Species diversity

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

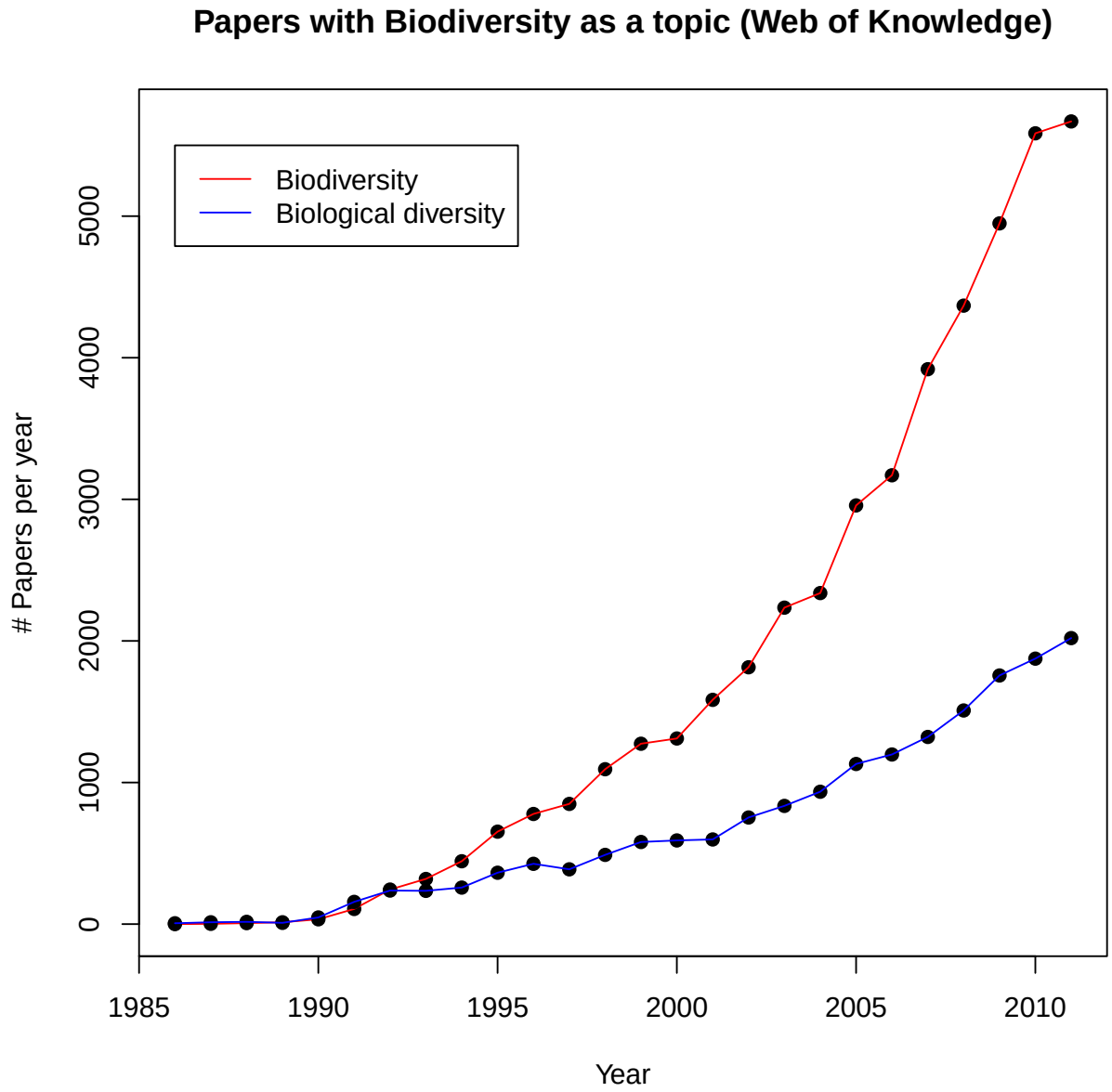
March 15, 2018

Contents

Statistics for the life sciences

1. Statistics and Bioinformatics
2. Statistics and Biodiversity
3. Statistics and Health science

Scientific papers about Biodiversity



What is Biodiversity?

Definition of the United Nations Environment Program:

"Biological diversity means the variability among living organisms from all sources, inter alia, terrestrial, marine and other aquatic systems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems."

Biodiversity is, according to the definition, observed at different levels:

- Genetic diversity (within-species diversity; diversity of genes)
- Species diversity (species richness and abundance)
- Ecological diversity (diversity of communities)

In this part of the course we focus only on **species diversity** and **genetic diversity**.

Outline block 2: Statistics and Biodiversity (1/2)

Species diversity:

- What is Biodiversity?
- How to measure Biodiversity?
- Basic concepts in ecology (species abundance, species evenness, assemblage, community, ecosystem, niche, habitat, biomass, ...)
- Graphics for representing abundance data: rank-abundance plot, abc curve, frequency-abundance plot.
- Statistical models for abundance: Fisher's log series, the log normal model, MacArthur's broken stick model.
- Diversity measures: Species richness, diversity indices α & β, Shannon's index, Simpson's index.
- Comparing diversity estimates.

Outline block 2: Statistics and Biodiversity (2/2)

Genetic diversity:

- Basic concepts in genetics (genetic markers: RFLPs, microsatellites, SNPs, allele frequency, dominant, co-dominant and recessive markers).
- Genetic equilibria: Hardy-Weinberg equilibrium & linkage disequilibrium. Wahlund effect.
- Genetic diversity: allelic richness, percentage of polymorphic loci, heterozygosity, gene diversity, genotypic diversity, Shannon's index, Simpson's index. Wright's F statistics, Nei's gene diversity indices, genetic diversity within and between populations.

Focus of block 2

- Theory: study of biological concepts and the related statistics.
- Practice: analysis of datasets of species counts and genetic markers with R and other specialized software.

Species diversity

- Species diversity refers to the "variety and abundance of species in a defined unit of study".
- Species diversity comprises two things:
  - Species richness (# of species in an area)
  - Species evenness (the relative abundance of the species)
- Basic assumptions underlying diversity measurement:
  - All species are equal (only abundance is used to weight species).
  - All individuals are equal
  - Abundance is measured in appropriate, comparable units.

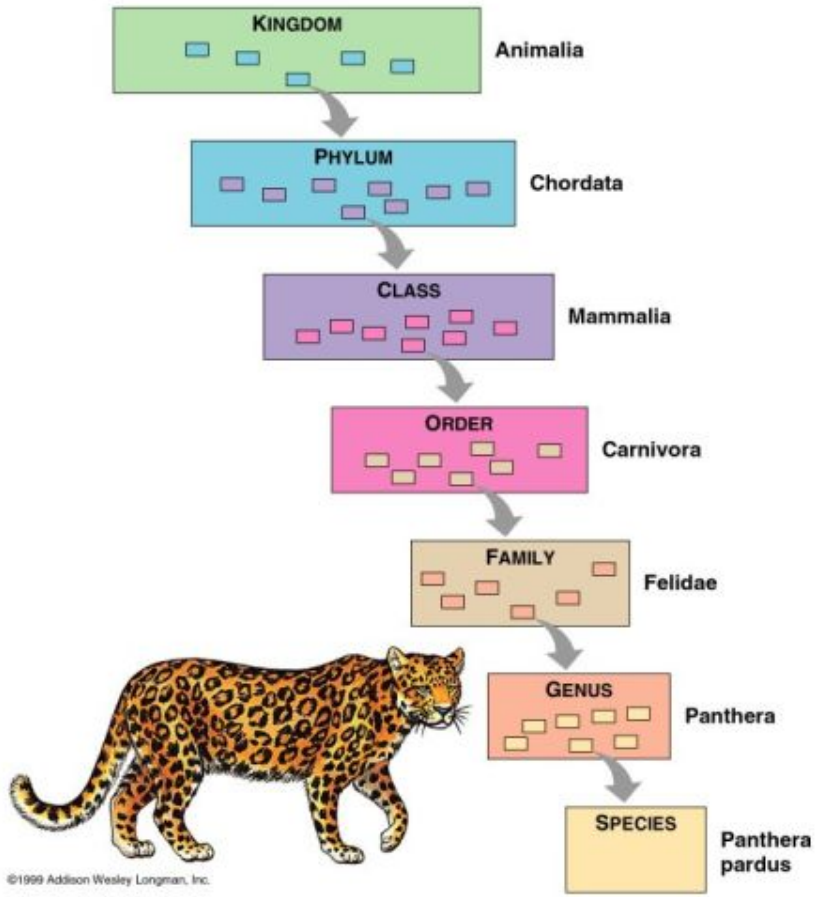
What is meant with a species?

- Individual organisms are considered to be of the same species if they are capable of interbreeding and producing fertile offspring.
- The precise definition of a species is a problem in biology, and different criteria are used, depending on the purpose of the study.
- In microbiology, the first definition is highly problematic.
- If two individuals are considered to be of the same species or not depends largely on the extent to which they share and can interchange genetic material.

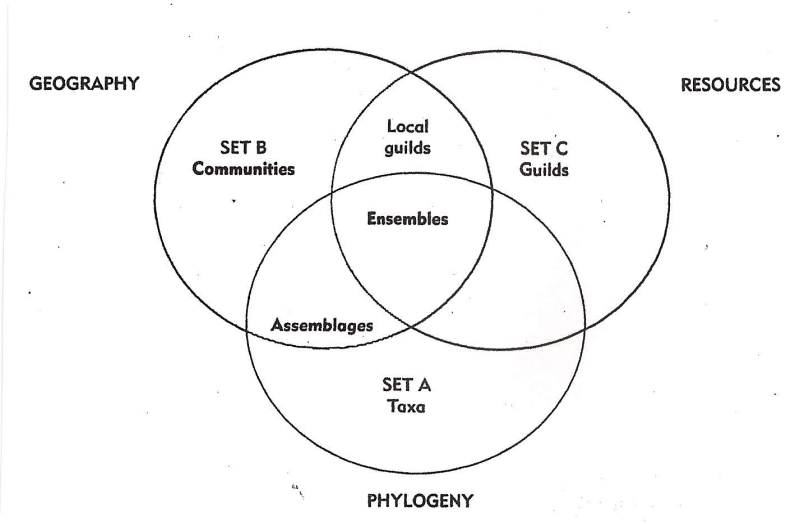
A bit of taxonomy

- Species are classified into groups known as **taxa**.
- A hierarchy of classifications has been created, joining similar **species** into a **genus**, similar genus into a **family**, and so on.
- The basic levels of the classification system are: **species, genus, family, order, class, phylum** and **kingdom**.
- Intermediate levels (e.g. suborder, superorder) are also used.
- There are 5 kingdoms: Animalia, Plantae, Fungi, Protista, Archaea, and Bacteria.

