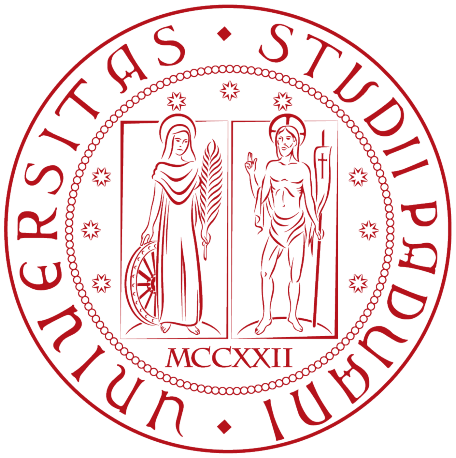
CONVOLUTIONAL NEURAL NETWORKS APPLIED TO VERY HIGH ENERGY GAMMA-RAY DATA

NICOLA MARINELLO



DEI - Department of Information Engineering
Ingegneria delle Telecomunicazioni
Università degli Studi di Padova

Marzo 2019

Professor: Alessandro Chiuso
Assistant Professor: Michele Doro
Supervisor: Rubén Lopez Coto

[October 29, 2018 at 22:13 – classicthesis v4.6]

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.

1939 – 2005

ABSTRACT

Short summary of the contents in English. . . a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

ZUSAMMENFASSUNG

Kurze Zusammenfassung des Inhaltes in deutscher Sprache. . .

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— knuth:1974 [knuth:1974]

ACKNOWLEDGMENTS

Put your acknowledgments here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, Jörg Weber, Henri Menke, Claus Lahiri, Clemens Niederberger, Stefano Bragaglia, Jörn Hees, Scott Lowe, Dave Howcroft, José M. Alcaide, David Carlisle, Ulrike Fischer, Hugues de Lassus, Csaba Hajdu, Dave Howcroft, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was intially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and for the contributions to the original style.

1 Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

I SOME KIND OF MANUAL

1 INTRODUCTION 3

II THE SHOWCASE

2 MASTER EQUATION 7
- 2.1 Distribuzione binomiale negativa 7
- 2.2 Distribuzione logaritmica di Fisher 8
  - 2.2.1 La distribuzione di Fisher come caso particolare della binomiale negativa 9

3 MATH TEST CHAPTER 11

4 METODI DI UPSCALINGR 15
- 4.1 Metodo della binomiale negativa 15
  - 4.1.1 Proprietà di auto-somiglianza della distribuzione binomiale negativa 16
  - 4.1.2 Il numero di specie a scala totale 17
- 4.2 Metodo della distribuzione di Fisher 18
  - 4.2.1 Proprietà di auto-somiglianza della distribuzione logaritmica di Fisher 18
  - 4.2.2 Il numero di specie a scala totale 19
- 4.3 Metodo $Chao_{wor}$ 20
  - 4.3.1 Il numero di specie a scala totale 21

5 CONCLUSIONI 25

III APPENDIX

A APPENDIX TEST 29
- A.1 Appendix Section Test 29
- A.2 Another Appendix Section Test 29

LIST OF FIGURES

LIST OF TABLES

Table A.1 Autem usu id 29

LISTINGS

Listing A.1 A floating example (`listings` manual) 29

ACRONYMS

Part I

SOME KIND OF MANUAL

1

INTRODUCTION

Uno dei più importanti obiettivi in ecologia è quello di dedurre le proprietà generali di un ecosistema campionando solo una frazione di esso.
Per ragioni pratiche la biodiversità viene misurata o monitorata tipicamente a piccole scale, ma è importante poter conoscere la biodiversità dell'intero ecosistema
....

Uno strumento statistico comunemente usato per descrivere la normalità o rarità della presenza delle specie in una comunità ecologica è la **relative species abundance distribution** (**SAD** o **RSA**) cioè il numero di individui per specie presenti all'interno di una regione. Normalmente la SAD è misurata a scale locali

Part II

THE SHOWCASE

You can put some informational part preamble text here. Illo principalmente su nos. Non message *occidental* anglo-romanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

2

MASTER EQUATION

In questo capitolo vediamo come si possono ricavare la distribuzione binomiale negativa e la distribuzione logaritmica modellizzando la dinamica dell'abbondanza delle specie attraverso un'equazione che descrive la nascita e la morte degli individui : la *birth-death master equation*.
Entrambe le distribuzioni possono essere derivate da principi primi alla base dei processi biologici: sia $P_{n,s}(t)$ la probabilità che ad un certo tempo t, la specie s abbia esattamente n individui, con $s \in \{1, ..., S\}$. Assumiamo che la dinamica della popolazione di ogni specie sia governata da due termini: i rate di nascita e di morte, rispettivamente, $b_{n,s}$ e $d_{n,s}$, per una specie s con n individui. L'equazione che regola l'evoluzione di $P_{n,s}(t)$ per $n \geq 0$ è la seguente:

$$\frac{\partial P_{n,s}(t)}{\partial t} = P_{n\text{-}1,s}(t)b_{n\text{-}1,s} + P_{n+1,s}(t)d_{n+1,s} - P_{n,s}(t)b_{n,s} - P_{n,s}(t)d_{n,s}. \quad (2.1)$$

