ, i.e. $\theta > 20$ instead of $\theta > 10$.

In the example of the large fish dataset of Vaudor et al. (submitted), the range of values θ is of course difficult to assess, because it can only be assessed through the observed range of values θ, which we know tend to be quite imprecise and biased. In particular, 23.3% of the samples with $n = 20$ (i.e. 68 samples out of 280) are such that $\hat{\theta} \geq 10$, and 23.1% of the samples with $n = 50$ (i.e. 45 samples out of 195) are such that $\hat{\theta} \geq 20$. Still, this is not in contradiction with the hypothesis that values of θ are generally $\leq 10 (n = 20)$ or $\leq 20 (n = 50)$, considering the high standard deviation of Θ.

In ranges of values θ that are not too high considering sample sizes at hand, the quality of estimation is all the better that t is high. For $n = 20$ and under the hypothesis $\theta \leq 5$, for instance, the RMSE is ≤ 10.2 for t in 2-20, whereas it is ≤ 5.5 for t in 10-20. Hypothesizing that θ might be higher, for instance $\theta \leq 10$, high values t would correspond to a smaller gain in precision: RMSE = 11.7 for t in 2-20, and RMSE ≤ 11.3 for t in 10-20. For $n = 50$ and under the hypothesis $\theta \leq 5$, the RMSE is ≤ 19.2 for t in 2-20 and ≤ 5.0 for t in 10-20. Hypothesizing that θ might be higher, for instance $\theta \leq 10$, high values t correspond to a smaller but still substantial gain in precision: RMSE ≤ 23.8 for t in 2-20, and RMSE ≤ 10.1 for t in 10-20.

So far, existing simulation studies about the properties of the estimator of dispersion generally calculated estimates for all simulated samples for which it was possible, including some uninformative samples. Our study allows to put into perspective these previous simulation results: it highlights the fact that they might be pessimistic as to the reliability of estimation of dispersion on fixed-size samples, because they are partly based on really uninformative samples that would, in practice, be discarded from inference.

Low mean, high dispersion and small sample sizes lead to frequent occurrence of low values t (Fig. II.2.5- II.2.6), which correspond to particularly poor estimation (Fig. II.2.2 - II.2.4). This is consistent with other studies that showed the joint effect of these three factors on estimation quality (e.g. Clark and Perry, 1989). It should be noted that low values t occur frequently in case the dispersion is high, even though the mean is not particularly low. In other words, although the maximum likelihood estimator of mean T/n is unbiased, under high dispersion conditions its distribution is particularly skewed, i.e. the mean is most of the time underestimated, and sometimes highly overestimated. This reflects that, when dispersion is high, a bad estimate of θ often goes together with a bad estimate of μ.

Our results illustrate the fact that the notion of information is at the heart of the problems of estimation. Samples comprising too few individuals bring too little information about the mean and the dispersion of the populations. At best, they provide imprecise estimates (for instance when both mean and dispersion are low). At worst, they provide misleading estimates (for instance when only few individuals have been sampled, although mean and dispersion are actually high). Information, however, is hard to define and quantify in a unique way. Though we showed the influence of T on the properties of estimators, other statistics than T could also reflect how informative a sample is likely to be and would deserve further focus. For example, the number of non-null counts is another feature of samples which, intuitively, may suggest that estimation is more or less reliable. In particular, very few non-null counts correspond to strong values $\hat{\theta}$, which suggest that dispersion is probably quite high, and maybe too high in respect of sample sizes at hand. T and number of non-null counts, although related to some extent, are far from being equivalent: for instance, a sample comprising only zeros except one high abundance has a low number of non-null counts but a high t. Such a sample is also quite uninformative, as it is likely to provide an overestimate of mean and variance. Such a sample, however, occurs less frequently than samples with low values t that lead to generally underestimate both mean and dispersion.

Obviously, the higher the sampling effort, the less likely it is to observe uninformative samples. When dispersion is low, increasing the sampling effort will rapidly increase observed values t. When dispersion is high, in contrast, the sampling effort might have to be increased by a huge amount before causing a substantial increase in t. Under high dispersion conditions, luck plays an important role: an increase in t relies mainly on the unlikely occurrence of a non-null abundance. Whatever the sample size is, if it is fixed a priori, there is a risk that few or even no individuals are sampled. To avoid this, it is actually sensible, rather than fixing a sample size a priori, to carry out inverse or sequential sampling (i.e. to go on sampling while the sample does not verify a certain condition, for instance t superior to a certain threshold value). Willson et al (1984) proposed a sequential sampling plan to estimate the dispersion of the NB and showed that such a sampling method improved the estimation of dispersion compared to fixed-sample size plans. They used the methods of moments to calculate an estimate of $1/\theta$ and "continued to add observations to the sample until the estimate of $[1/\theta]$ began to converge in some specified manner". Their stopping criterion was that the difference between the two last estimates of $1/\theta$ was below 0.05, which is actually not a guarantee for convergence of $1/\theta$. Actually, the gain in estimation quality going with the use of inverse or sequential sampling could be even more important if inverse sampling was based on other stopping criterions. Our results suggest that a stopping criterion

based on t could be tested for improving inverse sampling strategies.

In a fixed-sample size context, the bias and standard deviation variations of estimators suggest to rely as much as possible on multiple sources of information. Former studies' results, field work, and general knowledge about the populations of interest might be taken into account to complete the information brought by the data, and to limit errors of estimation. In a Bayesian framework, this corresponds to proposing a prior probability distribution for the values of the parameters, such that estimation of the populations' characteristics based on the observed sample can not provide estimates in total contradiction with prior knowledge. Al-Saleh and Al-Batainah (2003) and Lord and Miranda-Moreno (2008) showed for instance that Bayesian estimators of $1/\theta$, with prior distributions favouring smaller values of $1/\hat{\theta}$ (i.e. high estimates of dispersion), could improve the reliability of the estimates. The choice of a prior on distributions' parameters is of course always arduous. Here, it is particularly difficult because prior knowledge of θ is likely to be biased itself.

It is probably easier for scientists to a priori characterize the population they study in terms of mean, than in terms of dispersion, because mean corresponds to a much more intuitive quantity, and because the estimator of mean is unbiased. Interestingly, specifying a prior for the estimation of μ does not only influence the value of the estimate $\hat{\mu}$ but also the value of $\hat{\theta}$. For instance, observing few individuals might imply two different things: either that mean abundance is really low, or that dispersion is really high. The maximum likelihood estimator of μ is the arithmetic mean of the observed sample. The maximum likelihood method thus always concludes that samples with low t result from a population of low mean abundance (and thus underestimates dispersion), while samples with low t could as well result from a population with higher mean abundance but high dispersion. Interestingly, the method of Conditional Maximum Likelihood (with the likelihood conditioned on T), used by Anraku and Yanagimoto (1990) and Robinson and Smyth (2008) are to some extent related to this idea, stating that problems with classical estimators of the dispersion of NB counts are due to the fact that "they fail to adjust for the fact that the mean is estimated from the same data". Our study provides support for the need to further explore of the properties of these estimators.

Acknowledgements

We thank Georg Heinze for his interest and advice, and Verena Trenkel for commenting on this paper.

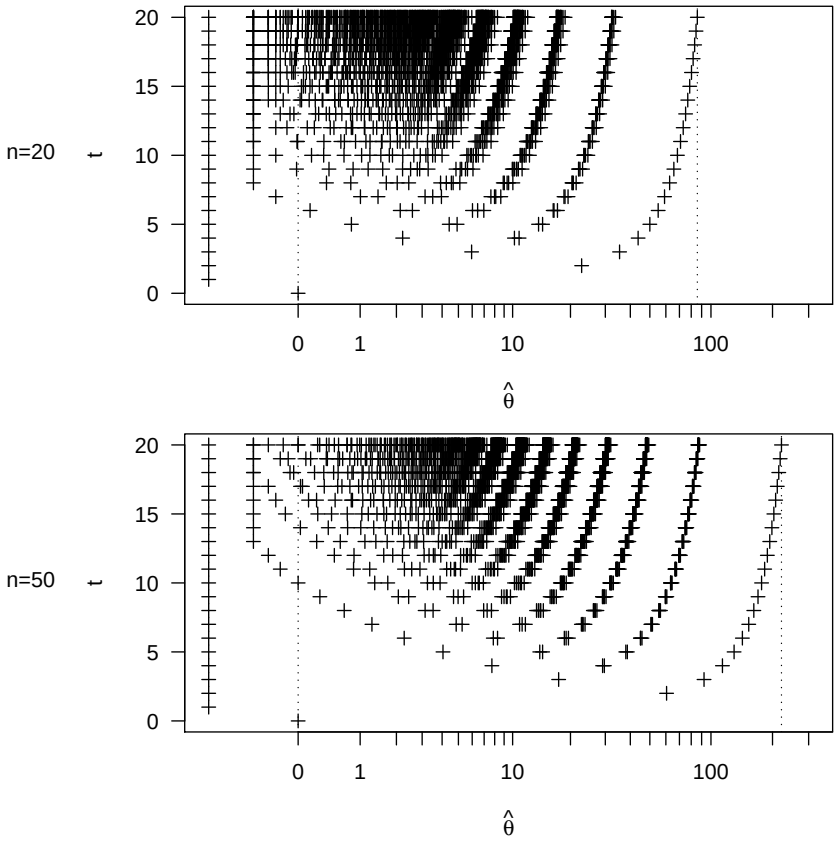


Figure II.2.1: Observable values of $\hat{\theta}$ according to the observed value t of T ($t \leq 20$), for $n = 20$ and $n = 50$. Note the use of a logarithmic scale for positive values of $\hat{\theta}$ on the x-axis.

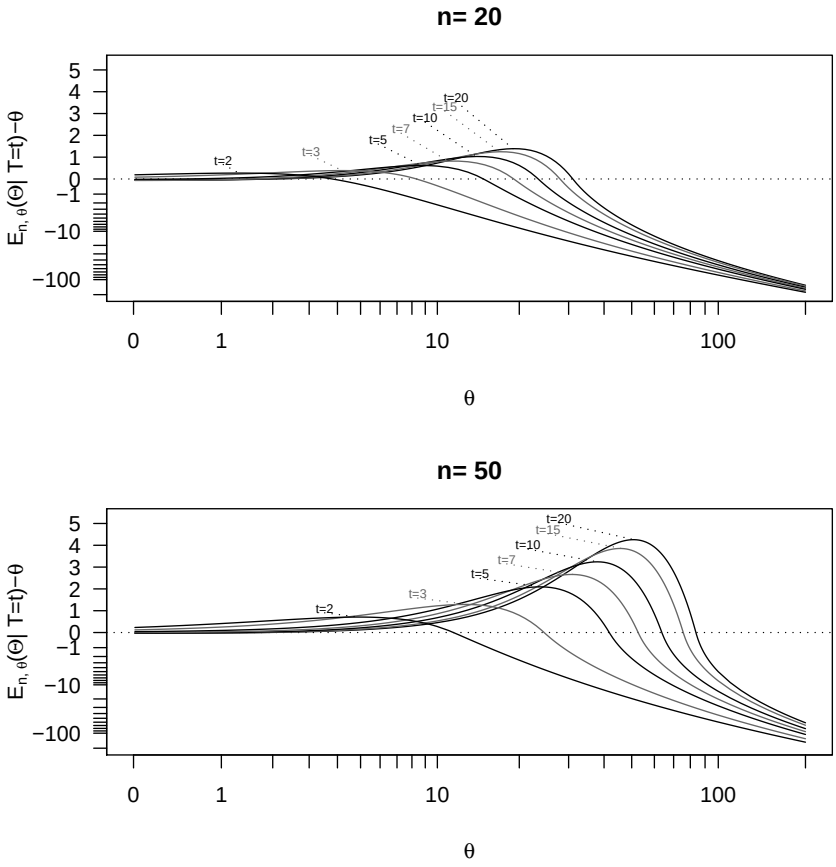


Figure II.2.2: Bias of the maximum likelihood estimator Θ, conditional on T, as a function of n and θ. Note the use of a logarithmic scale for the x-axis. The two dotted lines have the equation $y = 0$.

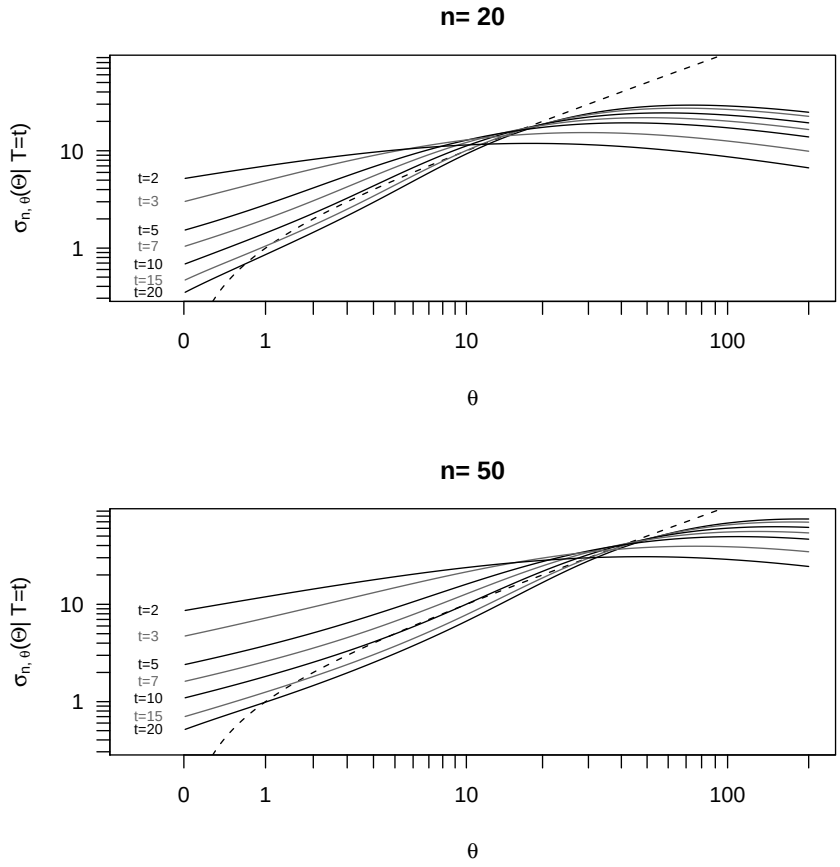


Figure II.2.3: Standard deviation of the maximum likelihood estimator Θ conditional on T, as a function of n and θ. Note the use of a logarithmic scale for the x-axis. The two dashed curves have the equation $x = y$.