Example


KINGDOM
Animalia
PHYLUM
Chordata
CLASS
Mammalia
ORDER
Carnivora
FAMILY
Felidae
GENUS
Panthera
SPECIES
Panthera pardus
©1999 Addison Wesley Longman, Inc.

The scale of a biodiverstity study



Some terminology:

- Community: collection of species occurring at a certain time and place
- Guild: group of species exploiting the same resources
- Assemblage: group of phylogenetically related species.
- Ensemble: interacting species that share ancestry and resources.

Definitions may vary across authors.

Purpose of a biodiversity study

Biodiversity studies are often of comparative nature, and the purpose of the study is often to compare or rank communities, or to asess if diversity has changed over time.

Diversity measures

A wide range of diversity measures has been proposed.

- Species richness
- Shannon's index (H')
- Simpson's index (D)
- Parameters estimated from parametric models
  - Fisher's α
  - k of a geometric series.
  - ....
- Parametric and non-parametric Species richness estimates (S_{max}, S_{chao}).
- ...

Shannon index (H')

1	2	3	4	$\cdots$	S
n_1	n_2	n_3	n_4	$\cdots$	n_S

$$p_i = \frac{n_i}{N} \qquad H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

Notes:

- S is the total number of categories, N the total count $N = \sum_{i=1}^{S} n_i$.
- $H' \geq 0$
- H' has its origin in information theory, and is widely used in many contexts.
- For empirical ecological data, H' often varies from 1.5 to 3.5.
- $H_{max} = \ln(S)$.
- Abundance 0 does not contribute.
- The maximum possible value, H_{max}, occurs when all species are equally abundant.
- $V(H') = \frac{\sum_{i=1}^{S} p_i(\ln(p_i))^2 - \left(\sum_{i=1}^{S} p_i \ln(p_i)\right)^2}{N} + \frac{S-1}{N^2}$
- Has a normal distribution for large samples.
- Captures species richness and evenness.
- Shannon evenness measure

$$J' = \frac{H'}{H_{max}} = \frac{H'}{\ln(S)}$$

Simpson's index (D)

1	2	3	4	$\cdots$	S
n_1	n_2	n_3	n_4	$\cdots$	n_S

$$D = \sum p_i^2 \text{ for a infinitely large community}$$

$$D = \sum \frac{n_i(n_i - 1)}{N(N-1)} \text{ for a finite community}$$

- The larger D, the less diversity.
- Typically $1 - D$ or $1/D$ is used as an index of diversity.
- Abundant species make a large contribution to the index.
- $1 \leq 1/D \leq S$.
- A related popular evenness measure

$$E_{1/D} = \frac{(1/D)}{S}$$

- $0 < E_{1/D} < 1$

To get some idea ...

Sample	Sp1	Sp2	Sp3	Sp4	Sp5	N	S	H'	D	$1/D$
1	10	10	10	10	10	50	5	1.609	0.184	5.444
2	50	0	0	0	0	50	1	0.000	0.000	1.000
3	49	1	0	0	0	50	2	0.098	0.960	1.042
4	48	1	1	0	0	50	3	0.196	0.921	1.086
5	30	10	5	3	2	50	5	1.156	0.403	2.480

Example: bird counts at three sites in Ireland

	Species	Derrycunniy.oakwood	Muckross.yew.wood	Sitka.spruce.plot
1	Chaffinch	35	9	14
2	Robin	26	20	10
3	Blue tit	25	10	0
4	Goldcrest	21	21	30
5	Wren	16	5	4
6	Coal tit	11	14	6
7	Spotted flycatcher	6	0	0
8	Tree creeper	5	3	0
9	Siskin	3	2	7
10	Blackbird	3	6	3
11	Great tit	3	9	0
12	Long-tailed tit	3	2	0
13	Woodpigeon	3	0	0
14	Hooded crow	2	0	0
15	Woodcock	2	0	0
16	Song thrush	2	6	0
17	Redstart	1	0	0
18	Mistle thrush	1	0	0
19	Dunnock	1	0	0
20	Sparrow hawk	1	1	0
21	Long-eared owl	0	1	0
22	Jay	0	1	0
23	Chiff chaff	0	0	1
S		20	15	8
N		170	110	75
H'		2.41	2.35	1.71
J'		0.804	0.866	0.825
D		0.115	0.109	0.222
$1/D$		8.717	9.181	4.505
$E_{1/D}$		0.436	0.612	0.563

Uncertainty in the Shannon index

- 95% confidence intervals for H'

Site	CI
Derrycunniy.oakwood	(2.26,2.55)
Muckross.yew.wood	(2.20,2.49)
Sitka.spruce.plot	(1.52,1.91)

- Two sites do not seem to differ in diversity.
- Relies on asymptotic normality, may be unrealistic.

Plotting species abundance data in a biodiversity study

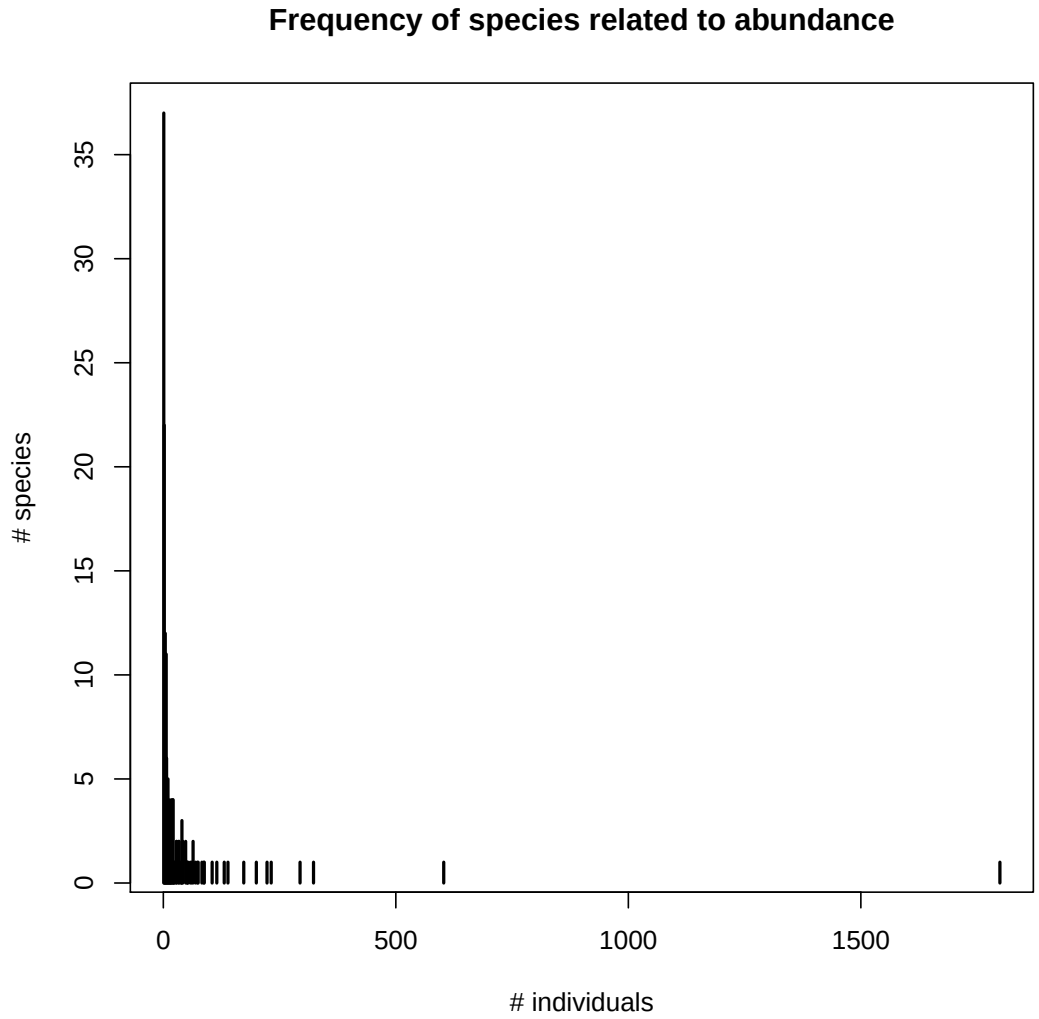
- Frequency distribution (number of species against abundance)
- Rank-abundance plot (Whittaker plot)
- ...

Example data set: *Macrolepidoptera*

Number individuals per species caught in light traps in the UK.

individuals	species	individuals	species
1	37	39	1
2	22	40	3
3	12	42	2
4	12	48	2
5	11	51	1
6	11	52	1
7	6	53	1
8	4	58	1
9	3	61	1
10	5	64	2
11	2	69	1
12	4	73	1
13	2	75	1
14	3	83	1
15	2	87	1
16	2	88	1
17	4	105	1
18	2	115	1
19	4	131	1
20	4	139	1
21	1	173	1
22	1	200	1
23	1	223	1
25	2	232	1
28	2	294	1
29	2	323	1
33	2	603	1
34	2	1799	1
38	1		

Frequency distribution



Identifies outlying abundant species as well as the number of rare species

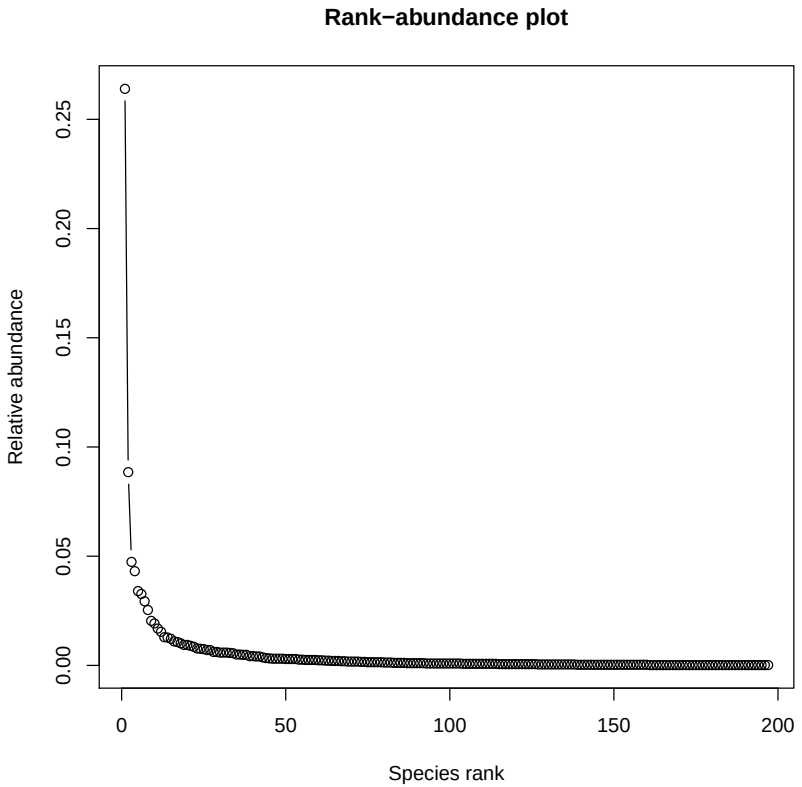
R-code for making a frequency distribution