Imponendo condizioni al contorno riflettenti, $b_{\text{-}1,s} = d_{0,s} = 0$, la (2.1) è valida anche per $n = 0$ e n=1. Per $n > 0$ la soluzione stazionaria è:

$$P_{n,s} = P_{0,s} \prod_{i=0}^{n-1} \frac{b_{i,s}}{d_{i+1,s}} \quad (2.2)$$

dove il termine $P_{0,s}$ è il fattore di normalizzazione che può essere trovato imponendo la condizione $\sum_{n=1}^{\infty} P_{n,s} = 1$. Notiamo che la somma inizia da n=1 in quanto non si considerano specie con abbondanza nulla.

2.1 DISTRIBUZIONE BINOMIALE NEGATIVA

Assumiamo ora che $b_{n,s}$ dipenda da un termine indipendente dal numero di individui b_s, cioè il tasso di nascita pro capite, e dal termine r_s, che tiene conto di eventi di immigrazione o di interazioni intraspecifiche:

$$b_{n,s} = b_s(n + r_s). \quad (2.3)$$

Analogamente assumiamo che il termine $d_{n,s}$ dipenda da d_s, cioè dal tasso di morte pro capite:

$$d_{n,s} = d_s n. \quad (2.4)$$

Queste supposizioni sono ragionevoli in ecologia.
Sostituendo questi ultimi termini nella (2.2) e denotando con $\xi_s = b_s/d_s$, si ottiene:

$$P_{n,s} = P_{o,s}\binom{n+r_s-1}{n}\xi_s^n.$$

La costante di normalizzazione può essere calcolata imponendo:

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{o,s}\sum_{n=0}^{\infty}\binom{n+r_s-1}{n}\xi_s^n = P_{o,s}[1-(1-\xi_s)^{r_s}](1-\xi_s)^{-r_s}$$

Dunque la probabilità che una specie s abbia s individui all'equilibrio è data da una binomiale negativa di parametri (r_s, ξ_s) e normalizzata per abbondanze non nulle ($n \geq 1$):

$$P_{n,s}^{NB} = \frac{1}{1-(1-\xi_s)^{r_s}}\binom{n+r_s-1}{n}\xi_s^n(1-\xi_s)^{r_s}. \tag{2.5}$$

Sotto l'ipotesi della teoria neutrale,secondo la quale le specie sono considerate demograficamente equivalenti(cioè ogni individuo ha la stessa probabilità di procreare,morire e migrare), possiamo rimuovere l'indice s di specie dall'equazione sopra, ottenendo così una RSA per l'ecosistema in esame.

2.2 DISTRIBUZIONE LOGARITMICA DI FISHER

Notiamo che,scegliendo in modo diverso i termini $b_{n,s}$ e $d_{n,s}$,si può ottenere,partendo dalla *birth death master equation* (2.1), un'altra importante distribuzione: la distribuzione logaritmica di Fisher. Assumiamo che la dinamica della popolazione di una comunità sia governata dal corso ecologico e dalla speciazione casuale invece che dalla migrazione da comunità esterne (?). Allora possiamo porre:

$$b_{n,s} = b_s n + \delta_{n,o}\nu \tag{2.6}$$

Aggiungendo la condizione al contorno riflettente $b_{o,s} = \nu$ si ha che il tasso di nascita tiene conto della riproduzione e della speciazione. In particolare, il parametro ν assicura che, se le specie si estinguono, la comunità rimane sempre popolata da un individuo. Dunque sostituendo la (2.4) e la (2.6) nella (2.2) e definendo $x_s = b_s/d_s$, si trova la seguente soluzione stazionaria:

$$P_{n,s} = P_{o,s}\frac{\nu}{b_s}\frac{x_s^n}{n}. \tag{2.7}$$

La costante di normalizzazione $P_{o,s}$ si determina imponendo:

$$1 = \sum_{n=1}^{\infty} P_{n,s} = P_{o,s}\frac{\nu}{b_s}\sum_{n=0}^{\infty}\frac{x_s^n}{n} = P_{o,s}\frac{\nu}{b_s}[-\log(1-x_s)].$$

Dunque abbiamo:

$$P_{n,s}^{LS} = -\frac{1}{\log(1-x_s)}\frac{x_s^n}{n}. \tag{2.8}$$

Anche in questo caso assumiamo che le specie siano equivalenti e possiamo dunque omettere l'indice s.