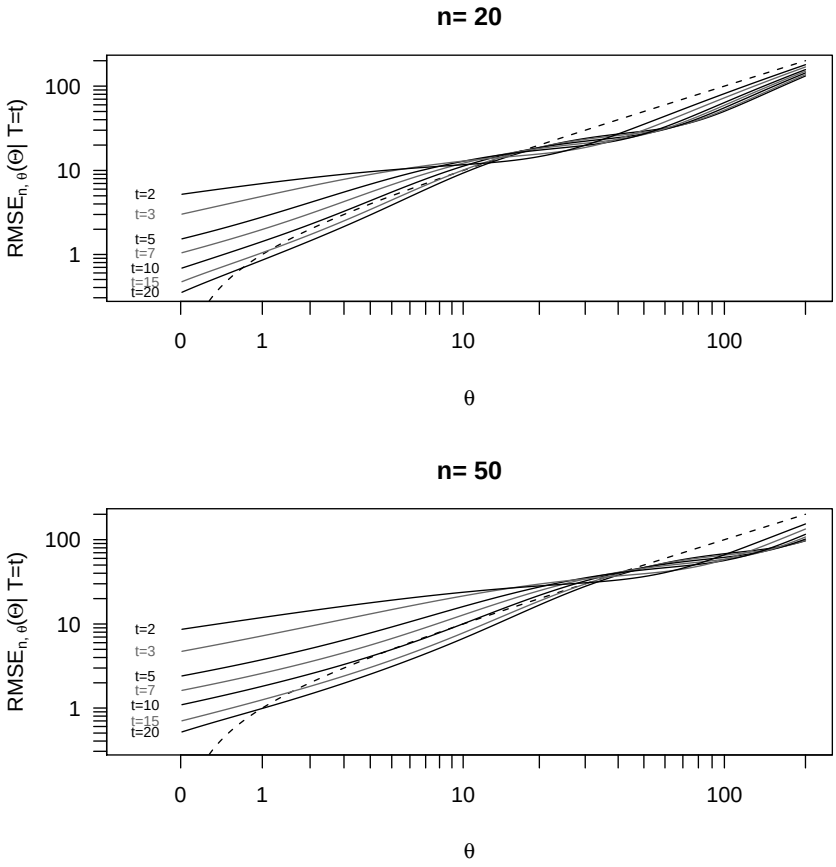


Figure II.2.4: RMSE of the maximum likelihood estimator Θ conditional on T, as a function of n and θ. Note the use of a logarithmic scale for the x-axis. The two dashed curves have the equation $x = y$.

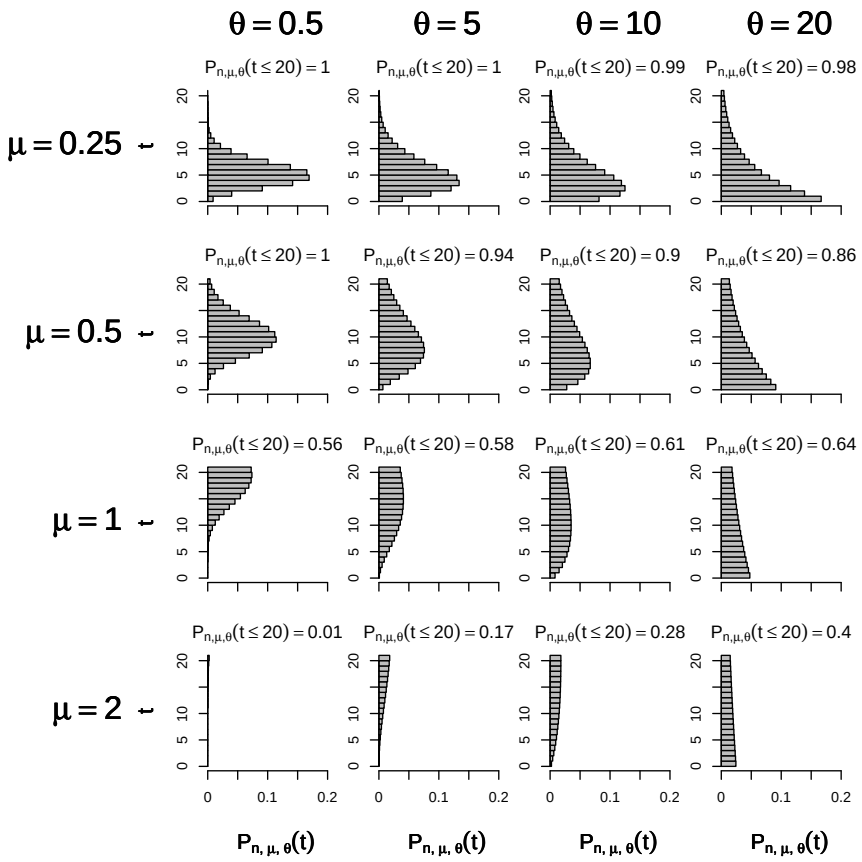


Figure II.2.5: Probability of occurrence of $T = t$, as a function of μ and θ, and for sample size $n = 20$.

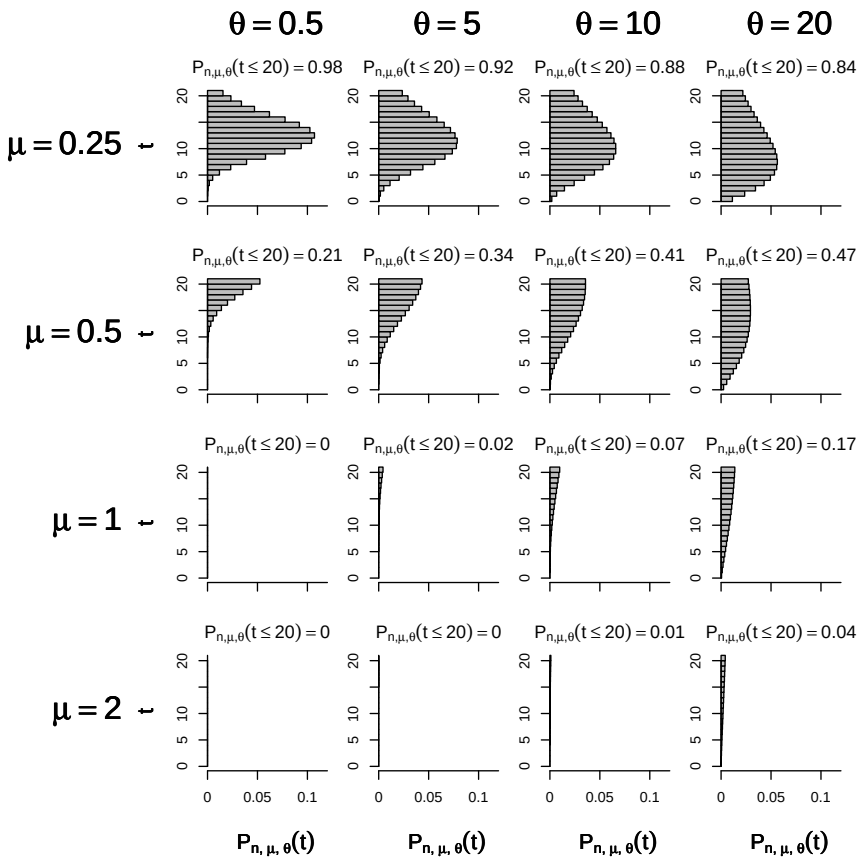


Figure II.2.6: Probability of occurrence of $T = t$, as a function of μ and θ, and for sample size $n = 50$.

Chapter 3

Confidence intervals for the mean abundance: which method is best suited to small samples of overdispersed count data?

Vaudor, L., Ecochard, R.

Abstract

The calculation of a confidence interval for the mean of abundance is a basic way to account for the uncertainty concerning the estimate of mean. The methods used for deriving confidence intervals might be problematic, in particular when the studied populations have an overdispersed abundance distribution. In this case, conventional ways to calculate confidence intervals often underestimate uncertainty in the estimate of mean. In this simulation study, we compared four methods used to calculate confidence intervals for the mean by examining the coverage probability, the average length and other properties of the calculated confidence intervals. We carried out our simulations under the hypotheses that abundance was distributed as a negative binomial with low mean (between 0.2 and 5), and high dispersion (between 2 and 30), and that samples had small sizes (20 to 50 counts). We showed that confidence intervals based on likelihood (e.g. profile likelihood intervals) provide good results, in particular when dispersion is high.

Keywords: Negative binomial; Frequency distribution; Student; Bernstein; Profile likelihood; Skewness

II.3.1 Introduction

Estimating the mean abundance of a population is one of the most basic and yet important problems in statistics, with applications in many fields, e.g. ecology, finance, actuaries, epidemiology. Random sampling is probably the most classical way of sampling populations in order to study their mean abundance. But while the arithmetic mean of random samples is seen as a reasonable and quite unquestionable estimate of population mean abundance, assessing the uncertainty of this estimate might be more challenging. Confidence intervals (CIs) are widely used to assess this uncertainty in an easily interpretable way (Newcombe, 1998; Di Stefano et al, 2005). Still, in some contexts, interval estimation itself is particularly challenging, and CIs provide results that dramatically differ depending on the method used to calculate them, and on the modelled mean-variance structure in particular.

Overdispersion, or Poisson overdispersion (i.e. the fact that variance $>$ mean, in contradiction to the mean=variance hypothesis underlying a Poisson distribution model) is a really common feature of abundance count data (Taylor, 1984). The negative binomial (NB) distribution is a Gamma-Poisson mixture, with a mean-variance structure such that variance $>$ mean. It is widely recognised as appropriate for modelling overdispersed count data (Anscombe, 1949; Bliss and Fisher, 1953; Taylor et al, 1979; Johnson et al, 1992). A possible parameterization of the negative binomial (NB) distribution for a random variable X is

$$\forall x \in \mathbb{N} \quad pr(X = x) = \frac{\Gamma(x+\theta^{-1})}{x!\Gamma(\theta^{-1})}(\mu\theta)^x(1+\mu\theta)^{-(x+\theta^{-1})} \tag{3.1}$$

where μ is the parameter of mean ($\mu > 0$) i.e. $E(X) = \mu$, θ is the parameter of dispersion ($\theta > 0$), and Γ is the Gamma function. The mean-variance structure of the NB distribution is $var(X) = \mu + \mu^2\theta$ i.e. $var(X) > E(X)$.

A very classical way of calculating a $100(1-\alpha)\%$ CI for a parameter μ consists in inverting a Wald test of the hypothesis $\{\hat{\mu} = \mu_0\}$. The test rejects the hypothesis when the Wald statistic $(\hat{\mu} - \mu_0)/se(\hat{\mu})$ (where $se(\hat{\mu})$ is the standard error of the maximum likelihood estimate) exceeds in absolute value the $(1-\alpha)/2$ quantile of a normal distribution N(0,1) -e.g. 1.96, in the case $\alpha = 0.05$-. Wald intervals, though, are known to be problematic in several aspects (Cook and Weisberg, 1990; Meeker and Escobar, 1995; Fears et al, 1996; Pawitan, 2000; Brown et al, 2003) and in particular in terms of coverage probability (the probability that the interval contains the true value of the parameter). Cook and Weisberg (1990) note that "there is a long history of dissatisfaction with Wald intervals, primarily because they [...] may not accurately reflect the actual, often asymmetric, uncertainty in an estimate.

In addition they can have true coverage rates that are far from the nominal values". Along with Wald-type CIs, Student-type CIs are another very classical way of estimating the mean parameter (Boyles, 2008; Rosenblum and Van der Laan, 2008), and their shape is similar to that of Wald CIs since they also derive from the hypothesis that the distribution of $\hat{\mu}$ around μ_0 is Gaussian. Still, they have the advantage of being a little bit more conservative, adjusting to the fact that generally both mean and variance are estimated from the same set of data.

Problems with the validity of Wald or Student CIs occur when the hypotheses underlying their construction are not verified. These hypotheses are either that the distribution of the data is Gaussian, (which is not the case in our context), or that the sample size is “large enough" to approximate the distribution of $\hat{\mu}$ by a Gaussian distribution (through the Central Limit Theorem). In this latter case, the Wald or Student CIs rely on asymptotical results. The sample size necessary to satisfactorily approach asymptotical results depends on the nature of the methods considered, as well as on the nature of the data. In particular, the convergence of $\hat{\mu}$ towards a Gaussian distribution is quite fast when the distribution of the data itself is close to a Gaussian distribution. On the contrary, when the distribution of the data is really different from a Gaussian (for instance when it is highly overdispersed), the convergence of $\hat{\mu}$ towards a Gaussian distribution is particularly slow (Shilane et al, 2008). For instance, for a sample size of 20, under a NB distributional hypothesis with a quite high mean abundance ($\mu = 5$) and a high dispersion ($\theta = 10$), Shilane et al (2008) showed that the distribution of $\hat{\mu}$ was right-skewed, with about 5% of the observed $\hat{\mu} > 11.0$ and about 5% of the observed $\hat{\mu} < 1.2$. Shilane et al (2008) also showed that despite a nominal coverage value of 95%, Wald-type CIs for the mean of $n = 20$ data points (respectively, $n = 50$) from a $NB(\mu = 5, \theta = 10)$ had a coverage probability of 75% (respectively, 84%) only. This confirmed that the validity of such CIs for quite small samples of overdispersed count data should be questioned and compared to alternative methods.

Along with Wald or Student methods, the inversion of a likelihood ratio test is another quite classical way of estimating CIs. These intervals, called profile likelihood intervals or likelihood ratio intervals, have, in particular, gained favour among statisticians. In practice, though, they are still less commonly used than Wald-type intervals: Meeker and Escobar (1995) suggest that this is because asymptotically, these two methods are equivalent, such that “when the focus is on theory, students are often left with the mistaken impression that the Wald and likelihood approaches provide equally accurate approximations".

Besides, some recent studies have proposed non-asymptotic methods (which will be described in

the Methods section) to calculate confidence intervals for the mean of finite samples. Rosenblum and Van der Laan (2008) and Shilane et al (2008), in particular, suggested that a confidence interval based on Bernstein inequalities provided good results in terms of coverage probability. Brown et al (2003) also worked on a NB and showed that a confidence interval based on a Jeffreys equal-tailed test had good coverage probability compared to a standard Wald confidence interval.

In this simulation study, we compared four methods to construct CIs for the mean of small samples of data from a NB distribution. Two methods were based on asymptotical results: Student and profile likelihood CIs. Two were non-asymptotical methods: Bernstein CIs, as defined in Rosenblum and Van der Laan (2008) and Shilane et al (2008); and Jeffreys CIs, as defined by Brown et al (2003), and Cai (2005). This comparison was based on several criteria (e.g. coverage probability, balance between left and right non-coverage), following the recommendations of Vos and Hudson (2005) and Swift (2009).

II.3.2 Methods

Our simulation ranges (sample size and parameter values) were motivated by an available extended data base of repeated freshwater fish counts (Vaudor et al. submitted), comprising more than 2200 small samples of overdispersed data. Sample sizes were between 20 and 180, but 88% comprised less than 50 counts. Samples were also characterized by low observed mean (1st quantile=0.24, median=0.60 and 3rd quantile=1.52) and high observed dispersion (1st quantile=3.14, median=6.66 and 3rd quantile=13.90). For these samples, the Wald and profile likelihood methods were shown to give notably different results (Vaudor et al. submitted).

For each couple of parameters (μ,θ) with μ in (0.2, 0.5, 1, 2, 5) and θ in (2, 5, 10, 15, 30), we simulated 1000 samples $(x_1,x_2,\ldots,x_n)$, of size $n=20$ and $n=50$, as realizations of the independent and identically distributed (i.i.d.) random variables $(X_1,X_2,\ldots,X_n)$, following a distribution $NB(\mu,\theta)$ -i.e. all distributed as X, cf equation (3.1)-. For each simulated sample, we estimated the mean and constructed a 95% CI, according to each of the methods considered. We then assessed some properties (detailed hereafter) of the various methods considered for constructing a CI. We used the software package R (R Development Core Team, 2010) for all simulations and calculations.