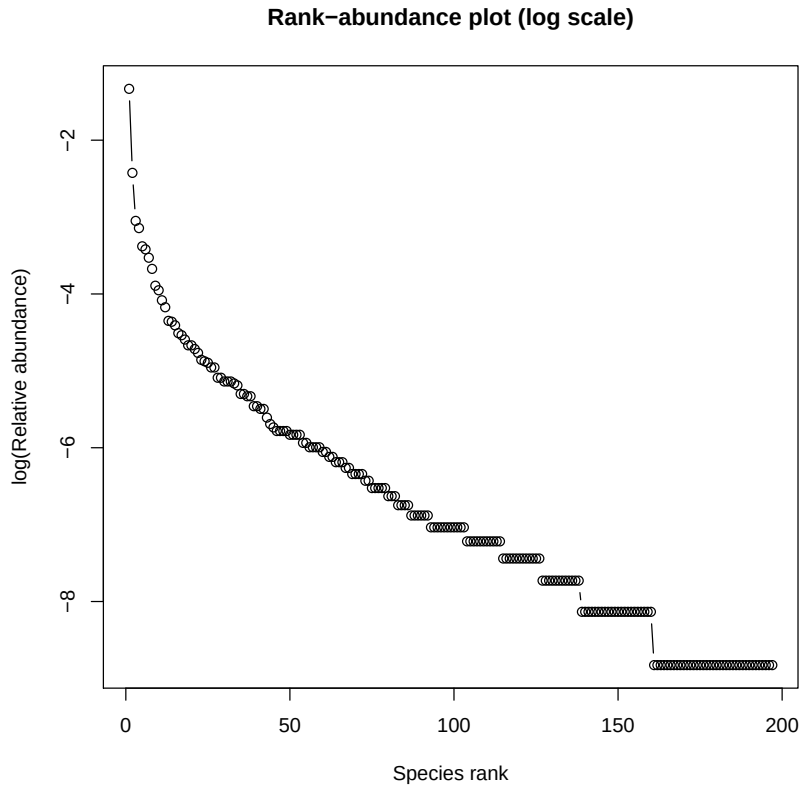
```
> X <- read.csv("http://www-eio.upc.es/~jan/data/biodiv/Lewis.csv",sep=";")
> head(X)
  individuals species
1           1      37
2           2      22
3           3      12
4           4      12
5           5      11
6           6      11
> S <- sum(X[,2])
> S
[1] 197
> N <- sum(X[,1]*X[,2])
> N
[1] 6815
> plot(X$individuals,X$species,type="h",lwd=2,xlab="# individuals",
       ylab="# species",main="Frequency of species related to abundance")
```

Rank-abundance plot

- A rank-abundance plot is a widely used graphical tool in ecological diversity studies.
- Species are ordered from most abundant to least abundant on the horizontal axis.
- The number of individuals per species is plotted on the vertical axis, often in a logarithmic scale

Rank-abundance plot *Macrolepidoptera*





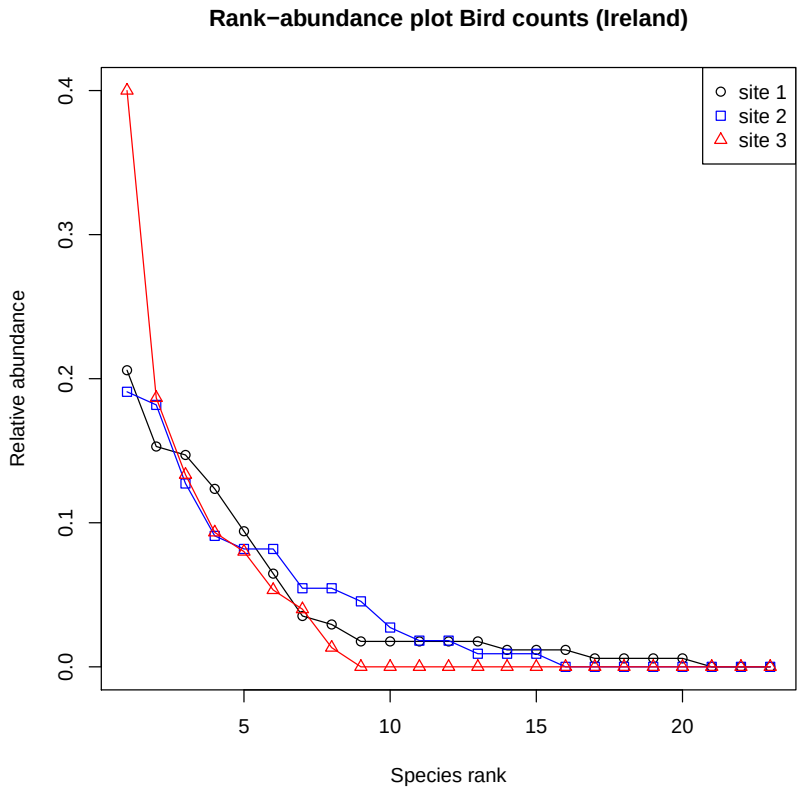
R-code for making a Rank-abundance plot

```
> abundance <- rep(X$individuals,times=X$species)
> N <- sum(abundance)
> N
[1] 6815
> S <- length(abundance)
> S
[1] 197
> index <- order(abundance, decreasing=TRUE)
> rel.abundance.sorted <- abundance[index]/N
> speciesrank <- 1:S
> plot(speciesrank,rel.abundance.sorted,type="b") # in a relative scale
>
> plot(speciesrank,log(rel.abundance.sorted),type="b") # in the log scale
>
> library(biodiversityR)
> abundance.sorted <- abundance[index]
> Z <- data.frame(1:S,abundance.sorted)
> rankabunplot(Z)                                  # in abundance scale
```

Stratification

A rank-abundance plot can be stratified, if, for instance,

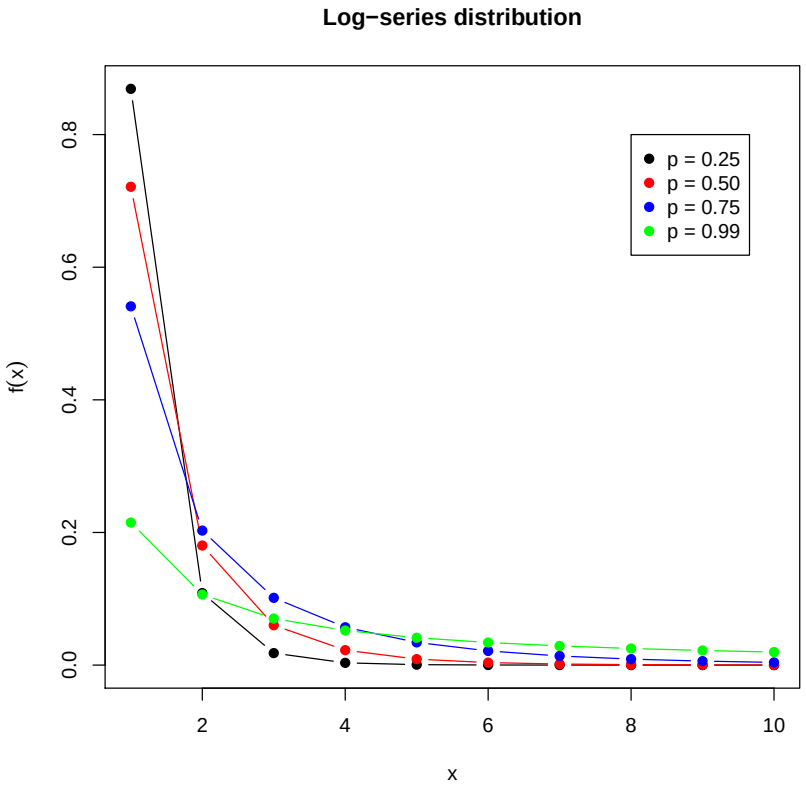
- The same community has been sampled on several occasions (over time)
- Comparable communities have been sampled at different locations



Models for species abundance

- Fisher's log-series model
- Geometric series
- MacArthur's broken stick model
- The log-normal model (truncated or not)
- ...

Logarithmic distribution (log-series distribution)



This distribution is a model for the frequency distribution.

$$f(x) = \frac{-1}{\ln(1-p)} \frac{p^x}{x} \quad x = 1, 2, 3, \ldots \quad 0 < p < 1$$

x	1	2	3	4	$\cdots$	n
$f(x)$	kp	$k\frac{p^2}{2}$	$k\frac{p^3}{3}$	$k\frac{p^4}{4}$	$\cdots$	$k\frac{p^n}{n}$

$$k = \frac{-1}{\ln(1-p)}$$

We define

- X = the number of species of which x individuals are observed ($X \sim f(x, p)$)
- N = total number of individuals
- S = total number of species

$$\alpha = Sk = \frac{-S}{\ln(1-p)}$$ is a popular measure of diversity

if $p \approx 1$ then α is the number of singletons (species occurring only once)

This model was proposed by Fisher (1943).

Log-series distribution

- Probability function

$$f(x) = \frac{-1}{\ln(1-p)} \frac{p^x}{x} \quad 0 < p < 1 \quad x = 1, 2, 3, \ldots$$

- Expectation and variance

$$E(X) = \frac{-1}{\ln(1-p)} \frac{p}{1-p} \qquad V(X) = -p \frac{p + \ln(1-p)}{(1-p)^2 (\ln(1-p))^2}$$

- Estimation of p by the method of moments:

$$\frac{-1}{\ln(1-\hat{p})} \frac{\hat{p}}{1-\hat{p}} = \overline{X}$$

- There is no closed form solution for $\hat{p}$
- $\overline{X}$ is the average number of individuals per species:

$$\overline{X} = \frac{N}{S}$$

- In empirical studies $\hat{p}$ is almost always > 0.9.