2.2.1 *La distribuzione di Fisher come caso particolare della binomiale negativa*

Osserviamo che la distribuzione binomiale negativa converge ad una distribuzione logarimtica nel limite di r che tende a zero:

$$\lim_{r\to 0} P_n^{NB} = \lim_{r\to 0} \frac{(1-\xi)^r}{1-(1-\xi)^r}\binom{n+r-1}{n}\xi^n = \frac{\xi^n}{-n\ln(1-\xi)}, \tag{2.9}$$

dove si è usato il fatto che:

$$\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1\Gamma(r))} \approx \frac{r}{n+1},$$

per $r \approx 0$.
Notiamo dunque che la (2.9) coincide con la (2.8) ponendo $x = \xi$.

3

MATH TEST CHAPTER

La ricchezza delle specie, cioè il numero di specie, è la misura più intuitiva e più frequentemente usata per caratterizzare la diversità di un dato insieme: questa possiede caratteristiche e proprietà matematiche intuitive ben visibili utili per costruire modelli di comunità. Negli studi biogeografici, le mappe di monitoraggio di specie, flora e fauna locali e regionali forniscono informazioni solo sull'assenza o presenza di una determinata specie in ogni località. Dunque, per questo tipo di studi, la ricchezza delle specie è l'unico dato disponibile per quantificare la diversità di un sistema ed è importante quindi sviluppare dei modelli per ottenere questo tipo di informazione partendo da dei campioni di popolazione ridotti.
La ricchezza di specie dipende fortemente dal metodo di campionamento e dalla completezza del campione, il modo di raccogliere le informazioni porta ad avere principalmente due tipi di dati: dati di abbondanza e dati di incidenza.
Per fissare la notazione: consideriamo una comunità costituita da N individui appartenenti a S specie distinte. Sia N_i il numero di individui della *i*-esima specie con *i*=1,2,...,S, $N_i > 0$ e $N = \sum_{n=1}^{S} N_i$. L'abbondaza relativa della specie *i*-esima è $p_i = N_i/N$, dunque $\sum_{n=1}^{S} p_i = 1$. Qui N, S, N_i e p_i rappresentano i valori veri, ma sconosciuti, dei parametri fondamentali dell'insieme in esame.
In base al metodo di campionamento si distinguono due tipi di strutture dati: dati di abbondanza e dati di incidenza.

3.0.0.1 *Dati di abbondanza*

In molti studi biologici o ecologici solitamente gli individui vengono osservati o osservati in un dato momento e vengono classificati in base alla specie di appartenenza. Si prenda ad esempio un campione di n individui dall'insieme in esame e si ipotizzi di osservare un totale di S_{obs} specie: questo è il *campione di riferimento*. Questo tipo di data-set può essere ottenuto usando due schemi di campionamento differenti:

1. *campionamento di tipo discreto* in cui l'unità campionaria è un individuo. Ad esempio, si campiona un numero fissato n di individui in una certa area. Qui la grandezza del campione n è fissata e ogni specie può essere rappresentata al massimo da n individui;

2. *campionamento di tipo continuo* nel quale il campione viene quantificato misurando su scale continue come tempo, area o volume d'acqua. Si fissa, per esempio, una certa area da studiare o un

certo periodo di tempo per il quale analizzare il sito in esame. Con questo protocollo di campionamento il numero di individui osservati è una variabile casuale e ogni specie può essere rappresentata da un numero qualsiasi di individui.

3.0.0.2 *Dati di incidenza*

In alcune indagini le unità di campionamento non sono gli individui, ma "trap net quadrat plot ?" o periodi di tempo: queste vengono campionate casualmente e indipendentemente. Ad esempio un'area di interesse può essere divisa in in un certo numero di "quadrats?" approssimativamente della stessa area, tra questi ne vengono selezionati alcuni in modo casuale sui quali eseguire l'indagine.
A volte risulta impossibile contare esattamente il numero di individui per ogni specie che appaiono in ogni campione (ad esempio per microrganismi, invertebrati o piante) e quindi viene registrata solo la loro incidenza (presenza o assenza) nel campione. Dunque le stime si basano su degli insiemi di unità di campionamento in cui è registrata solo la presenza o assenza di una certa specie in un dato campione invece che la sua abbondanza.

Avendo a disposizione questo tipo di dati si possono seguire due approcci per stimare la diversità del campione: quello parametrico e quello non parametrico. In questo lavoro useremo dati di abbondanza ottenuti con campionamento continuo, infatti consideriamo una frazione a di un'area A nella quale sono stati registrati il numero di individui presenti in corrispondenza della loro specie di appartenenza. D'ora in poi ci occuperemo solo di questo caso particolare.