Let $\bar{x}$ and s^2 be the arithmetic mean and standard deviation of the sample $(x_1,x_2,\ldots,x_n)$:

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ s &= \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{1/2} \end{aligned}$$

Let $\hat{\mu}$ and $\hat{\theta}$ be the maximum likelihood (ML) estimates of μ and θ : $\hat{\mu}$ and $\hat{\theta}$ are thus estimates based on the NB distribution model. In the particular case of the parameter μ of the NB distribution, the ML estimate $\hat{\mu}$ is actually equal to the arithmetic mean of $(x_1,x_2,\ldots,x_n)$, i.e. $\bar{x}$. The variable $\left(\sum_{i=1}^{n} X_i\right)/n$ is thus the ML estimator of the mean μ. As for $\hat{\theta}$, it was calculated through the use of the function "nlminb" of the statistical software package R (R Development Core Team, 2010), that carries out the maximization of the log-likelihood of the samples, using a Newton-type algorithm, over the parameter space $]0,+\infty[$.

Methods for constructing confidence intervals

In the following paragraphs we detail the methods for constructing CIs with a nominal level (or confidence level) of $1-\alpha$. Note that all our results are for the particular value $\alpha = 5\%$.

Student confidence interval

The method for calculating the Student confidence interval for the mean relies on the fact that $(\bar{X} - \mu)/(s/\sqrt{(n)})$ asymptotically follows a Student distribution with $n-1$ degrees of freedom.

$$CI_S = \left[\hat{\mu} + q_{S,\alpha/2,n-1}\frac{s}{\sqrt{n}}, \hat{\mu} + q_{S,1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] \tag{3.2}$$

where $q_{S,\alpha/2,n-1}$ and $q_{S,1-\alpha/2,n-1}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the Student distribution with $n-1$ degrees of freedom. For large values of n, for instance $n = 1000$, $|q_{S,\alpha/2,n-1}| = |q_{S,1-\alpha/2,n-1}| \approx 1.96$ for $\alpha = 0.05$.

The Student confidence interval is quite similar to the Wald confidence interval,

$$CI_W = \left[\hat{\mu} + q_{N,\alpha/2}se(\hat{\mu}), \hat{\mu} + q_{N,1-\alpha/2}se(\hat{\mu})\right]$$

where $se(\hat{\mu})$ is an estimate of the standard error of $\hat{\mu}$, and $q_{N,\alpha/2}$ and $q_{N,1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the Normal distribution ($N(0,1)$ ($|q_{N,\alpha/2}| = |q_{N,1-\alpha/2}| \approx 1.96$ for $\alpha = 0.05$). A classical estimate for the standard error $se(\hat{\mu})$ in the context of Wald intervals is $I(\mu)^{-1/2}$ where $I(\mu)$ is the Fisher information about the parameter μ. In accordance with the mean-variance relationship for the NB, another possible estimate of $se(\hat{\mu})$ is $\left(\hat{\mu} + \hat{\mu}^2\hat{\theta}\right)^{1/2}/\sqrt{n}$ (Brown et al, 2003; Cai, 2005). In an even simpler way, as we are dealing with the parameter of mean, a possible estimate of $se(\hat{\mu})$ is $s/\sqrt{n}$ (Rosenblum and Van der Laan, 2008; Shilane et al, 2008, e.g.).

In this study, we preferred using the Student CI rather than a Wald CI for two reasons: the first one is its simplicity, as it does not require the calculation of $I(\mu)^{-1/2}$. The second reason is that, compared to Wald CIs with estimate of $s/\sqrt{n}$, the Student CI adjusts to the fact that mean and variance are estimated using the same data.

Besides, the Student confidence interval for the mean is readily available in many statistical packages, for instance through the use of the function t.test in R (R Development Core Team, 2010).

Bernstein confidence interval

The method for calculating the Bernstein confidence interval for the mean relies neither on a distributional hypothesis for X, nor on asymptotical results. It relies on the two following hypotheses (Bennett, 1962; Rosenblum and Van der Laan, 2008; Shilane et al, 2008):

1. The random variable X is bounded, i.e. there exists a constant W such that $pr(|X - E(X)| \leq W) = 1$.

2. The variance of the variable X is bounded, i.e. there exists a constant S^2_{max} such that $var(X) < S^2_{max}$.

Under these two hypotheses, the Berstein inequality states that for any $\varepsilon > 0$,

$$pr(|\bar{X} - E(X)| > \varepsilon) \leq 2\exp\left(-\frac{1}{2}\frac{n\varepsilon^2}{S^2_{max} + \varepsilon W/3}\right) \tag{3.3}$$

In our case, $E(X) = \mu$, and the Bernstein CI for μ is

$$CI_B = [\hat{\mu} - \hat{\varepsilon}, \hat{\mu} + \hat{\varepsilon}] \tag{3.4}$$

where $\hat{\varepsilon}$ is the positive value that satisfies the second degree equation obtained through equating the right side of (3.3) with α.

$$\hat{\varepsilon} = \frac{log(2/\alpha)}{n}\left(\frac{1}{3}W + \sqrt{\frac{1}{9}W^2 + 2nS^2_{max}(log(2/\alpha))^{-1}}\right) \tag{3.5}$$

The construction of the Bernstein confidence interval requires W and S^2_{max} to be estimated. Rosenblum and Van der Laan (2008) and Shilane et al (2008), although they underline the need and complexity of accurately estimating them, suggest the use of the simple following estimates w and s^2_{max} of W and S^2_{max} respectively:

$$\begin{aligned} w &= \frac{n+1}{n}\max(x) \\ s^2_{max} &= s^2 \end{aligned}$$

The R code (R Development Core Team, 2010) for this method of calculation is available in Rosenblum and Van der Laan (2008).

As noted by Rosenblum and Van der Laan (2008) and Shilane et al (2008) themselves, the calculation of Bernstein CIs rely on the assumption (1) which is not necessarily valid as a NB variable has

by definition no upper bound. Besides, they acknowledged the fact that their proposed estimates w and s^2_{max} are questionable, as they rely on the supposition that the observed variance s^2 is the maximum observable variance s^2_{max}, and the observed maximum $\max(x)$ is really close to the maximum observable value of X, w. Nevertheless, Shilane et al (2008, 2010) calculated Bernstein CIs in the case of unbounded NB variables, considering that the hypothesis (1) was approximately verified. They thus showed that Bernstein CIs could be interesting in terms of coverage probability. Another CI construction method based on Bernstein's inequality has been recently detailed in Shilane et al (2010). It does not hypothesize that X has an upper bound, but the results it provides, in terms of coverage probability at least, are not as good as those of the Bernstein's method with the hypothesis of an upper bound.

Profile likelihood confidence interval

The method for calculating the profile likelihood confidence interval for the mean $\hat{\mu}$ relies on a distributional hypothesis for X (here a NB distributional assumption), and on asymptotical results. Its construction requires that the ML estimate of dispersion, $\hat{\theta}$, is calculated. The profile likelihood interval is obtained through the inversion of the likelihood ratio test which accepts the null hypothesis $H_t : \{\mu = \mu_t\}$ with significance level $\alpha = 5\%$. Let Λ be the likelihood ratio:

$$\Lambda = \frac{L(\mu, \hat{\theta})}{L(\hat{\mu}, \hat{\theta})}$$

$-2\log\Lambda$ asymptotically follows a $\chi^2(1)$ distribution. The CI comprises all values μ_t such that H_t is not rejected by the likelihood ratio test. The hypothesis H_t is not rejected as long as $-2\log\Lambda$ does not exceed the $1-\alpha$ quantile of the $\chi^2(1)$ distribution, $q_{\chi,1-\alpha}$. The bounds of CI_L are the values μ_t which minimize $\left|2log\Lambda - q_{\chi,1-\alpha}\right|$.

The function plkhci in package Bhat, in R (R Development Core Team, 2010) provides numerical solutions for these, using a modified Newton-Raphson algorithm -for more details see Venzon and Moolgavkar (1988)-.

Jeffreys confidence interval

The method for calculating the Jeffreys CI for the mean relies on the hypothesis that X is distributed as a NB. Its construction requires that an estimate of θ is calculated. Here, we used the ML estimate $\hat{\theta}$. The NB can also be parameterized as a function of (p, θ) with $p = (1+\mu\theta)^{-1}$. The Jeffreys prior (Jeffreys, 1946) for the parameter p is proportional to $(1-p)^{-1/2}p^{-1}$ (Johnson et al, 1995; Brown et al, 2003;

Cai, 2005). Its conjugate posterior distribution is a $Beta(n\theta^{-1}, n\bar{x}+0.5)$ (Fink, 1997). According to this posterior distribution, a Jeffreys CI for p has lower and upper bounds equal to the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Beta(n\theta^{-1}, n\bar{x}+0.5)$ distribution, $q_{B,\alpha/2}$ and $q_{B,1-\alpha/2}$. As $\mu = \theta^{-1}(1-p)/p$, relying on an estimate of θ (here, the ML estimate $\hat{\theta}$), we considered the corresponding Jeffreys CI for μ:

$$CI_J = \left[\frac{1}{\theta}\frac{1-q_{B,1-\alpha/2}}{q_{B,1-\alpha/2}}, \frac{1}{\theta}\frac{1-q_{B,\alpha/2}}{q_{B,\alpha/2}}\right] \tag{3.6}$$

Properties of the confidence intervals

Coverage probability

The coverage probability (called CP hereafter) of a CI is the actual probability that the calculated interval contains the value of the parameter. CIs aim to contain the value of an unknown parameter with a given probability $1-\alpha$ (nominal CP).

Expected length

Another desirable property of CIs is that they make optimal use of the information carried by the data. Typically, a CI is more informative if narrow: a desirable property of the methods for constructing CIs is that they lead to, in average, narrow CIs. Here, we assessed the median of the lengths of CIs (called MedL hereafter) rather than their mean in order to limit the influence of a few really wide CIs. Considering the fact that $\mu \geq 0$, we truncated CI_S and CI_B at 0 in case their lower bound was < 0, prior to calculating MedL.

Balance of left and right non-coverage

The CI has a probability of 1-CP not to contain the true value of the parameter. Let P_O be the probability the true value of the parameter is below the lower bound of the CI -i.e. the whole CI is an overestimate- and P_U be the probability that the true value of the parameter is above the upper bound of the CI -i.e. the whole CI is an underestimate-. We define an index of balance between left and right coverage (Bal) as:

$$Bal = \frac{P_u}{P_u + P_o} = \frac{P_u}{1-CP} \tag{3.7}$$

A value of Bal greater than 0.5 means that confidence intervals, when they fail to contain the true value of the parameter, are generally underestimates. Note that Bal is only defined when $CP \neq 1$.

Negative lower bound

Some methods for constructing confidence intervals for the mean allow for negative lower bounds. In this case, the lower bound of the confidence interval is totally uninformative because the parameter of mean is, of course, positive. The lower bounds of CI_S and CI_B might be negative in some cases, since they are based on a Gaussian approximation of the distribution of $\hat{\mu}$ around μ. In contrast, CI_L and CI_J always have positive lower bounds by construction (Shilane et al, 2008).

II.3.3 Results

Figure II.3.1 illustrates the characteristics of the CIs for the particular values $n = 20$, $\mu = 1$, and all tested values of θ. CI_S and CI_B frequently have negative lower bounds (here truncated at 0), while CI_J and CI_L always have a positive lower bound, and generally have a wide (and even extremely wide in some cases) right side. All methods tend to underestimate the mean more frequently than they overestimate it (the mean is often above and rarely below the constructed CIs), in particular for CI_S and CI_B.

Whichever method is used to construct CIs, the CPs are lower for higher values of dispersion θ (Fig. II.3.2). For all values (μ, θ), CI_S is the worst-performing in terms of coverage probability: in particular, for high values of mean and dispersion, the coverage probability is much lower than nominal value (for instance, it is 62% instead of 95‘% for $n = 20$, $\mu = 5$ and $\theta = 30$). In contrast, the coverage probability of CI_B is sometimes superior to the nominal value, in particular for low values (μ, θ) of mean and dispersion. There is at most a difference of 2% between the coverage probabilities of CI_L and that of CI_J, and both are generally closer to nominal value than the coverage probability of CI_B. In the few cases where the coverage probability of CI_B is closer to nominal value than the CP of CI_J and CI_L, it differs from CI_J and CI_L by at most 6%. Conversely, the coverage probabilities of CI_J and CI_L might be better than that of CI_B by 15% at most (for $n = 20$, $\mu = 5$ and $\theta = 30$). For all methods and for all cases except $\mu = 0.2$, CP is higher for $n = 50$ than for $n = 20$. It generally corresponds to a improvement of CP (i.e. to a decrease of the discrepancy between CP and nominal value), except for CI_B when CP for $n = 20$ is already above nominal value. For $\mu = 0.2$, for all methods except the Student method, the increase in sample size unexpectedly corresponds to a decrease of CP.

For all values (μ, θ), CI_S is, on average, the narrowest CI (Fig. II.3.3). CI_B is generally the second narrowest CI on average, except for low values ($n = 20$) or low to medium values ($n = 50$) of dispersion. In the case $n = 20$, CI_L and CI_J are generally longer than CI_B and CI_S for all values of $\theta \geq 5$, and the difference in median length is particularly important for high values of mean and dispersion. For instance, for $n = 20$, $\mu = 5$ and $\theta = 30$, the median length of CI_J is about 30 times that of CI_S. In the case $n = 50$, the differences in median length are less important and only affect highly dispersed data ($\theta \geq 15$). However, for $n = 50$, $\mu = 5$ and $\theta = 30$, the median length of CI_J is still about 5 times that of CI_S.

The fact that CI_L and CI_J tend to be quite wide, in particular for high values of dispersion, is actually

due to their right side being really wide in some cases (Fig. II.4.1). Across all tested parameter sets (μ, θ) and for $n = 20$, samples such that the upper bound of CI_L and CI_J is at least 100 times greater than arithmetic mean (i.e. 5.7% and 10.5% of the simulated samples, respectively) have mean values of $\hat{\theta}$ of 60.2 and 47.1 respectively, whereas the general mean is 11.2 Across all tested parameter sets (μ, θ) and for $n = 20$, samples such that the upper bound of CI_L and CI_J is at least 10 times greater than arithmetic mean (i.e. 25% and 32% of the simulated samples, respectively) have mean values of $\hat{\theta}$ of 30.8 and 26.4 respectively, whereas the general mean is 11.2.

When CI_B and CI_S fail to comprise the true value of μ, they are always or almost always underestimates, i.e. μ is above the upper bound of the interval (Fig. II.3.4). No value for CI_B, $n = 20$ and $\mu = 0.2$ is indicated because in this case CI_B never fails to comprise μ. For all values of $\mu \geq 0.5$ the proportion of underestimates and overestimates is generally more balanced (and even more so for $n = 50$) for CI_L and CI_J than for CI_B and CI_S. Still, they also tend to underestimate μ more frequently than they overestimate it (i.e. Bal is closer to 0.5, but it is still generally ≥ 0.5). In contrast, for the lowest value of mean ($\mu = 0.2$) and $n = 20$, CI_L and CI_J are always overestimates.