Estimating Fisher's α numerically

Roots of a non-linear equation $f(\alpha) = 0$ can be found by:

$$\hat{\alpha}_{n+1} = \hat{\alpha}_n + h_n \qquad h_n = -\frac{f(\hat{\alpha}_n)}{f'(\hat{\alpha}_n)}$$

An initial estimate for the parameter of interest is needed.
For the Macrolepidoptera:

$$S = 197 \qquad N = 6815 \qquad \overline{X} = 34.59$$

It.	h	p
0	-	0.99000000
1	0.00768267	0.99768267
2	-0.00141800	0.99626467
3	-0.00124517	0.99501951
4	-0.00049952	0.99451999
5	-0.00005060	0.99446939
6	-0.00000044	0.99446895

$$\hat{p} = 0.994469$$

$$\hat{\alpha} = \frac{-S}{\ln(1-\hat{p})} = \frac{-197}{\ln(1-0.994469)} = 37.904$$

Note this is close to the observed number of singletons (37).

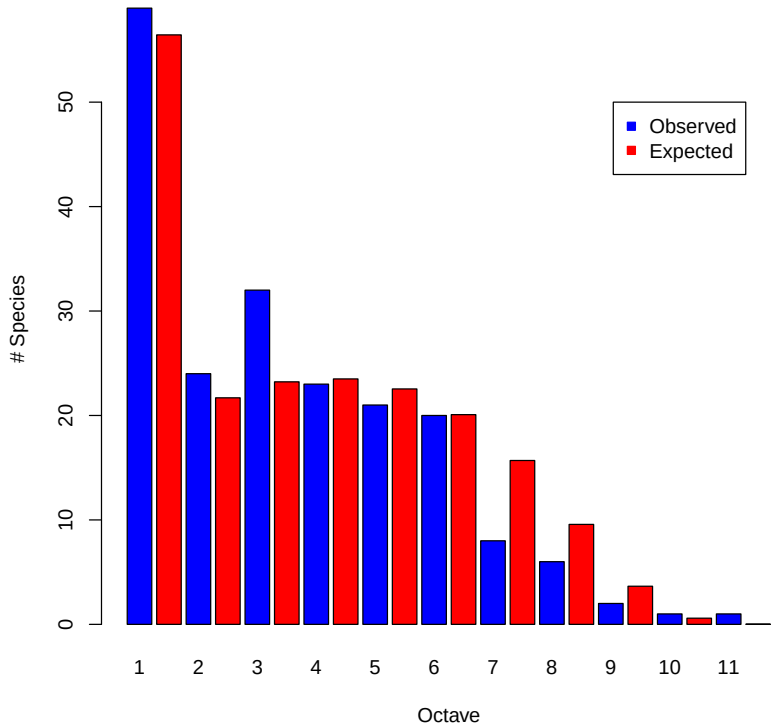
Estimating Fisher's α numerically in R

```
> library(vegan)
> X <- read.table("Macrolepidoptera.txt")
> head(X)
  individuals species
1           1      37
2           2      22
3           3      12
4           4      12
5           5      11
6           6      11
> y <- rep(X[,1],X[,2])
> y
  [1]    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
 [16]    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
 [31]    1    1    1    1    1    1    1    2    2    2    2    2    2    2    2
 [46]    2    2    2    2    2    2    2    2    2    2    2    2    2    2    3
 [61]    3    3    3    3    3    3    3    3    3    3    3    4    4    4    4
 [76]    4    4    4    4    4    4    4    4    5    5    5    5    5    5    5
 [91]    5    5    5    5    6    6    6    6    6    6    6    6    6    6    6
....
[181]   73   75   83   87   88  105  115  131  139  173  200  223  232  294  323
[196]  603 1799
> length(y)
[1] 197
> fisher.alpha(y)
[1] 37.90375
> fisherfit(y)

Fisher log series model
No. of species: 197
Fisher alpha:   37.90375
```

Goodness-of-fit

In order to compare observed and expected values under the log-series model, observations need to be grouped. We do this by creating **octaves**.

Nr	Octave	Observed	Expected
1	(0,2.5]	59	56.44
2	(2.5,4.5]	24	21.69
3	(4.5,8.5]	32	23.22
4	(8.5,16.5]	23	23.50
5	(16.5,32.5]	21	22.54
6	(32.5,64.5]	20	20.08
7	(64.5,128]	8	15.69
8	(128,256]	6	9.57
9	(256,512]	2	3.65
10	(512,1.024]	1	0.59
11	(1.024,—]	1	0.02



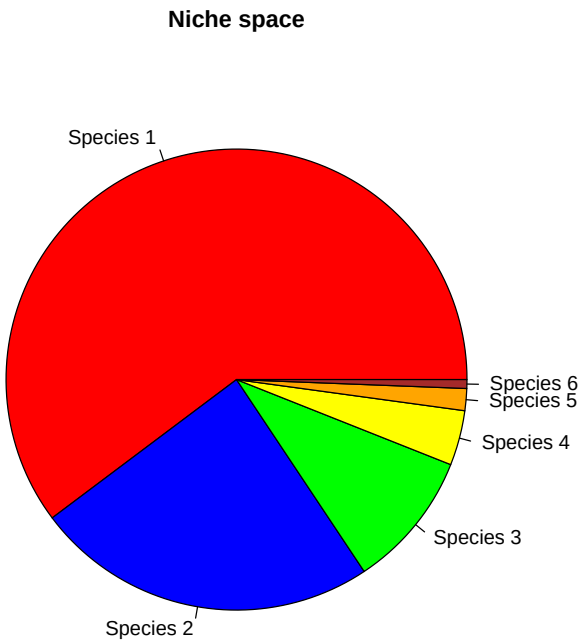
$$X^2 = \frac{(59 - 56.44)^2}{56.44} + \cdots + \frac{(4 - 4.26)^2}{4.26} = 8.91 \qquad \text{p-value} = P\left(\chi^2_7 \geq 8.91\right) = 0.259$$

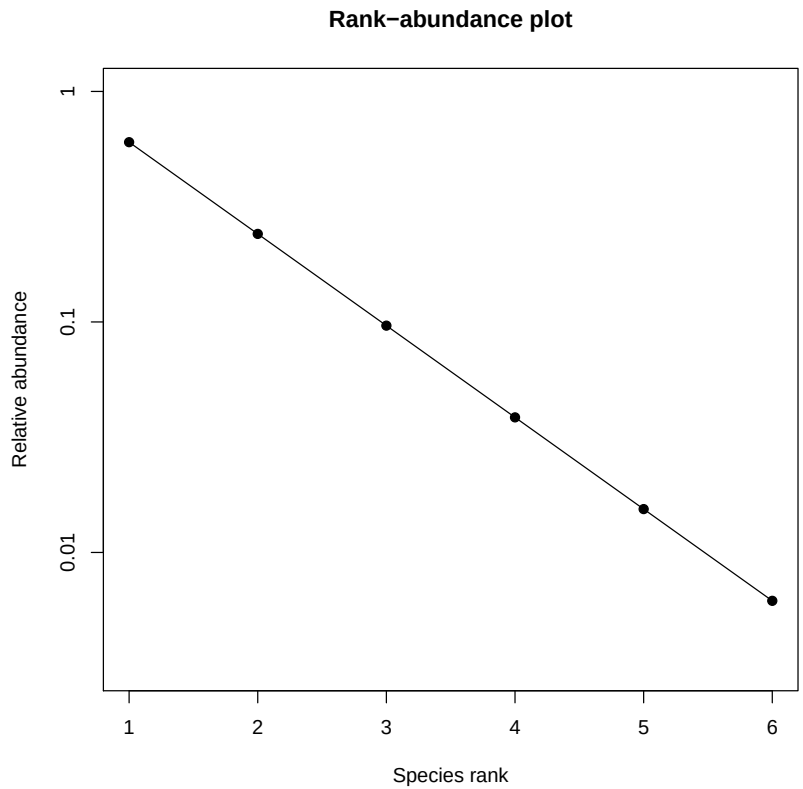
Niche and habitat

- **niche**: the way a species makins a living ("profession") role a species plays in a community.
- **fundamental niche**: what biotic a abiotic circumstances potentially allow a species to do
- **realized niche**: what the species actually does

The geometric series (niche pre-emption hypothesis 1/2)

- There is some resource which is limiting.
- The most dominant species takes fraction k of this resource.
- The second most dominant species takes fraction k of the remaining resource, and so on.