3.0.0.3 *Modelli parametrici e non parametrici*

Negli approcci parametrici che analizzeremo si assume che la distribuzione dell'abbondanza delle specie abbia una certa forma,governata da dei parametri. Facendo il fit della curva dell'abbondanza relativa delle specie dei dati osservati si ottengono i valori dei parametri che,secondo le caratteristiche e le proprietà della distribuzione ipotizzata, permettono di dedurre le informazioni sulla diversità del sistema osservato.
Negli approcci non parametrici, invece, non si fanno assunzioni sulla distribuzione sottostante alla curva dell'abbondanza delle specie. L'intuizione e concetto base su cui si fondano i metodi di stima non parametrici è che le specie abbondanti, cioè quelle a cui appartengono un elevato numero di individui, non danno alcuna informazione sulla ricchezza delle specie inosservate, mentre le specie rare, contengono quasi tutte le informazioni sulla biodiversità. Dunque, la maggior parte degli estimatori non parametrici si basa sulle frequenze di apparizione di basso ordine, specialmente sul numero di *singletons* e

doubletons, cioè sul numero specie che vengono registrare contenere uno o due individui.

4

METODI DI UPSCALINGR

In questa sezione vediamo in dettaglio come è possibile ricostruire la biodiversità di un intero ecosistema a partire da un campione ridotto di SAD.
Analizzeremo prima due metodi parametrici, quello della binomiale negativa e della distribuzione logaritmica di Fisher, e poi un metodo non parametrico, quello dell'estimatore di $Chao_{wor}$.

4.1 METODO DELLA BINOMIALE NEGATIVA

Di seguito analizzeremo in dettaglio le proprietà e i passaggi che ci permettono di ottenere le informazioni desiderate.
Quando facciamo upscaling siamo interessati alla SAD ed al numero totale di specie presenti a scala totale,cioè in tutta l'area della foresta A. Denotiamo con P($n \mid 1$) la probabilità che una specie abbia esattamente n individui a scala totale(qui con il numero 1 si denota l'intera foresta), anche nota come *abbondanza relativa delle specie* RSA. Notiamo che P($n \mid 1$) deve essere definita solamente per $n \geq 1$ poiché, a scala totale, una data specie deve avere almeno un individuo. In questo contesto si ipotizza che la SAD segua una distribuzione binomiale negativa, $P(n \mid r,\xi)$ di parametri (r,ξ):

$$P(n|1) = c(r,\xi)P(n|r,\xi) \tag{4.1}$$

con

$$P_n = \binom{n+r-1}{n}\xi^n(1-\xi)^r, c(r\xi) = \frac{1}{1-(1-\xi)^r} \tag{4.2}$$

dove c è la costante di normalizzazione. Quest'ultima è stata calcolata imponendo $\sum_{n=1}^{\infty} P(n|1)$, dove la somma parte da 1 poiché le specie con abbondanza nulla a scala totale saranno assenti anche a scale ridotte. Notiamo che $p(n \mid r,\xi)$ è normalizzata per $n \geq 0$: questo perché, nei sotto campioni, esiste una probabilità non nulla di trovare una specie, presente nell'intera foresta, avente n=0 individui.In questo modo si tiene conto del numero di specie mancanti nei sotto campioni. Consideriamo ora un campione di foresta di area a e definiamo p=a/A la scala del campione, cioè la frazione di foresta osservata. Come primo passaggio calcoliamo la RSA del campione assumendo che quest'ultima non sia influenzata da correlazioni spaziali. Quest'ipotesi è ben soddisfatta ed è stata verificata usando dati di foreste generati *in silico* a vari gradi di correlazione spaziale.(cit, ci devo tornare sopra??)

Sotto queste ipotesi la probabilità che una specie presenti k individui in un'area a=pA, condizionata dal fatto che presenta n individui nella regione totale A è data dalla distribuzione binomiale:

$$P_{binom}(k \mid p,n) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{se } k = 0, ..., n \\ 0 & \text{se } nk>n \end{cases} \tag{4.3}$$

Infatti,in assenza di correlazioni spaziali, la probabilità che uno degli individui di una specie si trovi in una regione di area a è esattamente p.(?controllare?)

Mostriamo ora un risultato chiave per il metodo di upscaling:

4.1.1 *Proprietà di auto-somiglianza della distribuzione binomiale negativa*

Sia P($n \mid 1$)=$c(r,\xi)P(n \mid r,\xi)$ la RSA della foresta a scala totale e denotiamo con $P(k \mid r,\xi)$ la probabilità che una specie abbia abbondanza k alla scala $p \in (0,1)$, condizionata dal fatto che alla scala totale A sono presenti n individui di quella specie. Se $P(k \mid n,p)$=$P_{binom}(n|r,\xi)$ segue una distribuzione binomiale, allora la RSA $P_{sub}(k|p)$ alla scala di campionamento p è ancora una binomiale negativa per $k \geq 1$ con il parametro ξ riscalato e lo stesso r:

$$P_{sub}(k|p) = \begin{cases} c(r,\xi)P(k|r,\xi), & k \geq 1 \\ 1 - c(r,\xi)/c(r,\hat{\xi}_p), & k=o \end{cases} \tag{4.4}$$

con

$$\hat{\xi}_p = \frac{p\xi}{1-\xi(1-p)} \tag{4.5}$$

DIMOSTRAZIONE?