CI_L and CI_J, by construction, have positive lower bounds so that the value of NLB for these methods is always 0. CI_B and CI_S, on the other hand, frequently have negative values as lower bounds (Fig. II.3.5). For $n = 20$, the proportion of cases in which CI_B has a positive lower bound is only 20% at best (i.e. for high mean and low dispersion values). For $n = 50$, except in case the dispersion is low ($\theta = 2$), the proportion of cases in which CI_B has a positive lower bound is only of 32% at best. CI_S frequently has a negative lower bound, too, in particular for low values of mean and high values of dispersion. For instance, for $n = 20$, $\mu = 1$ and $\theta = 30$, its lower bound is negative in 99% of the cases.

II.3.4 Discussion

If all assumptions used to calculate a CI were exactly met, the actual CP would be equal to the nominal CP. In practice though, these assumptions may be unverified. Some authors might require that CP is superior or equal to nominal value for all possible sets of parameter values. In this case, CIs are said to be exact or conservative (Swift, 2009). Exact CIs tend to be particularly wide, which might in some cases limit their practical interest (Brown et al, 2001). In our case, in particular, making sure that CP is not below nominal level for any set of parameter values would be particularly difficult, and would require setting lower and upper bounds on possible parameter values (μ, θ) as Rosenblum and Van der Laan (2008) did. Here, we did not consider exact CIs, and considered instead that a desirable property of CIs was that their coverage probability was as close as possible to (and possibly lower than) nominal CP. Obviously, small values of MedL tend to correspond to low values of CP, thus they are only desirable in case CP is at least equal to nominal level. If CP is below nominal value, then small values of MedL are not particularly appealing, as they at least partly correspond to the CIs being too narrow to account for the actual uncertainty about the mean.

Depending on the context, authors can be more concerned that the calculated CIs might be overestimates of the mean or, on the contrary, that they might be underestimates. Nevertheless, the use of a two-sided CI suggests that attention is paid to both upper and lower confidence limits for the parameter of interest, and that a desirable property of the CI is that both limits are equally informative. Hence, the CI should have the same probability (1-CP)/2 to overestimate the parameter, and to underestimate it. In other words, left and right non-coverage should be balanced (Newcombe, 1998; Swift, 2009), i.e. Bal should be as close as possible to 0.5. Similarly, negative lower bounds for the CIs are totally informative, as the probability to be above the true value of the parameter is zero. Consequently, a desirable property of CIs construction methods is that the proportion of CIs with negative lower bound (NLB) is as close to 0 as possible (Newcombe, 1998; Swift, 2009).

Overall, CI_J and CI_L should be preferred to CI_S and CI_B when calculating confidence intervals for the mean of small samples of overdispersed count data. They show better properties, in particular when dispersion θ is quite strong: their coverage probability is closer to nominal level than that of CI_S, their balance between right and left non-coverage is better, and they have more informative lower bounds than both CI_S and CI_B.

Our study illustrates the importance of basing the comparison of CIs on several criteria, as under-

lined by Newcombe (1998),Vos and Hudson (2005), and Boyles (2008). Basing this comparison on coverage only, or on coverage and average length, one could indeed conclude that CI_B is an appropriate method for constructing confidence intervals for small samples of overdispersed abundance data (Shilane et al, 2008). Here, we showed that although for certain ranges of (μ, θ) CI_B is better than other methods in terms of coverage probability, it is actually mainly linked to the lower bound of CI_B being really uninformative (such that the chance of underestimating μ is actually null). Rosenblum and Van der Laan (2008) and Shilane et al (2008) all underlined the fact that correctly estimating w and s with the Bernstein method might be difficult, and that sensitivity analysis should be carried out to assess the effect of these estimates on the estimation of CI_B. In particular, the estimators of w and s we used in this study are quite rough and they are likely to underestimate w and s, as highlighted by Rosenblum and Van der Laan (2008). Still, more accurate estimates of w and s could not really correct the main drawbacks of CI_B i.e. unbalanced right and left non-coverage, and frequent occurrence of negative lower bounds, because they would only modify the width of CI_B without modifying its shape. Indeed, it seems that the major flaw in CI_B is that, like CI_S, it is symmetric around the mean, whereas there is a certain asymmetry in the uncertainty of the mean estimate.

The present study confirms that likelihood-based methods for constructing CIs for mean abundance are better suited to overdispersed abundance data. In particular, following the recommendations of Cook and Weisberg (1990); Pawitan (2000); Boyles (2008), it is easily seen from the asymmetry of CI_L in Figure II.3.1 that the shape of the profile likelihood (with θ as a nuisance parameter) is not quadratic around the observed mean μ, and thus that CIs hypothesizing a Gaussian distribution of $\hat{\mu}$ around μ are not appropriate. Actually, the slope of profile likelihood, for a sample of size n with observed mean $\hat{\mu}$, and with a $NB(\mu, \alpha)$ distribution hypothesis, can easily be calculated and is equal to: $n(\hat{\mu}/\mu - 1)/(1 + \mu\hat{\theta})$. Thus, the slope is obviously not symmetric around $\hat{\mu}$: it tends to be lower for high values of μ (i.e., in particular, on the right of $\hat{\mu}$), especially when $\hat{\theta}$ is high. This is consistent with very wide CI_L occurring when $\hat{\theta}$ is particularly high (i.e. when the sample at hand suggests that the distribution is particularly overdispersed).

It might be seen as problematic, in practice, that CIs are extremely wide. Instances where the upper bounds of CI_L and CI_J are more than 100 times higher than observed mean $\hat{\mu}$ are not rare, in particular for $n = 20$: they represent 5.7 and 10.5%, respectively, of all simulated samples. In such cases, CI_L and CI_J barely provide information as to an upper confidence limit on the mean. Nevertheless, this lack of information is probably not due to the method of inference but due to the data itself: CI_S and CI_B tend

to be narrower not because they convey more information than CI_L and CI_J, but because they fail to reflect that the data is hardly informative.

More worrying are the cases where all types of CIs, comprising likelihood-based CIs, fail to account for the high uncertainty as for the estimate of mean. These cases are responsible for low values of CP, for CI_S, CI_L and CI_J in particular. As underlined by Rosenblum and Van der Laan (2008), there are two reasons why the actual CP of CI_S is lower than nominal value: 1. the assumption that the distribution of $\hat{\mu}$ around μ is close to a Gaussian is not appropriate, and 2. the variance (here, s^2) is often largely underestimated. Although CI_L and CI_J are not subject to the first problem, they are still subject to the second (although they do not rely on s^2, but on the estimate of dispersion $\hat{\theta}$ instead). Indeed, the estimators of θ (including the maximum-likelihood estimator) are known to be biased downwards (i.e., on average, they underestimate θ), in particular when μ and n are low, and θ is high (Saha and Paul, 2005, Vaudor and Lamouroux, in prep). Such a bias in the estimation of θ is thus particularly likely to occur in our case, and consistently the lowest CP values for CI_L and CI_J correspond to the lowest values of μ and n and the highest values of θ. In this respect, the fact that CIs for μ exhibit poor coverage is linked to the problems inherent to treating θ as a nuisance parameter, i.e. as a known constant $\theta = \hat{\theta}$. Confidence regions, as generalizations of confidence intervals to several parameters, could thus constitute a really interesting tool (though more complex, in terms in construction and calculation, than CIs) to account for the uncertainty in the estimation of both μ and θ in the case of small samples from overdispersed populations.

Although CI_J comes from Bayesian inference, and CI_L from frequentist inference, CI_J and CI_L are quite similar in terms of CP, MedL, Bal, and NLB, probably due to the common dependence of both methods on 1. the log-likelihood of samples under a NB distributional assumption, and 2. the estimate of its dispersion parameter $\hat{\theta}$. Besides, we chose to use the same estimator of θ for both methods (the ML estimator). Due to both its simple calculation, and slightly better coverage and balance between left and right non-coverage (cf. Figures II.3.1 and II.3.3), CI_J should actually be preferred to CI_L to calculate CIs for the mean of small samples from populations with overdispersed distributions.

CI_J is based on a Bayesian posterior probability for μ. Naturally, when using CI_J the corresponding estimate of μ should be the value maximizing this posterior likelihood, $\hat{\mu}_J$, rather than the classical maximum-likelihood estimate $\hat{\mu}$. The mode of the posterior distribution of $\mu\theta$ is $(n\hat{\mu} - 0.5)/(n\theta^{-1} + 1)$. Considering θ as a known constant (for instance $\theta = \hat{\theta}$), we have: $\hat{\mu}_J = (n\hat{\mu} - 0.5)/(n + \hat{\theta})$. Whatever the values of n, $\hat{\mu}$ and $\hat{\theta}$, $\hat{\mu}_J < \hat{\mu}$. The higher the dispersion θ, the weaker $\hat{\mu}$ and n, the

greater the relative difference: $(\hat{\mu} - \hat{\mu}_J)/\hat{\mu}$. However, the mode of the posterior distribution of μ is overall always quite close to the arithmetic mean, such that in our simulations the arithmetic mean always lied inside the bounds of CI_J. Besides, as noted by Brown et al (2001), the use of Jeffreys priors (Jeffreys, 1946) has good frequentist properties (Wasserman, 1991).

Acknowledgements

We would like to thank Nicolas Lamouroux for advice and careful reading of a first version of this draft.

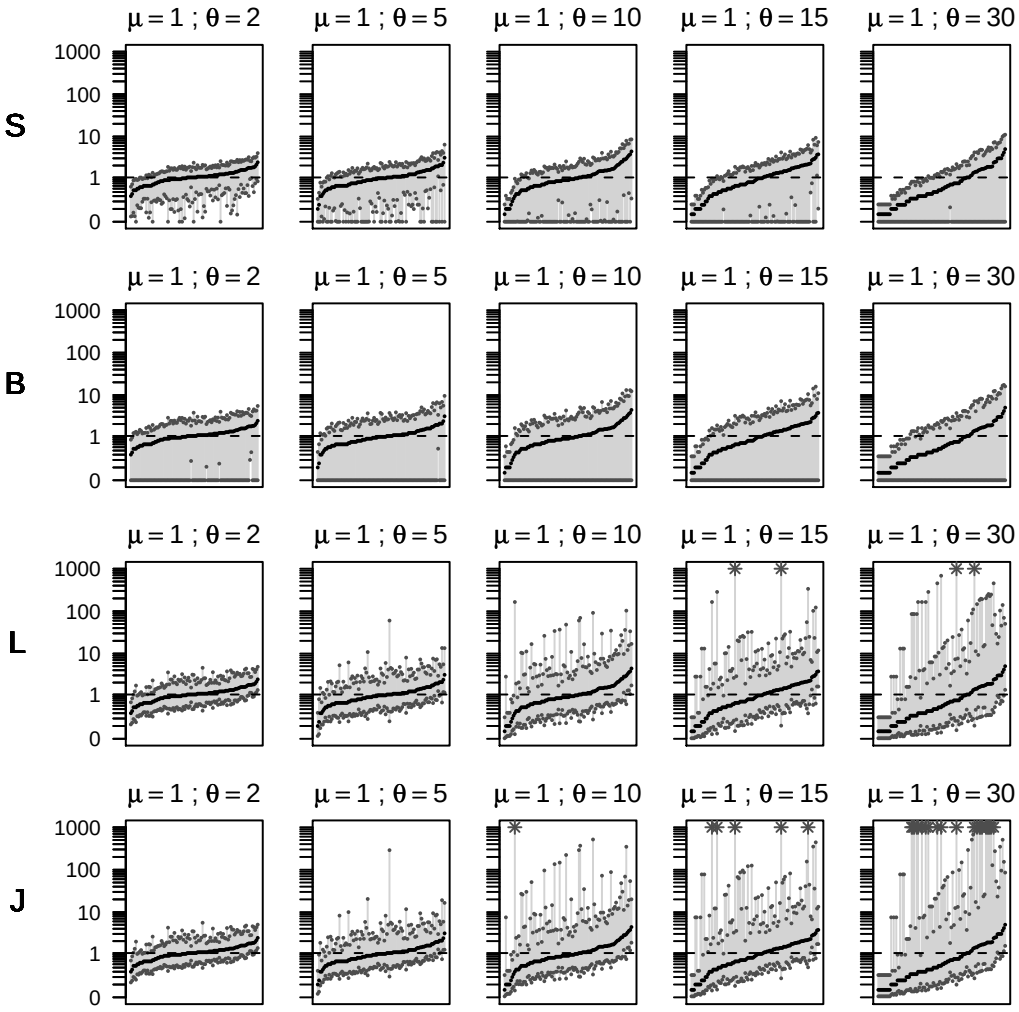


Figure II.3.1: CIs for the mean of simulated samples, for the four tested methods, in the case $\mu = 1$ and $n = 20$ (S: Student, B: Bernstein, L: profile likelihood, J: Jeffreys). The dark grey points stand for the bounds of the CIs, and the black points stand for the sample means. The horizontal dashed line stands for the true value of μ. For better readability, results are displayed for only 100 samples (randomly selected) out of the 1000 simulated samples, samples are ordered according to the sample means, and the upper and lower bounds of the CIs are represented by asterisks when they are above 1000.

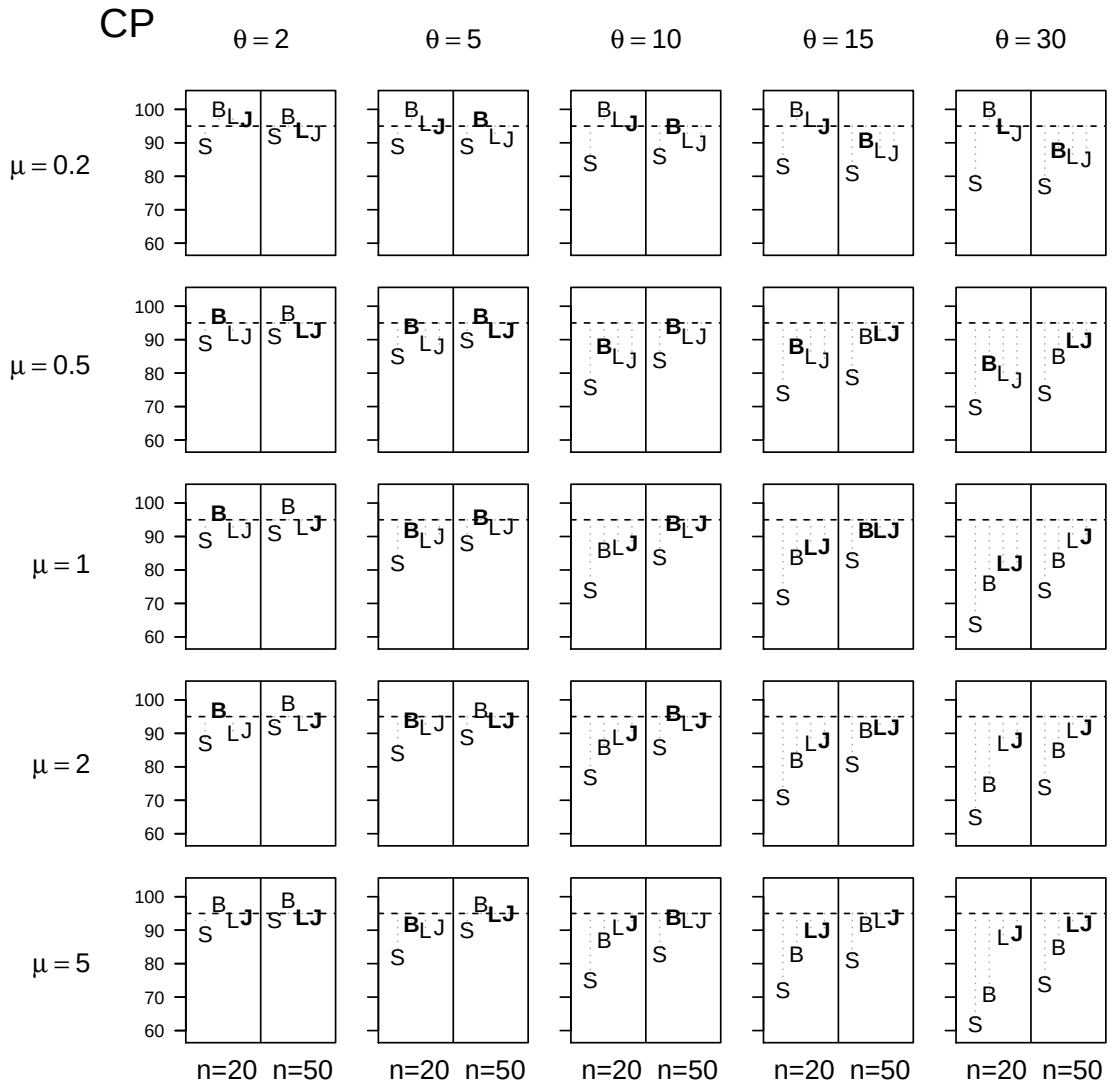


Figure II.3.2: Coverage probability (CP), of the different methods for constructing CIs (%), according to the parameter values (μ, θ) (S: Student, B: Bernstein, L: profile likelihood, J: Jeffreys). The horizontal line stands for the nominal value of CI i.e. 95%. The bold font indicates the methods with CP closest to nominal value.