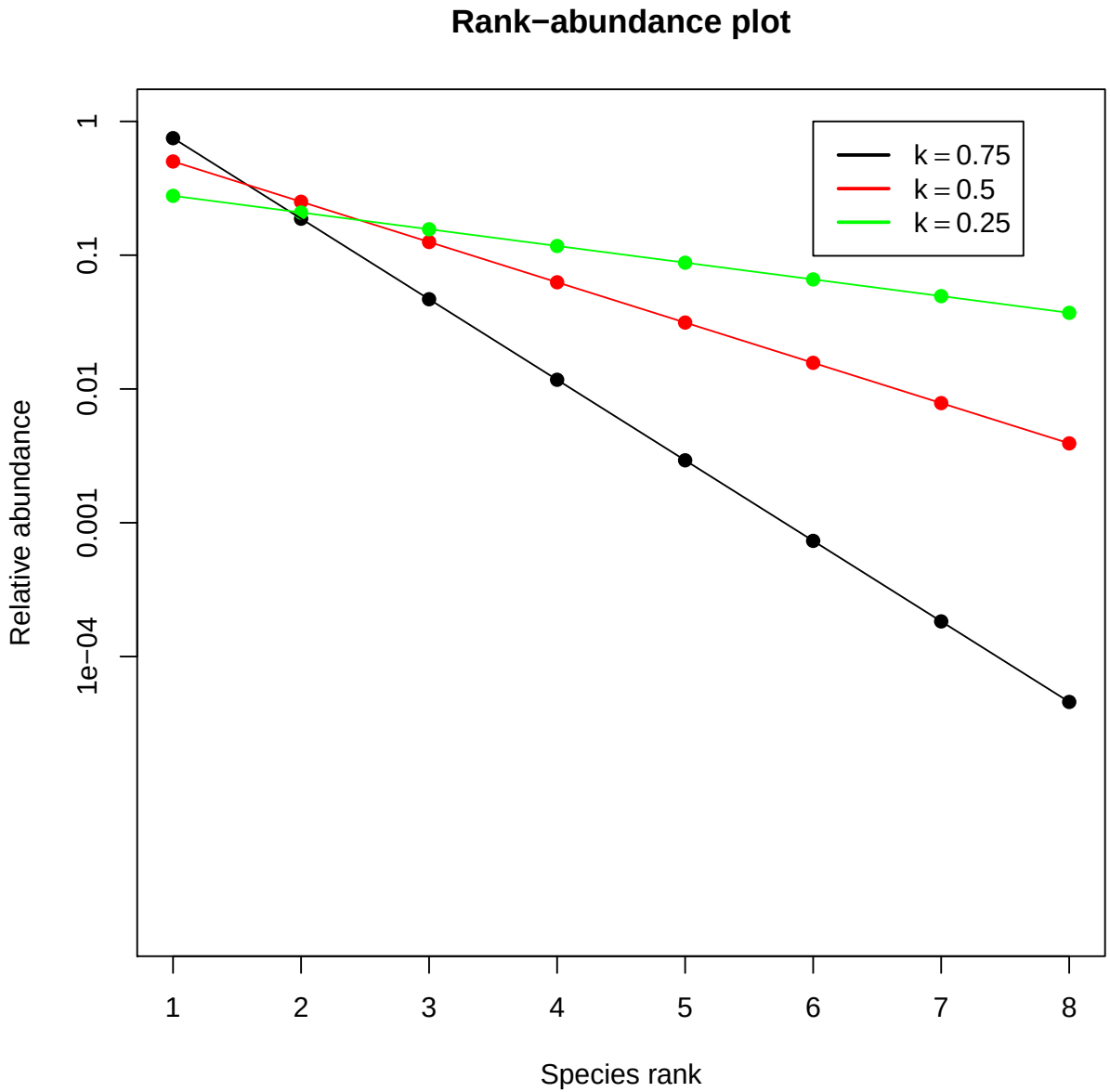
The geometric series (niche pre-emption hypothesis 2/2)

- The geometric series is a model for poor-species assemblages.
- The abundance is proportional to the niche a species occupies.
- $$ck, ck(1-k), ck(1-k)^2, \ldots, ck(1-k)^{S-1} \quad 0 < k < 1$$
- Model for the abundance of the species

$$n_i = \frac{Nk(1-k)^{i-1}}{1-(1-k)^S} \qquad i = 1, \ldots, S.$$

- To fit the geometric series to a data set, parameter k must be estimated.
- Note that $\ln(n_i)$ is linear in i, with slope $\ln(1-k)$.
- Thus, k can be estimated by linear regression.
- Alternative estimators of k have been described in the literature.

Diversity and the geometric series

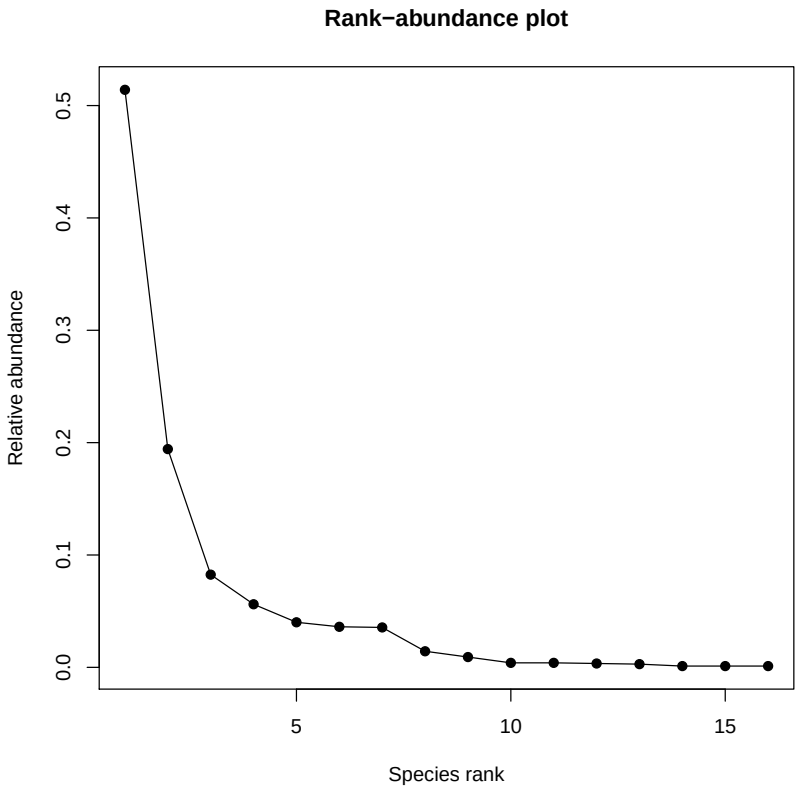


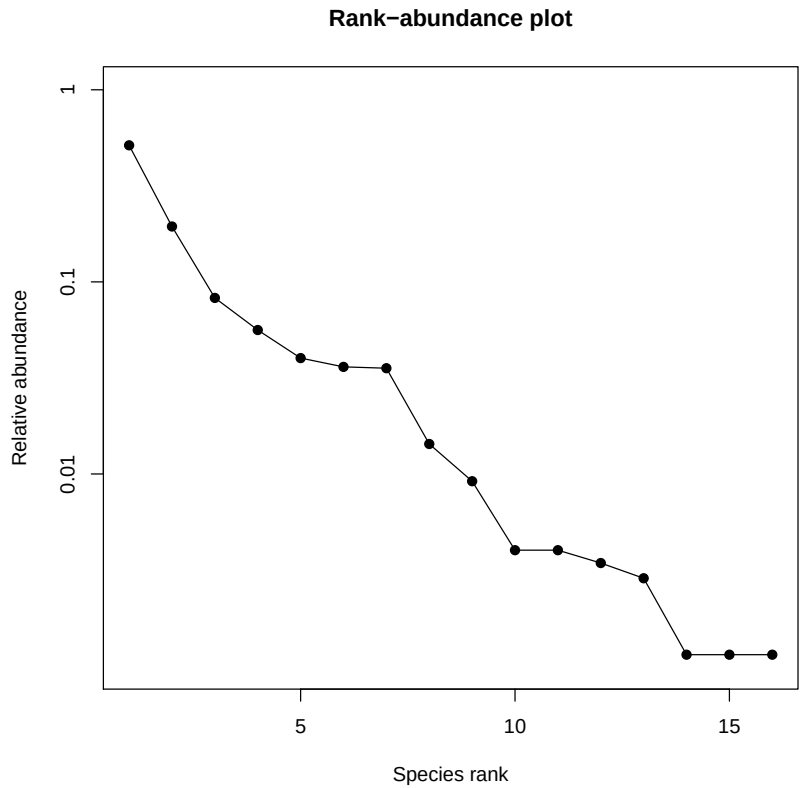
Example: dung beetles

	Species	Abundance
1	Onthophagus truncaticornis	897
2	Caccobius meridionalis	339
3	Onthophagus rectecornutus	144
4	Oniticellus cinctus	98
5	Onitis philemon	70
6	Onthophagus dama	63
7	Drepanocerus setosus	62
8	Caccobius unicornis	25
9	Copris indicus	16
10	Oniticellus spinipes	7
11	Onthophagus tarandus	7
12	Liatongus rhadamistus	6
13	Onthophagus catta	5
14	Onthophagus pactolus	2
15	Onthophagus spinifex	2
16	Sisyphys sp.	2

$N = 1745 \qquad S = 16$

Rank-Abundances plots dung beetles





Regression results (natural log)

```
x <- 1:16
lny <- log(y)

Call:
lm(formula = lny ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59334 -0.26063 -0.03756  0.16880  0.72185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.47031    0.18910   34.22 6.78e-15 ***
x           -0.39311    0.01956  -20.10 1.00e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3606 on 14 degrees of freedom
Multiple R-squared:  0.9665, Adjusted R-squared:  0.9641
F-statistic:   404 on 1 and 14 DF,  p-value: 1.004e-11
```

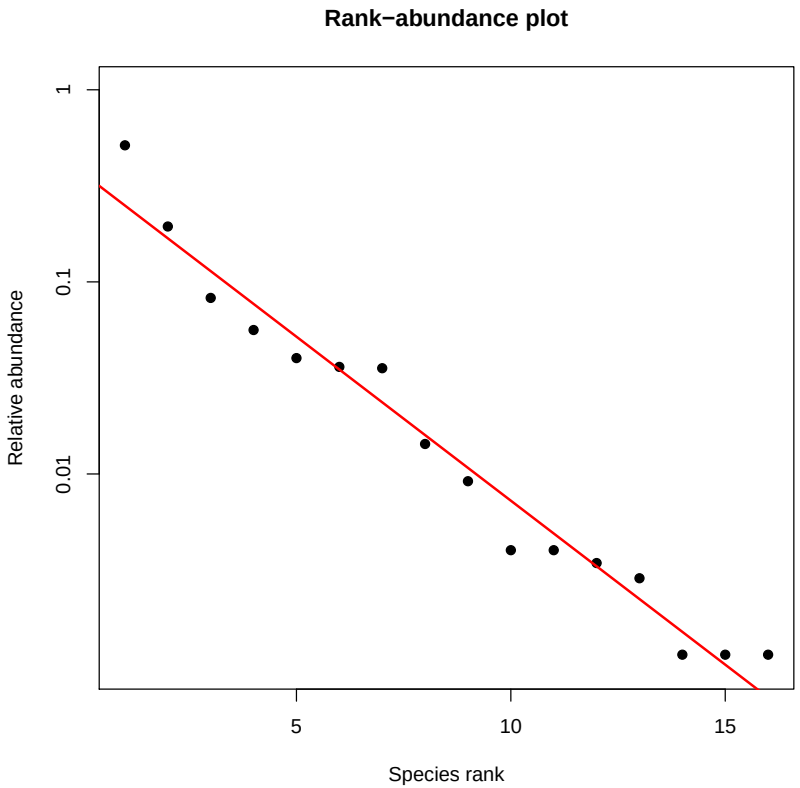
$$\ln(1-k) = -0.39311 \rightarrow k = 1 - e^{-0.39311} = 0.325$$