Ricordiamo che questo metodo fa uso solamente delle informazioni che si possono ottenere da un campione ad una certa scala p^*, infatti noi abbiamo informazioni solo sulle abbondanze delle $S^* \leq S$ specie presenti nel campione esaminato. Denotando il numero di specie di abbondanza k alla scala p^* con $S^*(k)$, otteniamo, per $k \geq 1$:

$$\frac{S^*(k)}{S^*} \equiv P(k|p^*) = \frac{P_{sub}(k|p^*)}{\sum_{k' \geq 1} P_{sub}(k'|p^*)} = \frac{P(k|r,\hat{\xi}_{p^*})}{\sum_{k' \geq 1} P(k'|r,\hat{\xi}_{p^*})} = c(r,\hat{\xi}_{p^*})P(k|r,\hat{\xi}_{p^*})$$

(4.6)

che, dalla (4.1), è una binomiale negativa normalizzata per $k \geq 1$, mentre $P(k|r, \hat{\xi}_{p^*}$ è normalizzata per $k \geq 0$. Per quanto detto sopra otteniamo dunque il seguente risultato: partendo da una distribuzione binomiale negativa per la RSA a scala globale, anche la RSA a scala ridotta risulta distribuita secondo una binomiale negativa di parametri lo stesso r e $\hat{\xi}^*_p$ riscalato. Una RSA avente la proprietà di avere la stessa forma funzionale a scale differenti è detta essere *invariante per forma*.

4.1.2 *Il numero di specie a scala totale*

Fittando la RSA dei dati alla scala p^* possiamo dunque trovare i parametri r e $\hat{\xi}^*_p$ e, invertendo l'equazione (4.5), troviamo:

$$\xi = \frac{\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(1 - p^*)} \tag{4.7}$$

Usando ancora la (4.5) per eliminare ξ dall'ultima equazione, otteniamo la seguente relazione per il parametro ξ alle due scale p e p^*:

$$\hat{\xi}_p = \frac{p\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(p - p^*)} \equiv U(p, p^*|\hat{\xi}_{p^*}) \tag{4.8}$$

dalla quale, ovviamente, è possibile riottenere sia la (4.5) che la (4.7) ponendo $\xi \equiv \hat{\xi}_{p=1}$.

Vogliamo ora determinare la relazione tra il numero totale di specie S alla scala totale p=1 e il numero totale di specie osservate localmente S_p alla scala p. D'ora in avanti per denotare il numero di specie alla scala locale useremo la notazione $S^* \equiv S_{p^*}$. Notiamo che:

$$P_{sub}(k{=}o|p^*) = \frac{S - S^*}{S} \tag{4.9}$$

$$P_{sub}(k{=}o|p^*) = \frac{S^*(k)}{S}. \tag{4.10}$$

Usando la seconda delle (4.4), il numero di specie presenti nell'intera foresta è dato, in termini dei dati del campione osservato, da:

$$S = \frac{S^*}{1 - P_{sub}(k = 0|p^*)} = S^* \frac{1 - (1 - \xi)^r}{1 - (1 - \hat{\xi}^*_p)^r} \tag{4.11}$$

Notiamo che, se si assume che la RSA segua una distribuzione binomiale negativa a scala globale, il valor medio dell'abbondanza totale riscala linearmente con l'area, infatti: (AGGIUNGERE EQ S26)

4.2 METODO DELLA DISTRIBUZIONE DI FISHER

Ora mostreremo che è possibile risalire al numero di specie anche quando si suppone che la SAD a scala globale sia distribuita secondo una log-series.
Supponiamo che la RSA a scala globale sia distribuita secondo una distribuzione logaritmica con parametro x:

$$P(n|1) = P^{LS}(n|x) = \alpha(x)\frac{x^n}{n}, \alpha(x) = -(\log(1-x))^{-1} \tag{4.12}$$

dove $\alpha(x)$ è la costante di normalizzazione. Assumendo anche questa volta che la RSA del campione non sia affetta da correlazioni spaziali si trova che anche la log-series soddisfa la proprietà di auto somiglianza.