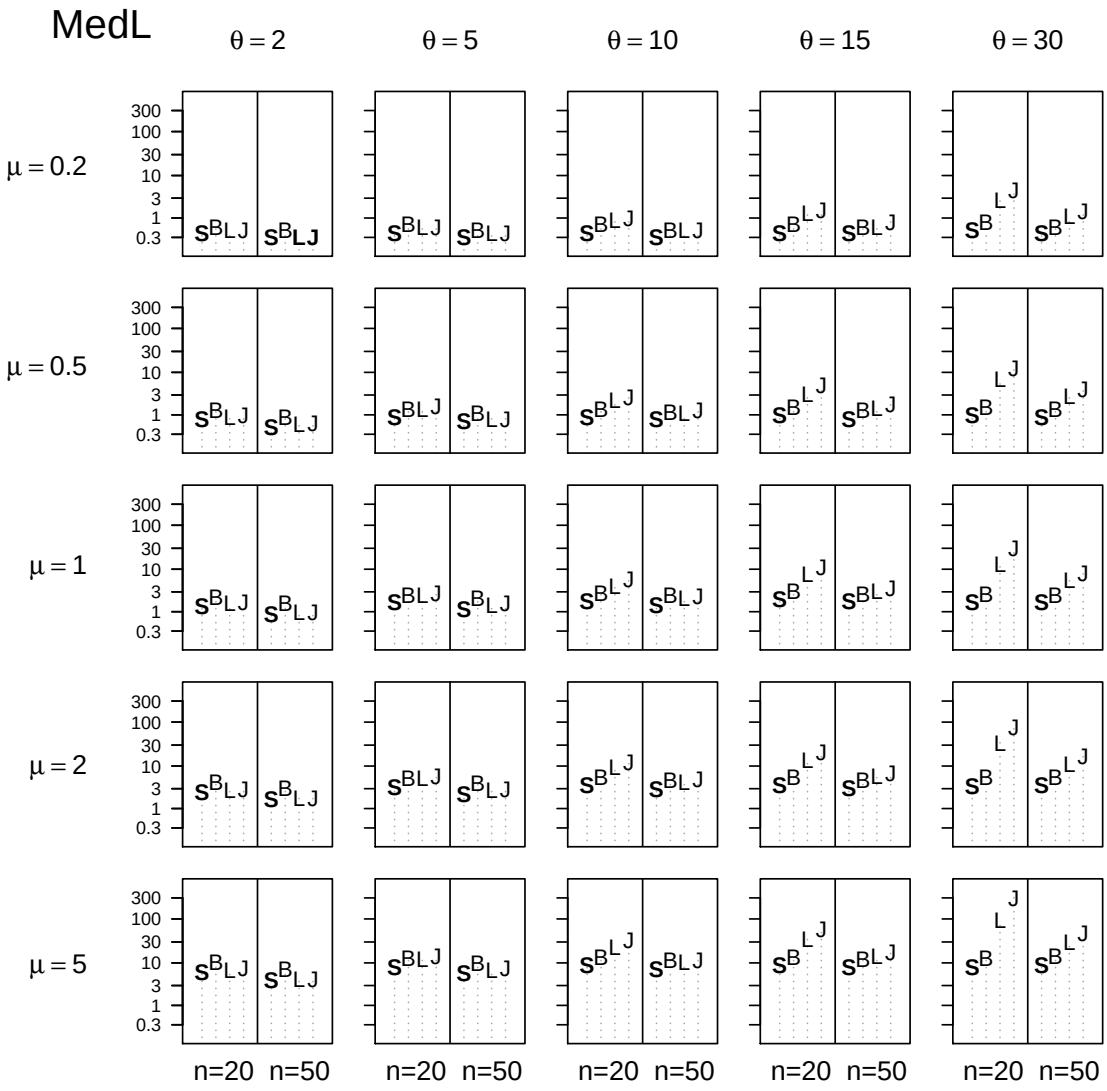


Figure II.3.3: Median length (MedL) of the different methods for constructing CIs, according to the parameter values (μ, θ) (S: Student, B: Bernstein, L: profile likelihood, J: Jeffreys). The bold font indicates the methods with narrowest MedL.

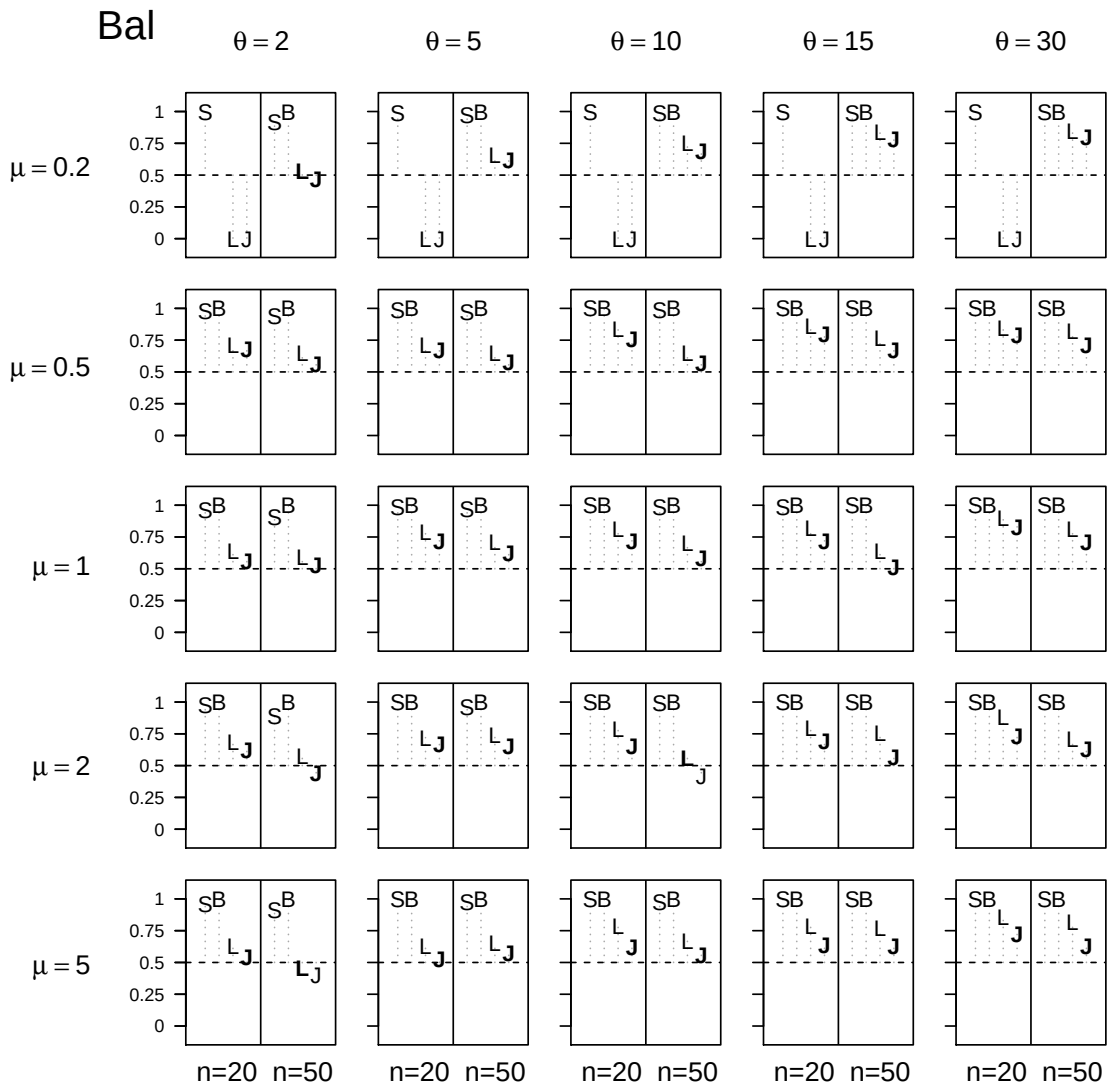


Figure II.3.4: Balance between left and right non-coverage (Bal) of the different methods for constructing CIs, according to the parameter values (μ, θ) (S: Student, B: Bernstein, L: profile likelihood, J: Jeffreys). The bold fonts indicate the methods with best balance (Bal closest to 0.5).

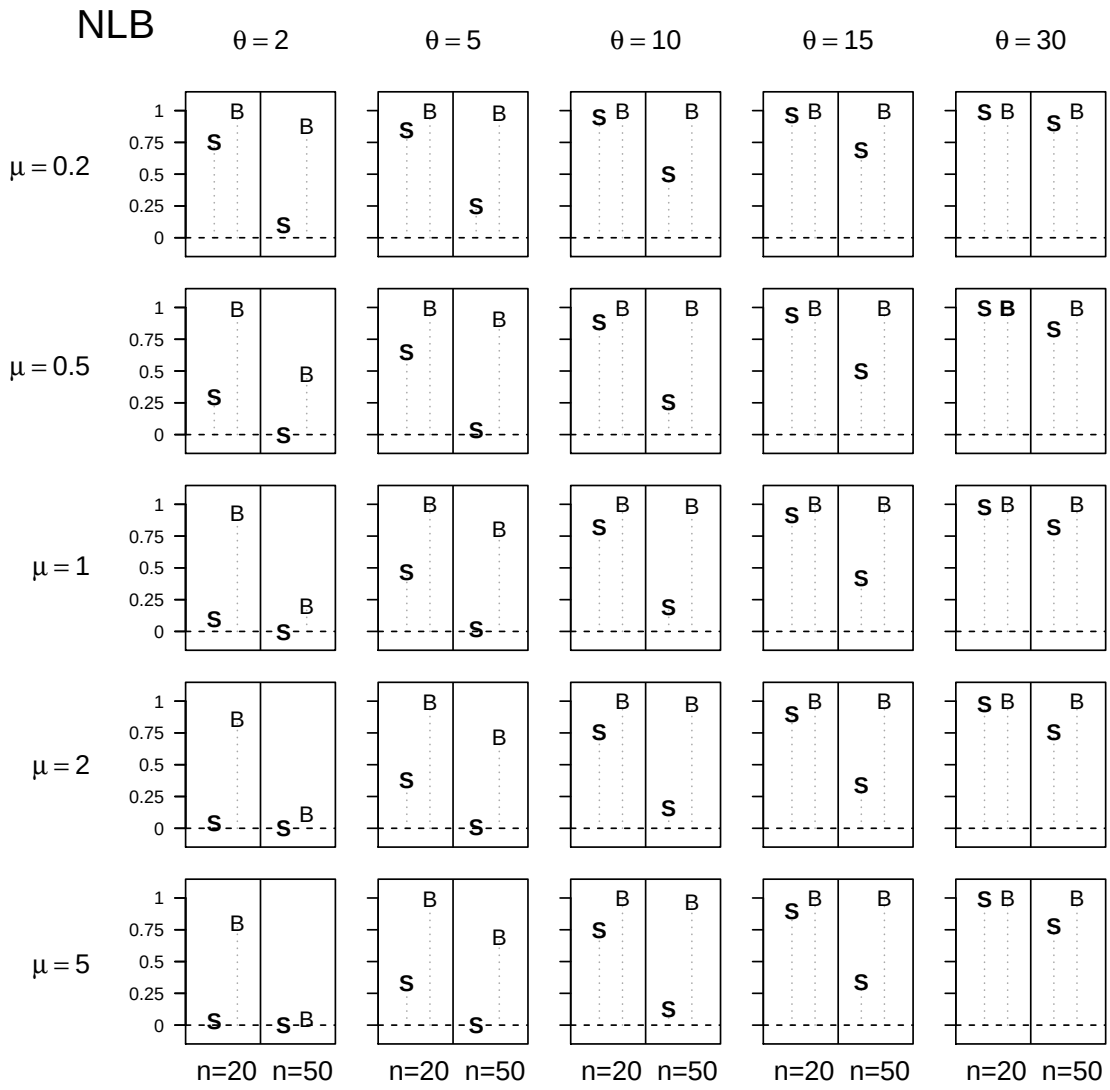


Figure II.3.5: Percentage of CIs with negative lower bound (NLB) for the CI_S and CI_B methods, according to the parameter values (μ, θ) (S: Student, B: Bernstein, L: profile likelihood, J: Jeffreys). CI_L and CI_J are not represented here because they always have a positive lower bound, by construction. The bold font indicates the method (among CI_S and CI_B) with NLB closest to 0.

Chapter 4

Estimation of mean and dispersion with random samples of overdispersed count data: comparison of sampling with fixed size, and sequential sampling

Vaudor, L., Lamouroux, N.