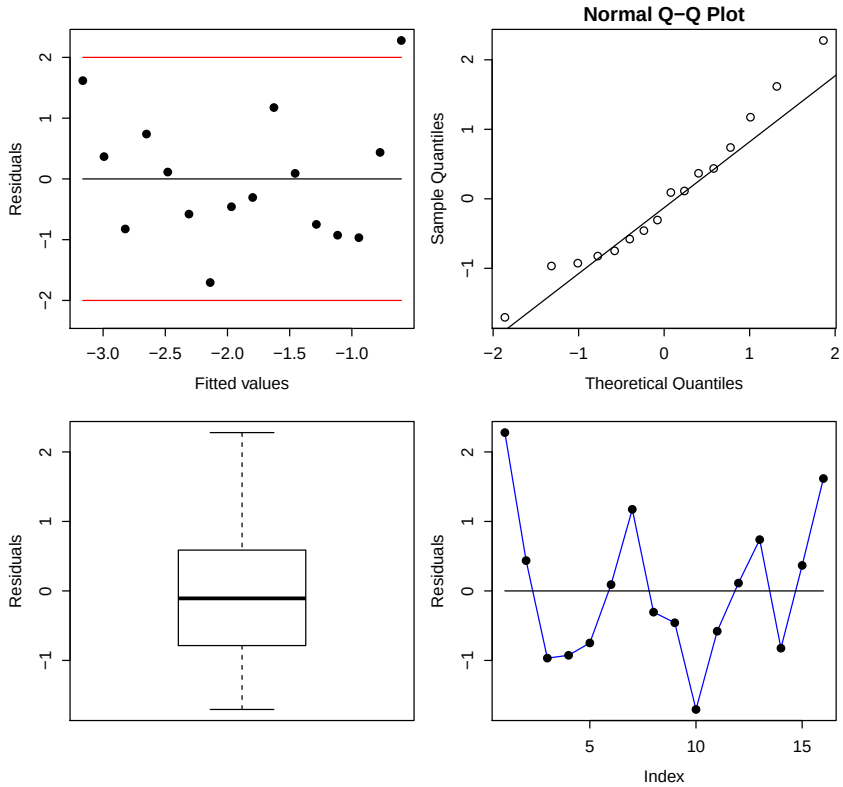
Regression results (log 10)

```
> x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
> log10y <- log10(y)
> out.log10 <- lm(log10y~x)
> summary(out.log10)

Call:
lm(formula = log10y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25768 -0.11319 -0.01631  0.07331  0.31350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.810019   0.082127   34.22 6.78e-15 ***
x           -0.170724   0.008493  -20.10 1.00e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1566 on 14 degrees of freedom
Multiple R-squared:  0.9665, Adjusted R-squared:  0.9641
F-statistic:   404 on 1 and 14 DF,  p-value: 1.004e-11
```

$$\log(1-k) = -0.170724 \rightarrow k = 1 - 10^{-0.170724} = 0.325$$

Fitted model and residuals





Broken stick model (random niche boundary hypothesis)

- Some resource is represented by a unit length stick.
- The stick is broken at random into S pieces.
- The abundance of a species is proportional to the length of a piece.
- Abundance given by

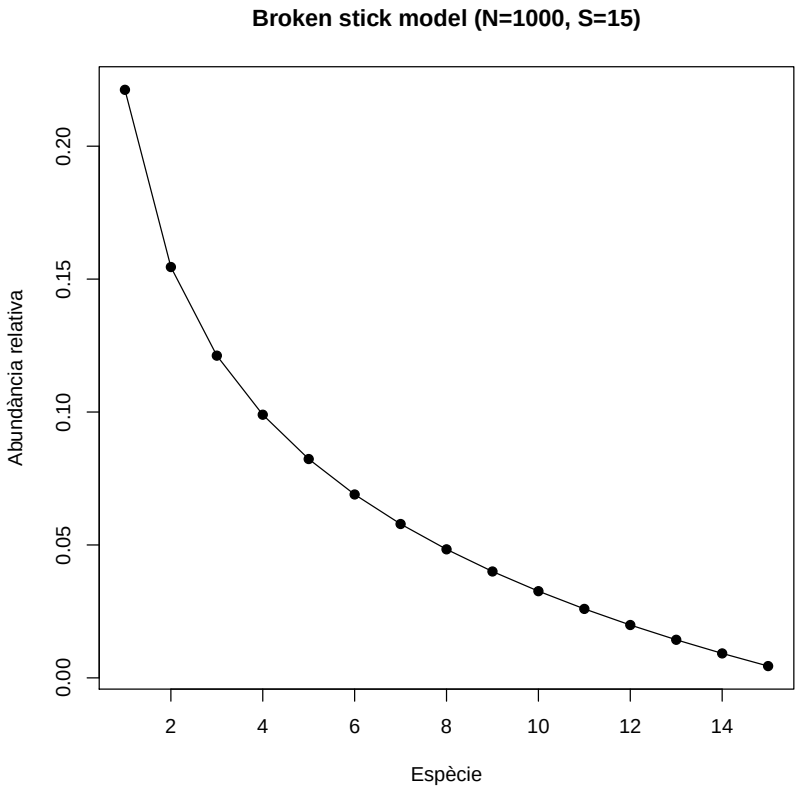
$$n_i = \frac{N}{S} \sum_{j=i}^{S} \frac{1}{j}$$

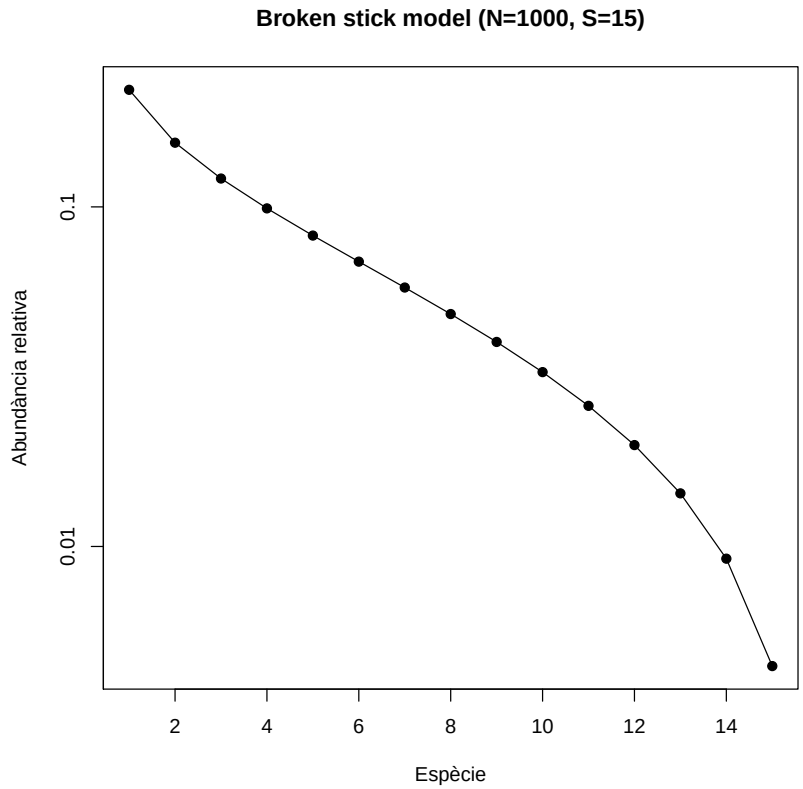
E.g.

$$\begin{aligned}
n_1 &= \frac{N}{S}\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots \frac{1}{S}\right) \\
n_2 &= \frac{N}{S}\left(\frac{1}{2} + \frac{1}{3} + \cdots \frac{1}{S}\right) \\
n_3 &= \frac{N}{S}\left(\frac{1}{3} + \cdots \frac{1}{S}\right) \\
&\vdots \quad = \quad \vdots \\
n_S &= \frac{N}{S}\frac{1}{S}
\end{aligned}$$

- Note: there are no parameters to be estimated!

Rank-abundance plot for the Broken Stick model

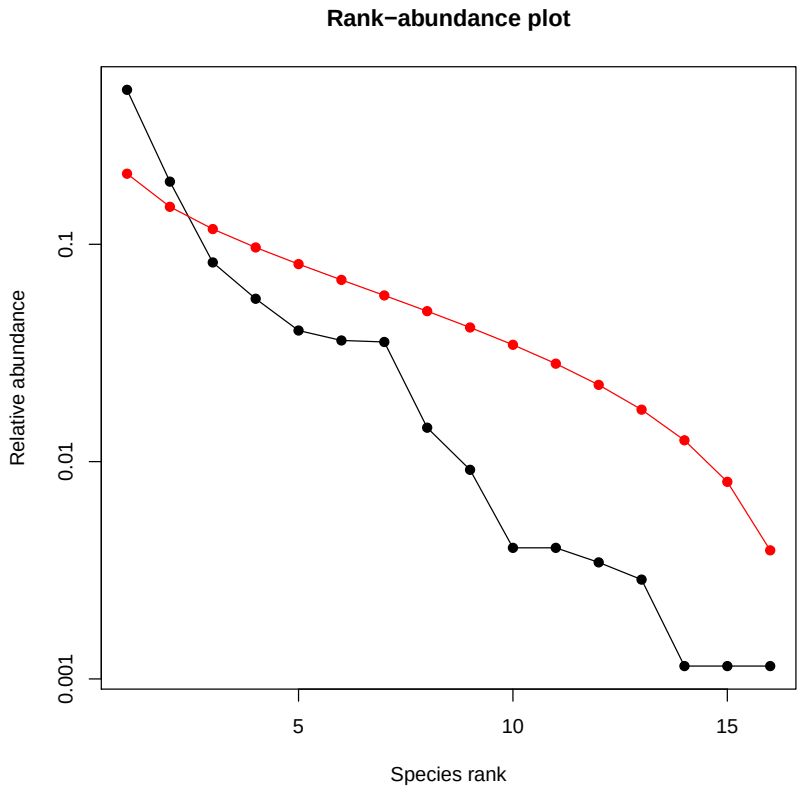




Example: fitting broken stick to dung beetles (1/2)

	Species	Abundance	Fitted
1	Onthophagus truncaticornis	897	368.71
2	Caccobius meridionalis	339	259.65
3	Onthophagus rectecornutus	144	205.12
4	Oniticellus cinctus	98	168.76
5	Onitis philemon	70	141.50
6	Onthophagus dama	63	119.68
7	Drepanocerus setosus	62	101.51
8	Caccobius unicornis	25	85.93
9	Copris indicus	16	72.29
10	Oniticellus spinipes	7	60.18
11	Onthophagus tarandus	7	49.27
12	Liatongus rhadamistus	6	39.36
13	Onthophagus catta	5	30.27
14	Onthophagus pactolus	2	21.88
15	Onthophagus spinifex	2	14.09
16	Sisyphys sp.	2	6.82