4.2.1 *Proprietà di auto-somiglianza della distribuzione logaritmica di Fisher*

Sia $P(n|1) = \alpha(x)P^{LS}(n|x)$ la RSA alla scala globale e denotiamo con $P(k)|n,p$ la probabilità che una specie abbia abbondanza k nel campione alla scala $p \in (0,1)$ condizionata dal fatto alla scala totale A la specie possiede n individui.
Se $P(k \mid n,p)$=$P_{binom}(k|n,p)$ è distribuita secondo una binomiale, allora la RSA alla scala del campione, $P^{LS}_{sub}(k|p)$, è ancora una log-series per $k \geq 1$ con il parametro x riscalato:

$$P^{LS}_{sub}(k|p) = \begin{cases} \alpha(x)P^{LS}(k|\hat{x}_p) & k \geq 1 \\ 1 - \alpha(x)/\alpha(\hat{x}_p) & k{=}0 \end{cases} \tag{4.13}$$

con

$$\hat{x}_p = \frac{px}{1-x(1-p)} \tag{4.14}$$

DIMOSTRAZIONE??

Notiamo che (4.14) è analoga a (4.5). Dunque l'analogo di (4.7) è

$$x = \frac{\hat{x}_p}{p + \hat{x}_p(1-p)} \tag{4.15}$$

e l'equazione (4.8) vale anche in questo caso.
La RSA può essere ottenuta come nell'equazione (4.6) ed è data da:

$$P(k|p) = \frac{P^{LS}_{sub}}{\sum_{k' \geq 1} P^{LS}_{sub}(k'|p)} = \alpha(\hat{x}_p)\frac{\hat{x}_p^k}{k} = P^{LS}(n|\hat{x}_p) \tag{4.16}$$

Poiché la distribuzione logaritmica di Fisher è un caso particolare della binomiale negativa, è anch'essa invariante per scala.

4.2.2 *Il numero di specie a scala totale*

Il numero di specie con popolazione $k \geq 1$ presenti nel sotto-campione di area a=pA è dato da:

$$S_p(k) \equiv SP_{sub}(k|p) = S\alpha(x)\frac{\hat{x}_p^k}{k} = \hat{\alpha}\frac{\hat{x}_p^k}{k} \tag{4.17}$$

dove abbiamo unito le costanti S e $\alpha(x)$ in un unico termine $\hat{\alpha}$ che non dipende dalla scala p del campione. Quando ci riferiremo alla scala p^* useremo, per brevità di notazione, $S^*(k) \equiv S_{p^*}(k)$.
Allora il numero totale di specie S^* e l'abbondanza totale N^* (?) alla scala p^* sono date rispettivamente da:

$$S^* = \sum_{k=1}^{\infty} S^*(k) = -\hat{\alpha}\log(1-\hat{x}_{p^*}) \tag{4.18}$$

$$N^* = k\sum_{k=1}^{\infty} S^*(k) = \hat{\alpha}\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}} \tag{4.19}$$

Poiché S^* e N^* sono note dal campione, possiamo trovare $\hat{\alpha}$ risolvendo la seguente equazione:

$$N^* - \hat{\alpha}(\exp(\frac{S^*}{\hat{\alpha}}) - 1) = 0 \tag{4.20}$$

che si ottiene inserendo l'espressione di $\hat{x}_{p^*}$ da (4.18) nella (4.19).

Vogliamo ora dedurre le informazioni a scala globale p=1 dai dati disponibili alla scala p=p^*. Dalle considerazioni precedenti sappiamo che $\hat{\alpha}$ è un parametro indipendente dalla scala, dunque abbiamo le seguenti relazioni per S e N:

$$S = -\hat{\alpha}\log(1-x) \tag{4.21}$$

$$N = \hat{\alpha}\frac{x}{1-x} \tag{4.22}$$

dalle quali otteniamo:

$$S = \hat{\alpha}\log(1+\frac{N}{\hat{\alpha}}), \hat{\alpha} = S\alpha(x). \tag{4.23}$$

Dunque per dedurre la biodiversità a scala globale, S, è necessaria una stima dell'abbondanza totale N. Prendiamo N=N^*/p^*. Notiamo che questo è consistente con il nostro quadro teorico nel quale assumiamo che la RSA sia "form-invariant(?)": infatti si può dimostrare che, se si assume che la RSA segua una distribuzione di Fisher a scala globale, il valor medio dell'abbondanza totale riscala linearmente con l'area:

$$\mathbb{E}(N^*) = \sum_{k=1}^{\infty} kS^*(k) = \sum_{k=1}^{\infty} k\hat{\alpha}\frac{\hat{x}_{p^*}^k}{k} = \alpha\frac{\hat{x}_{p^*}}{1-\hat{x}_{p^*}} = \hat{\alpha}\frac{px}{1-x} = p^*\mathbb{E}(N), \tag{4.24}$$

dove abbiamo usato la (4.14).