Abstract

Estimating the mean and dispersion of populations with count data is fundamental in many areas of science. Still, it might be problematic in some cases, for instance when populations under study are rare and clumped, and when sampling effort is limited due to practical or economical constraints. In particular, such conditions might cause fixed-size samples to comprise very few individuals, or very few non-null counts, leading to poor estimation of mean and dispersion. With sequential sampling, the investigator goes on sampling while some conditions (set by a stopping rule) are not met, e.g. while statistics such as the number of individuals (T) or number of non-null counts (S) are not higher than a certain value. In this study, we compared the quality of maximum likelihood estimation of the mean and dispersion of overdispersed populations, either with fixed-size sampling, or with sequential sampling with stopping rules based either on T or S. We compared these two types of sampling at the same average sample size, i.e. with sample size n_{fix} (for fixed-size sampling) such that $n_{fix} = E(N_{seq})$, where N_{seq} is the sample size in the sequential sampling case. We hypothesized that the populations under study were distributied according to a negative binomial, and tested varying ranges of parameter values of mean and dispersion. Our results show that sequential sampling with stopping rule based on T or S improves the estimation of dispersion but worsens the estimation of mean. Based on the root mean square error of estimators, sequential sampling improves the estimation of dispersion by 35% but worsens the estimation of mean by 55% (on average, accross all tested parameter ranges and stopping rules). Still, when stopping rules are based on S only, sequential sampling improves the estimation of dispersion by 37% and worsens the estimation of mean by 28% (on average, across all tested parameter ranges and stopping rules based on S). Sequential sampling with stopping rules based on the number of non-null counts S might thus be an interesting way to improve inference when populations are rare and overdispersed.

Keywords: Sequential; Inverse sampling; Fixed size; Negative binomial; Sampling effort;

II.4.1 Introduction

In many areas of scientific research e.g. ecology, economy, epidemiology, populations are studied based on count samples. The mean and variability of abundance are some of the most basic characteristics of interest that the investigators seek to estimate. When exhaustive census is impossible, abundance is generally inferred through a random, fixed sample size sampling design, in which the required sample size n_{fix} is fixed before the sampling is carried out.

Abundance data often exhibits overdispersion i.e. abundance variance is generally higher than its mean (Taylor, 1984). The negative binomial (NB) distribution, which is a generalization of the Poisson distribution allowing for overdispersion, is widely recognised as appropriate for modelling overdispersed count data (Anscombe, 1949; Bliss and Fisher, 1953; Taylor et al, 1979; Johnson et al, 1992, Vaudor et al., in revision). Because of overdispersion, the estimation of mean and dispersion can be particularly imprecise with random samples of fixed size.

Of course, this imprecision tends to be more important when the sample size is small. Hence, the investigator has to find a compromise between the sampling effort and the estimate precision. As noted by Young and Young (1998), in the fixed sample size context, “the primary difficulty in determining the optimal sample size [...] is that, in each case, the optimal sample size depends upon one or more unknown parameters. A worker may put in some “reasonable" values for these unknown parameters to determine approximate sample sizes. However, one can never be sure whether a predetermined fixed sample size will result in estimates [...] with the desired level of precision."’

An alternative to fixed sample size sampling is sequential sampling. In a sequential sampling design, sampling continues until certain conditions (set by a stopping rule) are met, e.g. until a certain number of individuals have been sampled. Then the sample size N_{seq} is a random variable, with a distribution varying according to population characteristics and stopping rule.

In some contexts, some sequential stopping rules guarantee that estimation attains a pre-specified degree of reliability (Binns, 1975; Gerrard and Cook, 1972; Mukhopadhyay et al, 1992; Johnson et al, 1996; Madden et al, 1996; Mukhopadhyay et al, 1997; Young and Young, 1998). Nevertheless, as underlined by Binns (1975), Willson et al (1984), Johnson et al (1996), and Young and Young (1998) in the case of a NB variable, these methods rely on prior knowledge of the dispersion of populations. Hence, Johnson et al (1996) tried to estimate the mean of a NB variable using sequential sampling with pre-specified degree of reliability, after estimating the dispersion using the multistage method of

Willson et al (1984). They dealt with low mean and quite high dispersion ($\mu \in [0.2, 3.1], \theta \in [3.1, 5.9]$). They showed that the actual degree of reliability was not inferior or equal to the pre-specified degree, in particular when the specified reliability was quite low. For instance, when they specified that the coefficient of variation of mean should be 40%, i.e. for average sample sizes close to ours (ranging from 23 to 74), the actual coefficient of variation ranged from 43% to 65%.

In practice the risk to have no estimates or really imprecise and/or biased estimates of mean and dispersion is more important when samples are small, mean is low and dispersion is high (Clark and Perry, 1989). Such conditions favour the occurrence of samples with few individuals and non-null counts (Vaudor and Lamouroux, submitted).

The stopping rule specified frequently relies on statistics such as total number of individuals T (e.g. Cox, 1952; Binns, 1975), or number of non-null counts S (e.g. Ye, 1993). When the distribution of populations is highly variable, sample size is likely to be highly variable itself. Indeed, in this case the realization of the stopping criterion might occur sooner or later according to how “lucky" the sampling is.

In this study, we compared the precision of estimations of mean and dispersion either with fixed-size sampling or with sequential sampling. This comparison is irrelevant unless the sample sizes are equivalent in both designs. We thus fixed sample size n_{fix} (in the fixed-size sampling case) such that $n_{fix} = E(N_{seq})$, where N_{seq} is the sample size in the sequential sampling case. We tested several stopping rules, either based on T or S, and we compared the results for various ranges of mean and dispersion.

II.4.2 Methods

Hypothesized distribution of data, and stopping rules

The present study was motivated by a fish abundance dataset (obtained through random sampling with fixed sample size, differing across sites). This dataset comprised 2258 count samples, characterized by small sizes ($20 \leq n_{fix} \leq 180$, with in 88% of the cases $20 \leq n_{fix} \leq 50$). In a previous study, Vaudor et al. (submitted) showed that the negative binomial distribution (NB) was an appropriate distribution model for these data.

For the present study, we hypothesized that the data were distributed according to a NB distribution:

$$pr_{\mu,\theta}(X = x) = \frac{\Gamma(x + \theta^{-1})}{x!\,\Gamma(\theta^{-1})} (\mu\theta)^x (1 + \mu\theta)^{-(x+\theta^{-1})} \tag{4.1}$$

where μ is the parameter of mean ($\mu > 0$), and θ is the parameter of dispersion ($\theta > 0$), and Γ is the Gamma function.

The samples of the fish abundance dataset were characterized by low observed mean $\hat{\mu}$ (1^{st} quantile=0.24, median=0.60 and 3^{rd} quantile=1.52) and high observed dispersion $\hat{\theta}$ (1^{st} quantile=3.14, median=6.66 and 3^{rd} quantile=13.90). The number of individuals in each sample, noted T, was ≤ 15 in 47% of the cases, and ≤ 30 in 64% of the cases. The number of non-null counts in each sample, noted S, was ≤ 5 in 50% of the cases, and ≤ 10 in 81% of the cases. For the present study, we tested stopping rules, and ranges of values μ and θ in accordance with these characteristics of the fish abundance dataset.

We tested 8 stopping rules of the form "Stop sampling when $Y_n \geq y_{min}$ and $N_{seq} \geq n_{min}$" where Y_n is a simple statistic of the sample of size n. We tested two different statistics: the number of observed individuals T, and the number of observed non-null counts S. The 8 tested stopping rules are: $\{T \geq 30$ and $N_{seq} \geq 10\}$, $\{T \geq 15$ and $N_{seq} \geq 10\}$, $\{S \geq 10$ and $N_{seq} \geq 10\}$, $\{S \geq 5$ and $N_{seq} \geq 10\}$, $\{T \geq 30$ and $N_{seq} \geq 2\}$, $\{T \geq 15$ and $N_{seq} \geq 2\}$, $\{S \geq 10$ and $N_{seq} \geq 2\}$, and $\{S \geq 5$ and $N_{seq} \geq 2\}$.

We considered that the investigator, when faced to the problem of choosing a sampling design, generally deals with a variety of mean and dispersion parameter values (several species, several sites, etc.). We hypothesized that both μ and θ were distributed as a log-Normal, with parameters ($m_\mu, sd_\mu = 1.2$) for μ and ($m_\theta, sd_\theta = 1$) for θ. We considered two possible values of m_μ ($m_\mu = -0.4$ or $m_\mu = 1$) and m_θ ($m_\theta = 2$ or $m_\theta = 1$). We thus dealt with four different parameter ranges, defined by ($m_\mu = -0.4$, $m_\theta = 2$), ($m_\mu = -0.4$, $m_\theta = 1$), ($m_\mu = 1$, $m_\theta = 2$), ($m_\mu = 1$, $m_\theta = 1$). We specified the parameter range

($m_\mu = -0.4$, $m_\theta = 2$) according to the values of mean and dispersion observed in the fish abundance dataset that motivated the study. Indeed, using ($m_\mu = -0.4$, $m_\theta = 2$) to simulate 2258 values of (μ, θ), and then using these values (μ, θ) to simulate fixed-size samples (with same sizes as in the fish data set) results in estimated values of mean and dispersion really close to those observed (mean : 1^{st} quantile=0.26, median=0.60 and 3^{rd} quantile=1.44, and dispersion: 1^{st} quantile=2.55, median=6.05 and 3^{rd} quantile=13.37).

Distribution of N_{seq}

The distribution of $Y_n = T$ is (Anraku and Yanagimoto, 1990):

$$pr_{\mu,\theta}(T=t) = \frac{\Gamma(t+n\theta^{-1}1)}{t!\,\Gamma(n\theta^{-1})}\left[\frac{1}{1+\mu\theta}\right]^{n\theta^{-1}}\left[\frac{\mu\theta}{1+\mu\theta}\right]^{t} \tag{4.2}$$

The distribution of $Y_n = S$ is:

$$\begin{aligned} pr_{\mu,\theta}(S=s) &= \left[pr_{\mu,\theta}(X=0)\right]^{n-s}\left[1-pr_{\mu,\theta}(X=0)\right]^{s} \\ &= \left[(1+\mu\theta)^{-}\theta^{-1}\right]^{n-s}\left[1-(1+\mu\theta)^{-\theta^{-1}}\right]^{s} \end{aligned} \tag{4.3}$$

For any (μ, θ) and any stopping rule $\{Y_n = y_{min} \text{ and } N \geq n_{min}\}$, equations (4.2) or (4.3) can be used to calculate the distribution of N_{seq}. Indeed,

$$\begin{aligned} &pr_{\mu,\theta}(N_{seq}=n) = pr_{\mu,\theta}(Y_n \geq y_{min}) - pr_{\mu,\theta}(Y_{n-1} \geq y_{min}), && \text{if } n > n_{min} \\ &pr_{\mu,\theta}(N_{seq}=n) = 0, && \text{if } n < n_{min} \\ &pr_{\mu,\theta}(N_{seq}=n) = pr_{\mu,\theta}(Y_n \geq y_{min}), && \text{if } n = n_{min} \end{aligned} \tag{4.4}$$

Thus the distribution of N_{seq} corresponding to any of the specified stopping rules, and any parameter range is:

$$pr(N_{seq}=n) = \int_0^{+\infty}\int_0^{+\infty} pr_{\mu,\theta}(N_{seq}=n)\,pr(\mu,\theta)\,d\mu\,d\theta \tag{4.5}$$

We calculate it approximately, considering discrete values μ_i and θ_j, regularly spaced on a logarithmic scale with base 2 (12 values μ_i comprised between 0.025 and 51.2, and 10 values θ_j comprised between 0.2 and 102.4):

$$pr(N_{seq}=n) = \sum_{i=1}^{12}\sum_{j=1}^{10} pr_{\mu_i,\theta_j}(N_{seq}=n)\int_{l_i}^{l_{i+1}}\int_{l_j}^{l_{j+1}} pr(\mu,\theta)\,d\mu\,d\theta \tag{4.6}$$

where the boundaries l_i and l_j are regularly spaced on a logarithmic scale with base 2 (such that l_{i+1} is the geometric mean of μ_i and μ_{i+1}, and l_{j+1} is the geometric mean of θ_j and θ_{j+1}).

We can then calculate $E(N_{seq}) = \sum_{n=1}^{+\infty} pr(N_{seq} = n)$ corresponding to any hypothesized parameter range and any stopping rule.

Simulations and calculation of root mean square error

For each of the 120 couples (μ_i, θ_j), each of the 8 specified stopping rules, and each of the 4 hypothesized parameter ranges, we simulated samples either through sequential sampling, or through fixed-size sampling with sample size n_{fix}. All simulations and subsequent estimations were carried out using R (R Development Core Team, 2010).

In the fixed-size sampling case, n_{fix} is set to the rounded value of $E(N_{seq})$, which depends on the parameter range - cf equations (4.5) and (4.6)-. We thus simulated 1000 samples, for 8 stopping rules, for 120 couples (μ_i, θ_j), in 5 cases (sequential sampling, or fixed-size sampling with size n_{fix} corresponding to each of the four parameter ranges), i.e. a total of 4,800,000 samples.

For each simulated sample k ($k = 1, 2, \ldots, 1000$), we calculated the maximum likelihood estimates $\hat{\mu}_{i,j,k}$ and $\hat{\theta}_{i,j,k}$ of μ_i and θ_j. The maximum likelihood estimate $\hat{\mu}_{i,j,k}$ actually corresponds to the arithmetic mean of the sample (Anraku and Yanagimoto, 1990). We determined the maximum likelihood estimate $\hat{\theta}_{i,j,k}$ using the function “nlminb" in R (R Development Core Team, 2010). This function carries out the maximization of the log-likelihood over the parameter space $]0, +\infty[$, using a Newton-type algorithm. Some of the samples generated contained only zeroes (null samples), in which case, θ can not be estimated. These cases were excluded from further calculations. Some of the samples generated are underdispersed (such that variance = mean) in which case log-likelihood has no maximum over $]0, +\infty[$. In this case, we set $\hat{\theta}_{i,j,k} = 0.001$.

We then calculated the root mean square error (RMSE) of $\hat{\mu}$ and $\hat{\theta}$ for each couple (μ_i,θ_j):

$$RMSE_{\mu_i,\theta_j}(\hat{\mu}) = \sqrt{\sum_{k=1}^{1000} (\hat{\mu}_{i,j,k} - \mu_i)^2} \tag{4.7}$$

$$RMSE_{\mu_i,\theta_j}(\hat{\theta}) = \sqrt{\sum_{k=1}^{1000} (\hat{\theta}_{i,j,k} - \theta_j)^2} \tag{4.8}$$

We used our hypothesized ranges of μ and θ to calculate a global approximate RMSE($\hat{\mu}$) and

RMSE($\hat{\theta}$), either in the case of sequential sampling or in the case of fixed-size sampling through:

$$RMSE(\hat{\mu}) = \sqrt{\sum_{i=1}^{12}\sum_{j=1}^{10}\left[RMSE_{\mu_i,\theta_j}(\hat{\mu})\right]^2 \int_{l_i}^{l_{i+1}}\int_{l_j}^{l_{j+1}} pr(\mu,\theta)d\mu d\theta} \tag{4.9}$$

$$RMSE(\hat{\theta}) = \sqrt{\sum_{i=1}^{12}\sum_{j=1}^{10}\left[RMSE_{\mu_i,\theta_j}(\hat{\theta})\right]^2 \int_{l_i}^{l_{i+1}}\int_{l_j}^{l_{j+1}} pr(\mu,\theta)d\mu d\theta} \tag{4.10}$$

II.4.3 Results

Across all tested populations characteristics and sampling plans, $E(N_{seq})$ ranges from 13 to 97 (Table 1 and 2). On average $E(N_{seq}) = 59$ for μ in its low range, whereas $E(N_{seq}) = 24$ on average for μ in its high range. The effect of the range of μ is particularly strong for stopping rules based on T (Fig.2), for which on average $E(N_{seq}) = 72$ for μ in its low rage and $E(N_{seq}) = 23$ for μ in its high range (Tables 1 and 2). On average $E(N_{seq}) = 46$ for θ in its high range, whereas $E(N_{seq}) = 38$ on average for θ in its low range. The effect of the range of θ is particularly strong for stopping rules based on S (Fig.2), for which on average $E(N_{seq}) = 43$ for θ in its high range and $E(N_{seq}) = 29$ for θ in its low range. The stopping rules $T = 15$ and $S = 5$ correspond to average sample sizes corresponding to about half the average sample sizes with the stopping rules $T = 30$ and $S = 10$ respectively. The variability of N_{seq} is important: on average, $SD(N_{seq})/E(N_{seq}) = 1.23$. The variability of N_{seq} is more important when stopping rules are based on T (Fig.2): on average, $SD(N_{seq})/E(N_{seq}) = 1.50$ for stopping rules based on T while on average, $SD(N_{seq})/E(N_{seq}) = 0.97$) for stopping rules based on S (Tables 1 and 2). The minimum sample size n_{min} has little influence on $E(N_{seq})$: the difference between $E(N_{seq})$ for $n_{min} = 2$ and $E(N_{seq})$ for $n_{min} = 2$ is ≤ 3 (Tables 1 and 2).