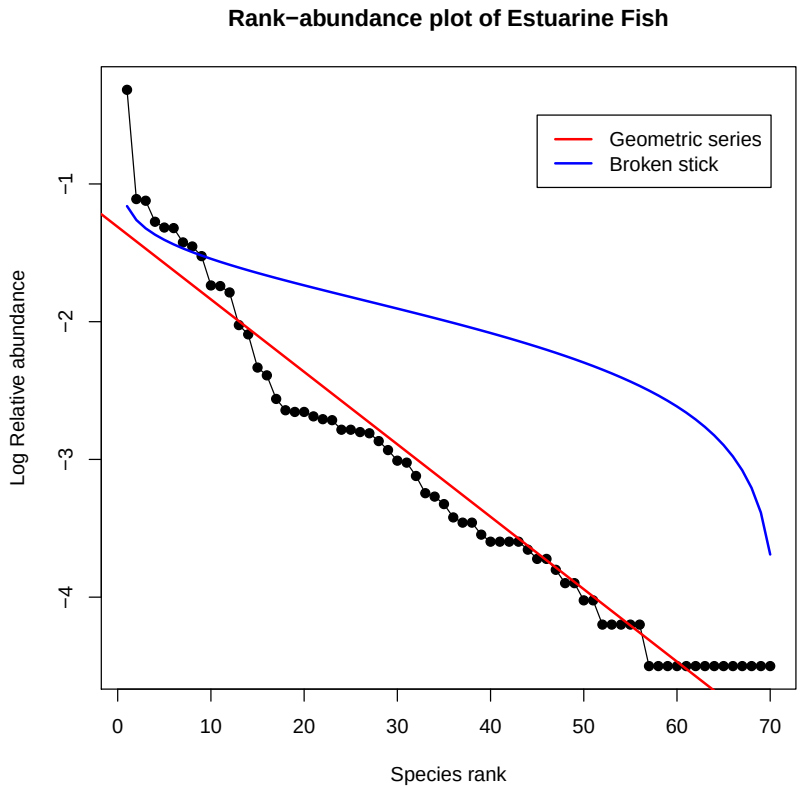
Example: fitting broken stick to dung beetles (2/2)



Fitting the broken stick model in R

```
> install.packages("vegan")
> library(vegan)
> y
 [1] 897 339 144  98  70  63  62  25  16   7   7   6   5   2   2   2
> out <- rad.null(y)
> yhat <- fitted(out)
> yhat
 [1] 368.710756 259.648256 205.117006 168.762839 141.497214 119.684714
 [7] 101.507631  85.927274  72.294461  60.176406  49.270156  39.355383
[13]  30.266841  21.877418  14.087240   6.816406
>
```

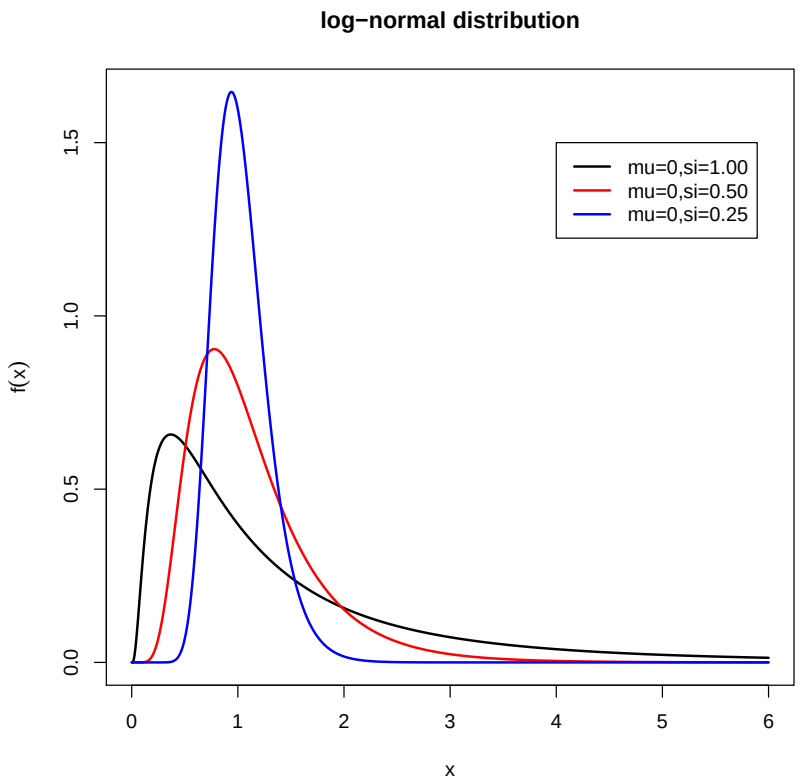
Example: fitting geometric series and broken stick to estuarine fish



Lognormal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{e^{-(\log x - \mu)^2/(2\sigma^2)}}{x}, \quad 0 \leq x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

$$E(X) = e^{\mu + \frac{\sigma^2}{2}}, \quad V(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$$



ML estimators of μ and σ

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(x_i)}{n}$$

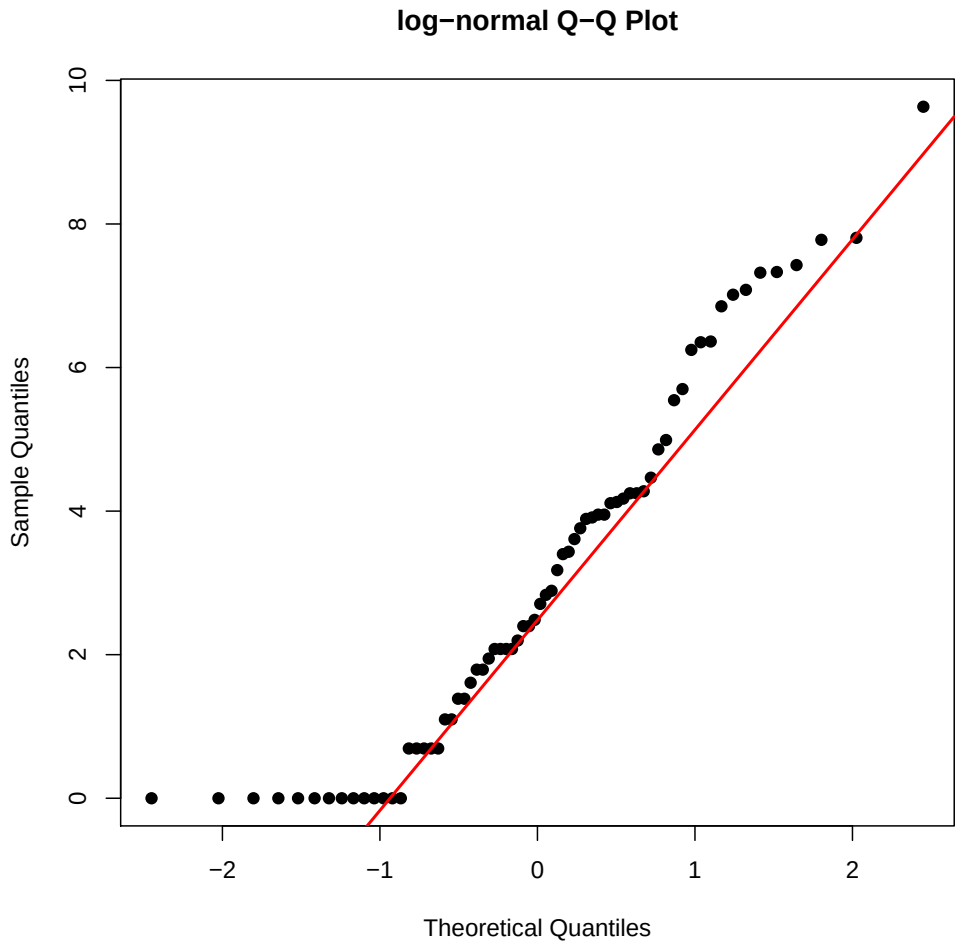
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\ln(x_i) - \hat{\mu})^2}{n}$$

Example: estuarine fish

Nr.	Individuals	Species	Nr.	Individuals	Species
1	1	14	23	61	1
2	2	5	24	62	1
3	3	2	25	65	1
4	4	2	26	70	2
5	5	1	27	72	1
6	6	2	28	87	1
7	7	1	29	129	1
8	8	4	30	147	1
9	9	1	31	256	1
10	11	2	32	299	1
11	12	1	33	516	1
12	15	1	34	574	1
13	17	1	35	580	1
14	18	1	36	947	1
15	24	1	37	1113	1
16	30	1	38	1191	1
17	31	1	39	1513	1
18	37	1	40	1527	1
19	43	1	41	1682	1
20	49	1	42	2391	1
21	50	1	43	2458	1
22	52	2	44	15272	1

$$\hat{\mu} = 3.040781 \qquad \hat{\sigma}^2 = 6.384154$$

Example: estuarine fish



A truncated log-normal is often used.