4.3 METODO $Chao_{wor}$

Vediamo ora il metodo non parametrico sviluppato da Chao, nato nell'ambito di uno schema di campionamento senza reinserimento. Questo è il sistema id indagine più usato quando si devono campionare individui come, ad esempio, insetti che vengono uccisi quando osservati: dunque nessun individuo può essere osservato ripetutamente. Inoltre viene applicato anche ad altri protocolli di campionamento, ad esempio nello studio delle foreste, nel quale gli alberi vengono censiti per 'plots o quadrats ??' che sono selezionati senza ripetizione. In questo schema di campionamento ogni individuo (o ogni unità di campionamento) può essere indagato solo una volta.
Assumiamo che in un ecosistema ci siano S specie indicizzate da 1 a S. Sia N_i (abbondanza assoluta) il numero di individui appartenenti alla i-esima specie, i=1,...,S, e $N_i > 0$. La popolazione totale dunque è data da $N = \sum_{i=1}^{S} N_i$. Assumiamo che la dimensione del campione N sia nota, cioè è nota la frazione di campionamento.
Supponiamo di prendere dall'intero ecosistema un sotto campione di n individui, campionandoli senza reinserimento. Sia X_i la frequenza della specie campionata cioè il numero di individui della i-esima specie osservati nel campione. Solo le specie con $X_i > 0$ sono osservabili nel campione. Sia f_k il numero di specie nel campione che sono rappresentate esattamente da k individui, dunque f_0 denota il numero di specie che non sono state osservate nel campione. Dunque abbiamo che $n = \sum_{i=1}^{S} X_i = \sum_{k>1} kf_k$. Definiamo $p^* = n/N$ la frazione di campionamento e S^* il numero di specie osservate nel sotto campione, $S^* = \sum_{k>1} f_k$.
Generalmente, la probabilità che una specie venga rilevata, o rate di rilevamento, dipende sia dall'abbondanza della specie nel campione sia da caratteristiche specifiche degli individui come ad esempio il modo di spostarsi e muoversi all'interno dell'ambiente , colore, forma

e dimensione.
Consideriamo dunque il caso generale in cui la probabilità di trovare un individuo possa variare a seconda della specie di appartenenza e indichiamola con $\theta_i > 0$ per la i-esima specie. Sotto queste ipotesi, definendo $q_i = N_i/N$ come l'abbondanza relativa, il rate di rilevamento per la i-esima specie diventa $\psi_i = \frac{N_i\theta_i}{\sum_{k=1}^{S} N_k\theta_k} = \frac{q_i\theta_i}{\sum_{k=1}^{S} q_k\theta_k}$ con i=1,...,S. Intuitivamente, il numero di individui che hanno la stessa possibilità di essere osservati è dato da $N_i\psi_i$, ma poiché questo potrebbe essere un numero non intero, definiamo una variabile a valori interi Z_i, che rappresenta il numero di individui che hanno la stessa possibilità di essere osservati per la i-esima specie. Siccome $Z \geq 1$ e la frazione di individui campionata è n/N, si può modellare il vettore $\mathbf{Z}$=$(Z_1, Z_2, \ldots, Z_S)$ attraverso una distribuzione multinomiale troncata con N celle totali e con celle di probabilità le $(\psi_1^*, \psi_2^*, \ldots, \psi_S^*)$, dove $\psi_i^* = \psi_1/\mathrm{Pz}, z_i \geq 1$, i=1,...,S, $\mathbf{z}$=$(z_1, z_2, \ldots, z_S)$ e $\sum_{i=1}^{S} z_i = N$. Per ogni dato valore di $\mathbf{z}$=$(z_1, z_2, \ldots, z_S)$, le frequenze con cui appaiono gli individui della specie i-esima nel campione, $(X_1, X_2, \ldots, X_S)$, seguono una distribuzione ipergeometrica generalizzata:

$$P(X_i = x_i, i = 1, 2, ..., S) = \binom{z_1}{x_1}\binom{z_2}{x_2}...\binom{z_S}{x_S} / \binom{N}{n} \tag{4.25}$$

$$z_i \geq 1, \sum_{i=1}^{S} z_i = N$$

Sulla base di questo modello generale, la distribuzione marginale per ognuna delle frequenze con le quali vengono individuate le specie è una distribuzione ipergeometrica:

$$P(X_i = x_i) = \binom{z_i}{k}\binom{N - z_i}{n - k} / \binom{N}{n} \tag{4.26}$$

4.3.1 *Il numero di specie a scala totale*

Vediamo dunque com'è possibile dedurre, sotto queste ipotesi, il numero di specie a scala totale a partire da un vettore di abbondanze ottenuto esaminando una frazione dell'intero ecosistema.