The proportion of null samples P_0 ranges from 0.11% to 3.27% across all tested populations characteristics and fixed-size sampling plans, whereas $P_0 = 0$ by construction in sequential sampling plans (Table 1 and Table 2). The proportion of underdispersed samples p_U ranges from 1.0% to 19.8%, and is generally larger for smaller sample sizes. Overall, P_U is larger for fixed-size sampling plans (average $P_U = 9.80\%$) than for sequential sampling plans (average $P_U = 6.51\%$). In fixed-size sampling plans, P_U is mainly influenced by the range of μ: on average, $P_U = 11.21\%$ and 8.40% respectively for μ in its low and high range. On the contrary, the range of θ has little influence on P_U: on average, $P_U = 9.78\%$ and 9.83% respectively for θ in its low and high range. In sequential sampling plans, P_U is influenced by both the ranges of μ and θ. On average, $P_U = 7.46\%$ and 5.56% respectively for μ in its low and high range, and $P_U = 8.05\%$ and 4.97% respectively for μ in its low and high range.

RMSE($\hat{\mu}$) is strongly influenced by the range of μ, whether the sampling plan has size fixed or is sequential. On average, RMSE($\hat{\mu}$) is 3.73 times higher for the high range of μ than for the low range of μ. It is also influenced by the range of dispersion θ: on average RMSE($\hat{\mu}$) is 1.53 times higher for the high range of θ than for the low range of θ. Across all sampling plans and characteristics of populations, RMSE($\hat{\mu}$) is higher in the case of sequential sampling, compared to fixed-size sampling.

The relative difference in $RMSE$, $D_{rel} = \left[RMSE_{seq}(\hat{\mu}) - RMSE_{fix}(\hat{\mu})\right] / RMSE_{fix}(\hat{\mu})$, ranges from -87% to -12%. D_{rel} is particularly strong (in absolute value) for the low range of μ (on average, $D_{rel} = -55\%$) compared to the high range of μ (on average, $D_{rel} = -36\%$). D_{rel} is about the same whether θ is in its high or low range ($D_{rel} = -47\%$ and -44% respectively). D_{rel} is a lot stronger in absolute value when stopping rules are based on T ($D_{rel} = -62\%$ on average) rather than on S ($D_{rel} = -28\%$ on average).

RMSE($\hat{\theta}$) is strongly influenced by the range of θ, whether the sampling design is fixed-size sampling or sequential sampling. On average, RMSE($\hat{\theta}$) is 2.46 times higher for the low range of θ than for the high range of θ. Across nearly all sampling plans and characteristics of populations, RMSE($\hat{\theta}$) is lower in the case of sequential sampling, compared to fixed-size sampling. The only exceptions are for the sampling rule $T = 15$ and the high range of μ. The relative difference in RMSE, $D_{rel} = \left[RMSE_{seq}(\hat{\theta}) - RMSE_{fix}(\hat{\theta})\right] / RMSE_{fix}(\hat{\theta})$, ranges from -1% to +77%. D_{rel} is particularly strong for the low range of μ (on average, $D_{rel} = 35\%$) compared to the high range of μ (on average, $D_{rel} = 10\%$). D_{rel} is a lot stronger when stopping rules are based on S ($D_{rel} = +37\%$ on average) rather than on T ($D_{rel} = +7\%$ on average).

II.4.4 Discussion

Many estimators of mean and dispersion of the NB exist, which might be more or less precise according to the sampling strategy and characteristics of the population of interest. In particular, it is disputable whether estimators used in the case of fixed-size sampling should be used in the case of sequential sampling. Indeed, Anscombe (1954) noted that "when the number of observations depends on the observations themselves, it can happen that a fixed-sample size analysis of the observations is grossly wrong [...] and it can also happen that a fixed-sample-size analysis is only very slightly in error". In this study, we used the same method (maximum likelihood) to estimate mean and dispersion for both fixed-size and sequential sampling. It is intuitive that in the sequential case, the fact that sampling is stopped when some individuals are observed tends to favour high estimates of the mean, and cause classical estimators of mean to be biased upwards (Whitehead, 1986), causing RMSE($\hat{\mu}$) to be generally higher in the case of sequential sampling. Still, the likelihood has the same form $L_{\mu,\theta} = \prod_{i=1}^{N} pr_{\mu,\theta}(X_i)$ whether the sampling has fixed size ($N = n_{fix}$) or not ($N = N_{seq}$ is a random variable). Lindsey (1997) underlined the fact that "some relevant information about the parameters [...] can be available from a given experiment, but is absent from the likelihood usually reported". According to this idea, de Cristofaro (2004) and Bunouf and Lecoutre (2006) argued that the information associated with the sampling design should be taken into consideration when carrying out inference on samples. As a result, Bunouf (2006) proposed a Bayesian estimator of the probability of success in a sequence of Bernoulli experiments that takes into account the nature of the design and have good properties. Still, we considered that comparing estimators was beyond the scope of the present study, and chose to use maximum likelihood as a simple, common estimator of mean and dispersion for both parameters and for all sampling strategies.

In this study, we hypothesised that sampling effort was proportional to sample size, which might be simplistic in some cases. In particular, it might be advantageous to know the sample size prior to sampling, in order to be able to plan efficiently the field or lab work, quantities of gear, spatial position of sampled units, etc. In such a case, even with a similar sample size, on average, than in the fixed-size case, sequential sampling corresponds to a greater effort. On the other hand, considering that there is a basic cost associated to a sampling operation (independent of the number of sample units collected during this operation), the investigator might prefer drawing samples in a few batches rather than one by one (accelerated sequential sampling or multistage sampling: Mukhopadhyay et al, 1992, 1997).

However, these practical constraints are highly variable according to the populations under study and are thus hard to formalize and quantify in a simple, general way. As a result, we considered that sample size could be used as a general and simple approximate measure of sampling effort.

Our results show that sequential sampling improves estimation of dispersion but worsens estimation of mean. Investigators generally seek to estimate both parameters through the same sampling design, hence it might be difficult to decide whether sampling design should be adopted. Nevertheless, a better estimate of dispersion could improve estimation of a confidence interval around the mean, as uncertainty in the estimate of mean tends to be underestimated due to poor estimation (underestimation) of dispersion (Vaudor and Ecochard, submitted). The comparison of interval estimation of mean and dispersion with fixed-size or sequential sampling, and the study of bias-corrected estimators for sequential sampling designs, could thus be interesting leads towards determining whether or not sequential sampling designs could improve estimation of both mean and dispersion in case populations are rare and overdispersed.

Acknowledgements

Table II.4.1: Properties of the estimators of μ and θ with sampling with fixed size sampling (fix) and sequential sampling (seq) according to hypothesized parameter range and stopping rule with $\{n_{min} = 10\}$. The properties summarized here are the expected value of N ($E(N_{seq}) = n_{fix}$) and standard deviation of N_{seq} ($sd(N_{seq})$), the proportion of null samples (P_0), of underdispersed samples (P_U), $RMSE(\hat{\mu})$ and $RMSE(\hat{\theta})$, and relative difference $D_{rel} = [RMSE_{seq} - RMSE_{fix}]/RMSE_{fix}$

Stopping rule	Param. range		N_{seq}		P_0	P_U	P_U	RMSE($\hat{\mu}$)			RMSE($\hat{\theta}$)		
	m_μ	m_θ	$E(N_{seq})$	$sd(N_{seq})$	fix (%)	fix (%)	seq (%)	fix	seq	D_{rel} (%)	fix	seq	D_{rel} (%)
$T \geq 30$	-0.4	2	97	146	0.18	10.8	1.0	1.01	3.22	-69	15.35	12.79	20
	-0.4	1	93	144	0.11	10.1	3.6	0.64	1.93	-67	6.07	4.87	25
	1	2	31	45	0.82	9.9	1.9	5.87	10.56	-44	14.66	14.11	4
	1	1	28	43	0.25	6.7	1.9	3.78	6.38	-41	5.97	5.52	8
$T \geq 15$	-0.4	2	52	76	1.03	9.2	2.4	1.37	3.16	-57	16.61	15.07	10
	-0.4	1	49	74	0.61	11.9	7.6	0.88	1.92	-54	6.98	6.15	13
	1	2	20	24	2.38	7.8	2.1	7.21	10.55	-32	14.19	15.47	-8
	1	1	17	22	0.90	7.1	3.6	4.85	6.37	-24	6.03	6.45	-7
$S \geq 10$	-0.4	2	72	70	0.46	9.4	5.2	1.18	1.82	-35	15.69	10.43	50
	-0.4	1	52	56	0.53	11.3	10.2	0.85	1.4	-39	6.95	3.94	77
	1	2	42	37	0.36	13.4	10.2	5.01	6.25	-20	15.15	13.6	11
	1	1	26	22	0.32	6.5	6.0	3.93	4.77	-18	5.92	4.66	27
$S \geq 5$	-0.4	2	36	38	2.31	10.9	9.2	1.62	2.56	-36	15.87	11.21	42
	-0.4	1	26	29	2.42	16.2	18.1	1.21	1.77	-32	7.14	4.8	49
	1	2	22	20	1.84	8.0	6.2	7.05	8.64	-18	14.26	12.23	17
	1	1	15	11	1.23	7.7	7.4	5.16	5.88	-12	6.07	4.59	32

Table II.4.2: Properties of the estimators of μ and θ with sampling with fixed size sampling (fix) and sequential sampling (seq) according to hypothesized parameter range and stopping rule with $\{n_{min} = 2\}$. The properties summarized here are the expected value of N ($E(N_{seq}) = n_{fix}$) and standard deviation of N_{seq} ($sd(N_{seq})$), the proportion of null samples (P_0), of underdispersed samples (P_U), $RMSE(\hat{\mu})$ and $RMSE(\hat{\theta})$, and relative difference $D_{rel} = [RMSE_{seq} - RMSE_{fix}]/RMSE_{fix}$

Stopping rule	Param. range		N_{seq}		P_0	P_U	P_U	RMSE($\hat{\mu}$)			RMSE($\hat{\theta}$)		
	m_μ	m_θ	$E(N_{seq})$	$sd(N_{seq})$	fix (%)	fix (%)	seq (%)	fix	seq	D_{rel} (%)	fix	seq	D_{rel} (%)
$T \geq 30$	-0.4	2	97	146	0.19	10.6	1.1	1.02	7.85	-87	15.66	12.96	21
	-0.4	1	93	144	0.11	10.0	3.9	0.64	4.53	-86	5.99	4.92	22
	1	2	29	46	1.00	9.2	1.8	6.09	25.37	-76	14.57	14.42	1
	1	1	26	44	0.31	6.5	3.3	3.93	14.73	-73	6.08	5.64	8
$T \geq 15$	-0.4	2	51	77	1.08	9.1	2.9	1.39	7.61	-82	16.07	15.47	4
	-0.4	1	48	74	0.61	11.8	9.0	0.89	4.42	-80	6.91	6.17	12
	1	2	17	26	3.27	7.7	3.2	8.06	23.96	-66	14.09	15.65	- 1
	1	1	14	23	1.44	8.2	8.2	5.34	13.95	-62	6.09	6.36	- 4
$S \geq 10$	-0.4	2	72	70	0.44	9.4	5.3	1.16	1.83	-37	15.59	10.46	49
	-0.4	1	52	56	0.53	11.4	10.2	0.85	1.40	-39	6.73	3.94	71
	1	2	42	37	0.35	13.2	10.2	4.97	6.30	-21	14.91	13.64	9
	1	1	26	22	0.29	6.4	5.9	3.92	4.79	-18	6.04	4.67	29
$S \geq 5$	-0.4	2	36	38	2.41	10.8	9.8	1.67	2.81	-40	15.79	11.17	41
	-0.4	1	26	30	2.47	16.4	19.8	1.22	2.09	-42	7.22	4.82	5
	1	2	21	20	2.02	7.8	7.0	7.07	9.62	-26	14.12	12.27	15
	1	1	13	12	1.68	8.3	10.1	5.60	7.07	-21	5.99	4.64	29

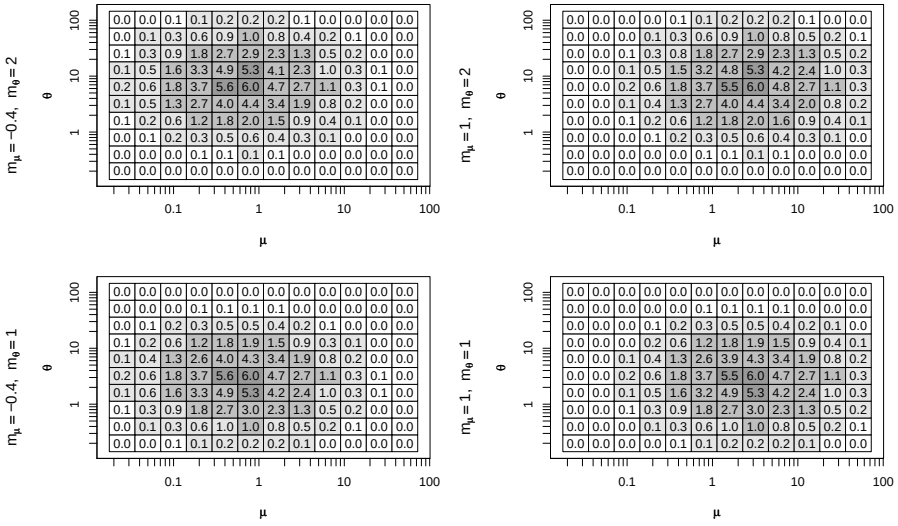


Figure II.4.1: Hypothesized parameter ranges ($m_\mu = -0.4$, $m_\theta = 2$), ($m_\mu = -0.4$, $m_\theta = 1$), ($m_\mu = 1$, $m_\theta = 2$) and ($m_\mu = 1$, $m_\theta = 1$). The numbers indicate the probability (%) that (μ, θ) lies inside each cell. The grey levels summarize these densities according to categories (dark grey: $\leq 5\%$, medium dark grey : $\leq 1\%$, medium light grey : $\leq 0.1\%$, light grey ($< 0.1\%$)