Il valore di aspettazione per i contatori di frequenze f_k usando la (4.26) è:

$$\mathbb{E}(f_k) = \sum_{i}^{S} P(X - i = x_x) = \sum_{i=1}^{S} \binom{z_i}{k}\binom{N - z_i}{n - k} / \binom{N}{n} \tag{4.27}$$

In particolare:

$$\mathbb{E}(f_o) = \sum_{i=1}^{S} \binom{N - z_i}{n} / \binom{N}{n}$$

$$\mathbb{E}(f_1) = \sum_{i=1}^{S} \binom{z_i}{1} \binom{N-z_i}{n\text{-}1} / \binom{N}{n} = \sum_{i=1}^{S} \frac{nz_i}{N-z_i-n+1} \binom{N-z_i}{n} / \binom{N}{n}$$

$$\mathbb{E}(f_2) = \sum_{i=1}^{S} \binom{z_i}{2} \binom{N-z_i}{n\text{-}2} / \binom{N}{n} = \sum_{i=1}^{S} \frac{n(n-1)z_i(z_i-1)}{2(N-z_i-n+1)(N-z_i-n+2)} \binom{N-z_i}{n} / \binom{N}{n}$$

Per la disuguaglianza di Cauchy-Schwarz si ha:

$$\left\{ \sum_{i=1}^{S} \frac{nz_i}{N-z_i-n+1} \binom{N-z_i}{n} / \binom{N}{n} \right\}^2 \leq$$

$$\left\{ \sum_{i=1}^{S} \binom{N-z_i}{n} / \binom{N}{n} \right\} \times \left\{ \sum_{i=1}^{S} \left(\frac{nz_i}{N-z_i-n+1} \right)^2 \binom{N-z_i}{n} / \binom{N}{n} \right\},$$

dove vale il segno di uguaglianza quando tutte le z_i sono uguali.

La parte sinistra della disuguaglianza è $\{\mathbb{E}(f_1)\}^2$ e la prima somma della parte destra è $\{\mathbb{E}(f_0)\}$. Per quanto riguarda la seconda somma di destra possiamo riscrivere:

$$\left(\frac{nz_i}{N-z_i-n+1} \right)^2 = \frac{n}{n-1} \left(\frac{n(n-1)z_i(z_i-1)}{(N-z_i-n+1)^2} \right) + \frac{n^2 z_i}{(N-z_i-n+1)^2}.$$

Dunque la seconda somma diventa:

$$\left\{ \sum_{i=1}^{S} \left(\frac{nz_i}{N-z_i-n+1} \right)^2 \binom{N-z_i}{n} / \binom{N}{n} \right\} \approx \frac{2n}{n-1} \mathbb{E}(f_2)$$

$$+ \sum_{i=1}^{S} \left[\frac{n}{N-z_i-n+1} \right] \frac{nz_i}{N-z_i-n+1} \binom{N-z_i}{n} / \binom{N}{n}$$

Il contributo delle specie con z_i grande all'ultimo termine dell'equazione sopra è trascurabile, per le specie con z_i molto più piccolo di N, abbiamo:

$$\frac{n}{N-z_i-n+1} = \frac{n/N}{(N-z_i-n+1)/N} \approx \frac{n/N}{1-n/N} = \frac{p^*}{1-p^*}.$$

Quindi otteniamo la seguente disuguaglianza:

$$\{\mathbb{E}(f_1)\}^2 \leq \{\mathbb{E}(f_0)\} \left\{ \frac{n}{n-1} 2\mathbb{E}(f_2) + \frac{p^*}{1-p^*} \mathbb{E}(f_1) \right\},$$

che è equivalente a:

$$\mathbb{E}(f_0) \geq \frac{\mathbb{E}(f_1^2)}{\frac{n}{n-1} 2\mathbb{E}(f_2) + \frac{p^*}{1-p^*} \mathbb{E}(f_1)}. \tag{4.28}$$

Sostituendo il valore di aspettazione con le frequenze osservate otteniamo come limite inferiore per la ricchezza delle specie:

$$S_{p=1} = S^* + \frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{p^*}{1-p^*}f_1}. \tag{4.29}$$

5 CONCLUSIONI

Part III

APPENDIX

APPENDIX TEST

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

More dummy text.

A.1 APPENDIX SECTION TEST

Test: Table A.1 (This reference should have a lowercase, small caps A if the option `floatperchapter` is activated, just as in the table itself → however, this does not work at the moment.)

LABITUR BONORUM PRI NO	QUE VISTA	HUMAN
fastidii ea ius	germano	demonstratea
suscipit instructior	titulo	personas
quaestio philosophia	facto	demonstrated

Table A.1: Autem usu id.

A.2 ANOTHER APPENDIX SECTION TEST

Equidem detraxit cu nam, vix eu delenit periculis. Eos ut vero constituto, no vidit propriae complectitur sea. Diceret nonummy in has, no qui eligendi recteque consetetur. Mel eu dictas suscipiantur, et sed placerat oporteat. At ipsum electram mei, ad aeque atomorum mea. There is also a useless Pascal listing below: Listing A.1.

Listing A.1: A floating example (`listings` manual)

```
for i:=maxint downto 0 do
begin
{ do nothing }
end;
```

DECLARATION

Put your declaration here.

Padova, Marzo 2019

Nicola Marinello

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Thank you very much for your feedback and contribution.