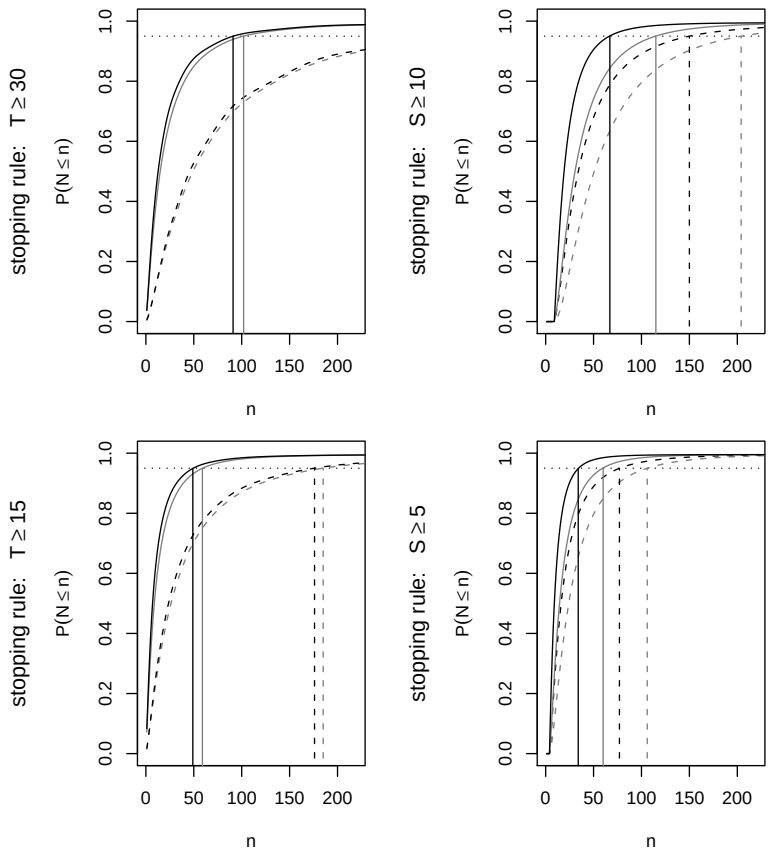


Figure II.4.2: Distribution of N according to the stopping rule with $\{n_{min} = 2\}$, and according to the hypothesized parameter range: ($m_\mu = -0.4$, $m_\theta = 2$) (dashed, light grey), ($m_\mu = -0.4$, $m_\theta = 1$) (dashed, dark grey), ($m_\mu = 1$, $m_\theta = 2$) (plain, light grey) and ($m_\mu = 1$, $m_\theta = 1$) (plain, dark grey). Dashed lines thus correspond to a low hypothesized range for μ and light grey corresponds to a high hypothesized range for θ. The vertical lines indicate the 95th percentiles of the distributions.

Bibliography

Al-Saleh MF, Al-Batainah FK (2003) Estimation of the shape parameter k of the negative binomial distribution. Applied Mathematics and Computation 143(2-3):431–441

Anraku K, Yanagimoto T (1990) Estimation for the negative binomial distribution based on the conditional likelihood. Communications in Statistics-Simulation and Computation 19(3):771–786

Anscombe FJ (1949) The statistical analysis of insect counts based on the negative binomial distribution. Biometrics 5(2):165–173

Anscombe FJ (1954) Fixed-sample-size analysis of sequential observations. Biometrics 10:89–100

Bartlett MS (1947) The use of transformations. Biometrics 3(1):39–52

Bennett G (1962) Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association 57(297):33–45

Binet FE (1986) Fitting the negative binomial distribution. Biometrics 42(4):989–992

Binns M (1975) Sequential estimation of the mean of a negative binomial distribution. Biometrika 62(2):433–440

Bland JM, Altman DG (1996) Transformations, means, and confidence intervals. British Medical Journal 312(7038):1079–1079

Bliss C, Fisher R (1953) Fitting the negative binomial distribution to biological data. Biometrics 9(2):176–200

Boos DD, Hughes-Oliver JM (2000) How large does n have to be for z and t intervals? The American Statistician 54(2):121–128

Boyles RA (2008) The role of likelihood in interval estimation. The American Statistician 62(1):22–26

Brown LD, Cai TT, DasGupta A, Agresti A, Coull BA, Casella G, Corcoran C, Mehta C, Ghosh M, Santner TJ, Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Statistical Science 16(2):101–133

Brown LD, Cai TT, DasGupta A (2003) Interval estimation in exponential families. Statistica Sinica 13(1):19–49

Bunouf P (2006) Lois bayésiennes a priori dans un plan binomial séquentiel. State thesis, Université de Rouen

Bunouf P, Lecoutre B (2006) Bayesian priors in sequential binomial design. Comptes Rendus Mathematique 343(5):339–344

Cai T (2005) One-sided confidence intervals in discrete distributions. Journal of Statistical Planning and Inference 131(1):63–88

Clark SJ, Perry JN (1989) Estimation of the negative binomial parameter kappa by maximum quasi-likelihood. Biometrics 45(1):309–316

Cook RD, Weisberg S (1990) Confidence curves in nonlinear regression. Journal of the American Statistical Association 85(410):544–551

Copp GH (1992) Comparative microhabitat use of cyprinid larvae and juveniles in a lotic floodplain channel. Environmental Biology of Fishes 33(1-2):181–193

Cox DR (1952) A note on the sequential estimation of means. Proc Camb Phil Soc 48:447–450

de Cristofaro R (2004) On the foundations of likelihood principle. Journal of Statistical Planning and Inference 126(2):401–411

Cunningham RB, Lindenmayer DB (2005) Modeling count data of rare species: Some statistical issues. Ecology 86(5):1135–1142

Di Stefano J, Fidler F, Cumming G (2005) Effect size estimates and confidence intervals: an alternative focus for the presentation and interpretation of ecological data. In: New Trends in Ecology Research, pp 71–102

Efron B (1987) Better bootstrap confidence intervals. Journal of the American Statistical Association 82(397):171–185

Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1(1):54–75

Evans DA (1953) Experimental evidence concerning contagious distributions in ecology. Biometrika 40:186–211

Fears TR, Benichou J, Gail MH (1996) A reminder of the fallibility of the wald statistic. American Statistician 50(3):226–227

Fink D (1997) A compendium of conjugate priors

Fletcher D, Faddy M (2007) Confidence intervals for expected abundance of rare species. Journal of Agricultural Biological and Environmental Statistics 12(3):315–324

Gayen AK (1949) The distribution of student's t in random samples of any size drawn from non-normal universes. Biometrika 36(3/4):353–369

Gerrard DJ, Cook RD (1972) Inverse binomial sampling as a basis for estimating negative binomial population densities. Biometrics 28(4):971–980

Gray BR (2005) Selecting a distributional assumption for modelling relative densities of benthic macroinvertebrates. Ecological Modelling 185(1):1–12

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72(358):320–338

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc Roy Soc A 186:453–461

Johnson GA, Mortensen DA, Young LJ, Martin AR (1996) Parametric sequential sampling based on multistage estimation of the negative binomial parameter k. Weed Science 44(3):555–559

Johnson N, Kotz S, Kemp A (1992) Univariate discrete distributions, 2nd edn. Wiley series in probability and mathematical statistics, Wiley, New York

Johnson N, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, Wiley series in probability and mathematical statistics, vol 2, 2nd edn. Wiley, New York

Johnson NJ (1978) Modified t tests and confidence intervals for asymmetrical populations. Journal of the American Statistical Association 73(363):536–544

Kvanli AH, Shen YK, Deng LY (1998) Construction of confidence intervals for the mean of a population containing many zero values. Journal of Business and Economic Statistics 16(3):362–368

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14

Lamouroux N, Capra H, Pouilly M, Souchon Y (1999) Fish habitat preferences in large streams of southern france. Freshwater Biology 42(4):673–673

Land CE (1972) An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics 14(1):145–158

Le Cam L (1986a) Asymptotic methods in statistical decision theory. Springer series in statistics, Springer-Verlag

Le Cam L (1986b) The central limit theorem around 1935. Statistical Science 1(1):78–91

Lewin W, Freyhof J, Huckstorf V, Mehner T, Wolter C (2009) When no catches matter: Coping with zeros in environmental assessments. Ecological Indicators 10(3):572–583

Lindsey JK (1997) Stopping rules and the likelihood function. Journal of Statistical Planning and Inference 59(1):167–177

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

Lord D, Miranda-Moreno LF (2008) Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: A bayesian perspective. Safety Science 46(5):751–770

Madden L, Hughes G, Munkvold G (1996) Plant desease incidence: inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. Crop Protection 15(7):621–632

Mante C, Durbec JP, Dauvin JC (2005) A functional data-analytic approach to the classification of species according to their spatial dispersion. application to a marine macrobenthic community from the bay of morlaix (western english channel). Journal of Applied Statistics 32(8):831–840

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8(11):1235–1246

McArdle BH, Anderson MJ (2004) Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. Canadian Journal of Fisheries and Aquatic Sciences 61(7):1294–1302

McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York

Meeker WQ, Escobar LA (1995) Teaching about approximate confidence regions based on maximum likelihood estimation. The American Statistician 49(1):48–53

Mukhopadhyay N, Bendel RB, Nikolaidis NP, Chattopadhyay S (1992) Efficient sequential sampling strategies for environmental monitoring. Water Resources Research 28(9):2245–2256

Mukhopadhyay N, Padmanabhan AR, Solanky TKS (1997) On estimating the reliability after sequentially estimating the mean: The exponential case. Metrika 45(3):235–252

Mullahy J (1986) Specification and testing of some modified count data models. Journal of Econometrics 33(3):341–365

Muus B, Dahlström P (1991) Guide des poissons d'eau douce et pêche. Les guides du naturaliste, Delachaux et Niestlé, Neuchâtel

Newcombe RG (1998) Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine 17(8):857–872

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods in Ecology and Evolution 1(2):118–122

Olivier JM, Carrel G, Lamouroux N, Dole-Olivier MJ, Malard F, Bravard JP, Amoros C (2009) The rhone river basin. In: Rivers of Europe, Academic Press, London, pp 247–295.

Olsson U (2005) Confidence intervals for the mean of a log-normal distribution. Journal of Statistics Education 13(1)

O'Neill MF, Faddy MJ (2003) Use of binary and truncated negative binomial modelling in the analysis of recreational catch data. Fisheries Research 60(2-3):471–477

Park BJ, Lord D (2008) Adjustment for maximum likelihood estimate of negative binomial dispersion parameter. Transportation Research Record 2061:9–19

Pawitan Y (2000) A reminder of the fallibility of the wald statistic: Likelihood explanation. American Statistician 54(1):54–56

Pearson E, Adyanthaya NK (1929) The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika 21(1):259–286

Persat H (1988) De la biologie des populations de l'ombre commun, thymallus thymallus (l. 1758), à la dynamique des communautés dans un hydrosysème fluvial aménagé, le haut-Rhône français. State thesis, Université Lyon 1

Persat H, Copp GH (1990) Electric fishing and point abundance sampling for the ichthyology of large rivers. Developments in Electric Fishing pp 197–209

Piegorsch WW (1990) Maximum-likelihood-estimation for the negative binomial dispersion parameter. Biometrics 46(3):863–867

Pieters EP, Gates CE, Matis JH, Sterling WL (1977) Small sample comparison of different estimators of negative binomial parameters. Biometrics 33(4):718–723

Potts JM, Elith J (2006) Comparing species abundance models. Ecological Modelling 199(2):153–163

Power JH, Moser EB (1999) Linear model analysis of net catch data using the negative binomial distribution. Canadian Journal of Fisheries and Aquatic Sciences 56(2):191–200

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Regis J, Pattee E, Lebreton JD (1981) A new method for evaluating the efficiency of electric fishing. Archiv Fur Hydrobiologie 93(1):68–82

Ridout MS, Demétrio CGB, Hinde JP (1998) Models for count data with many zeros. In: XIXth International Biometrics Conference, pp 179–192

Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to sage data. Biostatistics 9(2):321–332

Rosenblum M, Van der Laan MJ (2008) Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. UC Berkeley Division of Biostatistics Working Paper Series (Working Paper 237)

Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: A bivariate boxplot. American Statistician 53(4):382–387

Saha K, Paul S (2005) Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics 61(1):179–185

Shilane D, Hubbard AE, Evans SN (2008) Confidence intervals for negative binomial random variables of high dispersion. UC Berkeley Division of Biostatistics Working Paper Series Working Paper 242

Shilane D, Evans SN, Hubbard AE (2010) Confidence intervals for negative binomial random variables of high dispersion. International Journal of Biostatistics 6(1)

Sokal R, Rohlf F (1995) Biometry: the principles and practice of statistics in biological research, 3rd edn. W. H. Freeman and Co ., New York.

Sprott DA (1975) Application of maximum likelihood methods to finite samples. Sankhya: The Indian Journal of Statistics 37(Series B):259–270

Swift MB (2009) Comparison of confidence intervals for a poisson mean - further considerations. Communications in Statistics-Theory and Methods 38(5):748–759

Taylor LR (1984) Assessing and interpreting the spatial distributions of insect populations. Annual review of entomology 29:321

Taylor LR, Woiwod IP, Perry JN (1978) The density-dependence of spatial behaviour and the rarity of randomness. Journal of Animal Ecology 47(2):383–406

Taylor LR, Woiwod IP, Perry JN (1979) Negative binomial as a dynamic ecological model for aggregation, and the density dependence of k. Journal of Animal Ecology 48(1):289–304

Venables W, Ripley B (1999) Modern applied statistics with S-PLUS. Springer-Verlag, New York

Venzon DJ, Moolgavkar SH (1988) Method for computing profile-likelihood-based confidence intervals. Journal of the Royal Statistical Society Series C: Applied Statistics 37(1):87–94

Vos PW, Hudson S (2005) Evaluation criteria for discrete confidence intervals: Beyond coverage and length. The American Statistician 59(2):137–142

Wallace DL (1958) Asymptotic approximations to distributions. The Annals of Mathematical Statistics 29(3):635–654

Wang YN (1996) Estimation problems for the two-parameter negative binomial distribution. Statistics and Probability Letters 26(2):113–114

Warton DI (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16(3):275–289

Wasserman L (1991) An inferential interpretation of default priors. Technical Report

Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB (1996) Modelling the abundance of rare species: Statistical models for counts with extra zeros. Ecological Modelling 88(1-3):297–308

Wenger SJ, Freeman MC (2008) Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology 89(10):2953–2959

Whitehead J (1986) On the bias of maximum likelihood estimation following a sequential test. Biometrika 73(3):573–581

Willson LJ, Folks JL, Young JH (1984) Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k. Biometrics 40(1):109–117

Ye K (1993) Reference priors when the stopping rule depends on the parameter of interest. Journal of the American Statistical Association 88(421):360–363

Young LJ, Young JH (1998) Statistical Ecology: A Population Perspective. Kluwer Academic Publishers, Boston/Dordrecht/London

Zhou XH, Gao S (1997) Confidence intervals for the log-normal mean. Statistics in Medicine 16(7):783–790