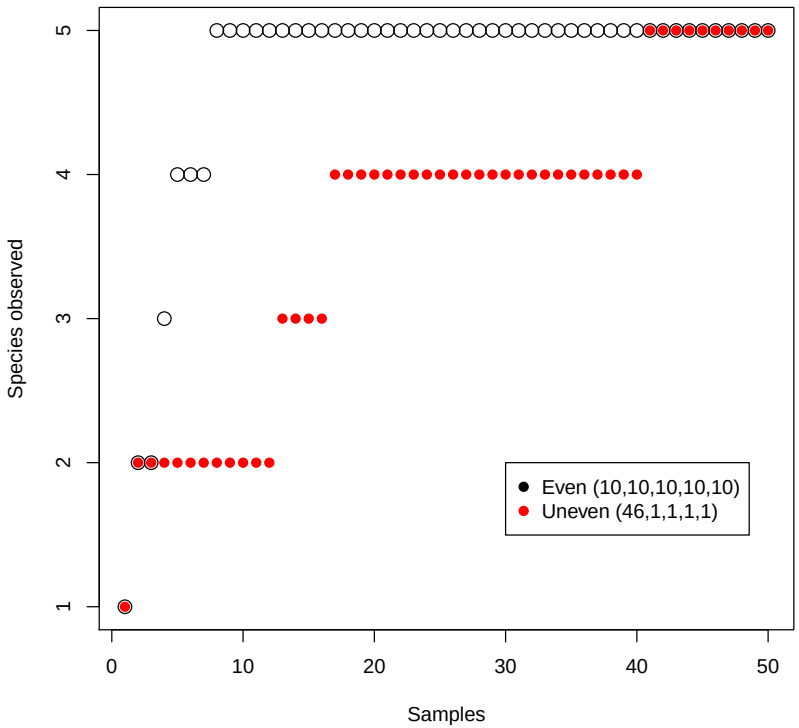
Measures of diversity: species richness

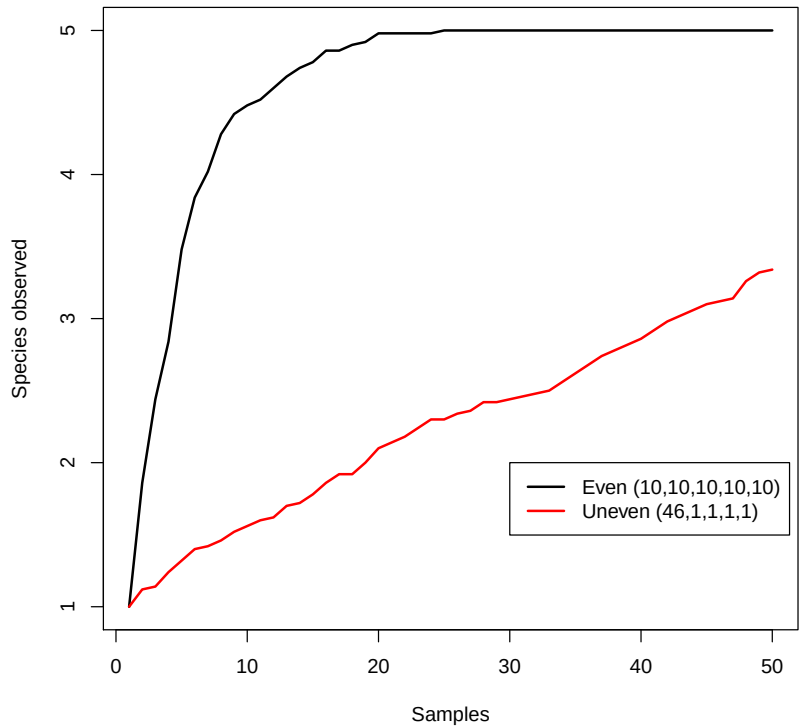
Species richness: number of species of a given taxon in a community

Problems associated with species richness:

- Taxonomical problems
- Sampling effort (space and time)
- Detectability
- Varying abundance
- ...

The effect of abundance on estimating richness





Taking samples of size 1 without replacement from a community of 50 individuals

Species richness indices

Margalef's diversity index

$$D_{Mg} = \frac{S-1}{\ln(N)}$$

Menhinick's index

$$D_{Mn} = \frac{S}{\sqrt{N}}$$

These measures try to correct for sample size

Estimating species richness

There are several methods:

- Extrapolation of the **species accumulation curve**

  The **species accumulation curve** or **collector's curve** is a graph of the cumulative number of species observed as a function of the number of samples.

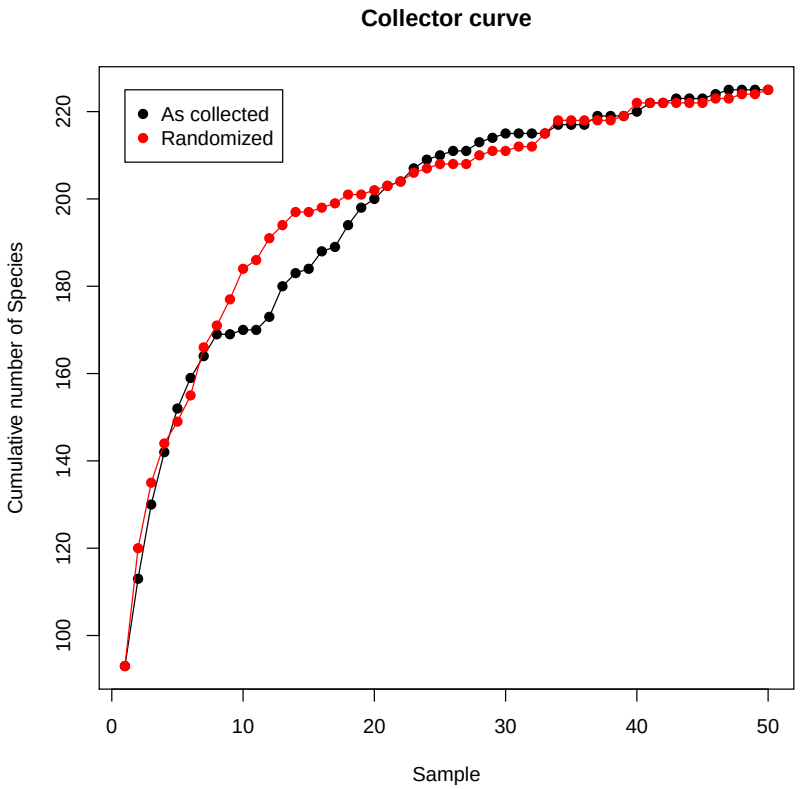
- Parametric: use the species abundance distribution to estimate richness
  - fitting a hyperbola to the species accumulation curve
  - estimating richness using Fisher's log series.
- Non-parametric estimators.
  - S_{chao}

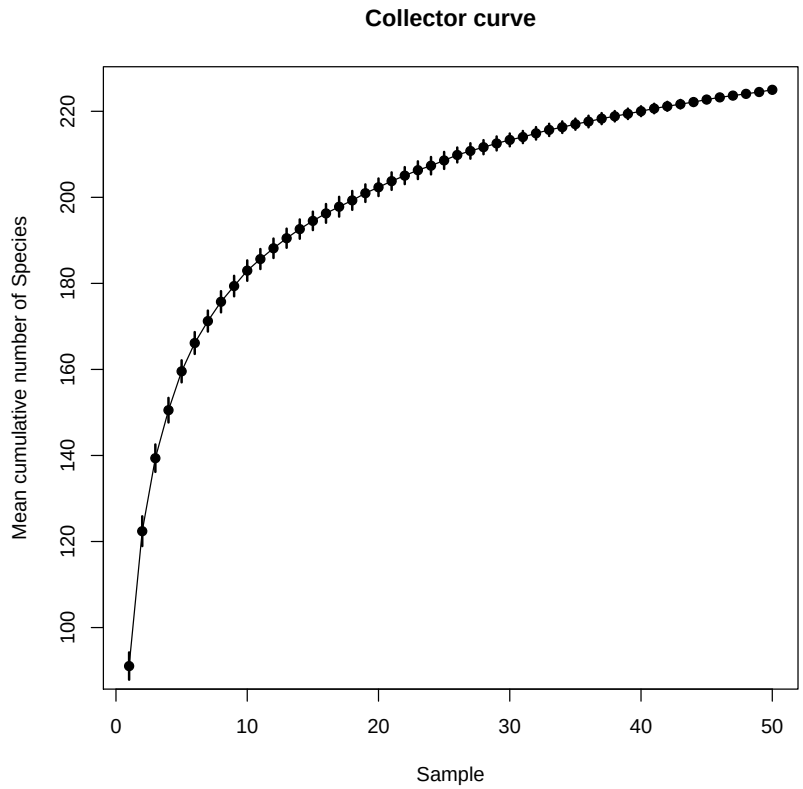
Example: Barro Colorado Island Tree Counts

	Abarema.mac	Acacia.mel	Acalypha.div	Acalypha.mac	Adelia.tri	
1	0	0	0	0	0	...
2	0	0	0	0	0	...
3	0	0	0	0	0	...
4	0	0	0	0	3	...
5	0	0	0	0	1	...
6	0	0	0	0	0	...
7	0	0	0	0	0	...
8	0	0	0	0	0	...
9	0	0	0	0	5	...
10	1	0	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮
47	0	0	0	0	2	...
48	0	0	0	0	1	...
49	0	0	0	0	0	...
50	0	0	0	0	1	...

- 50 plots of 1 hectare for which 225 species of trees were counted.
- The numbers of trees at least 10 cm in diameter at breast height was recorded.
- Data available in R package **vegan**.
- Multivariate count data, typically highly sparse.

Example: Barro Colorado Island Tree Counts





A parametric model: two-parameter hyperbola

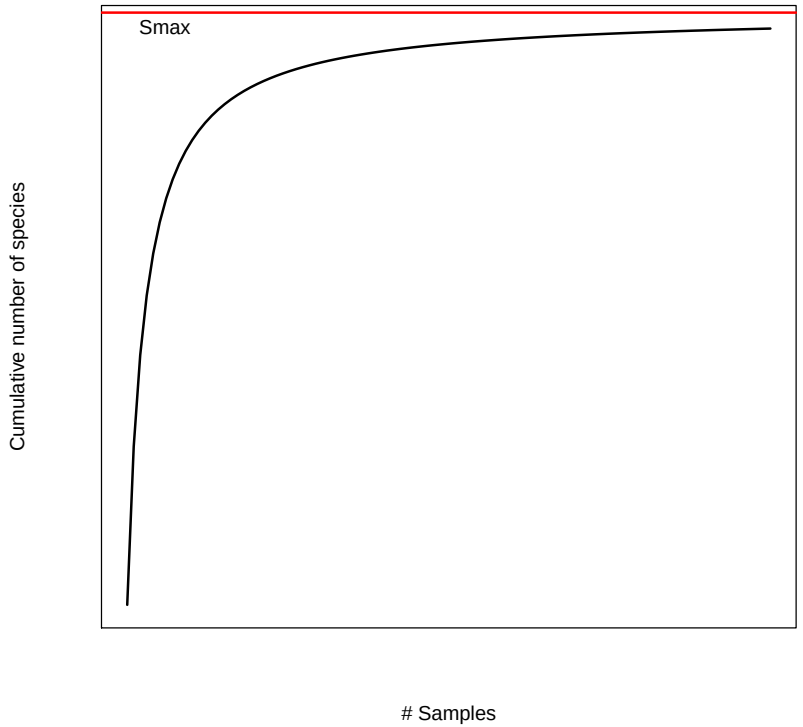
$$S_n = \frac{S_{max}\, n}{B + n}$$

ML estimates

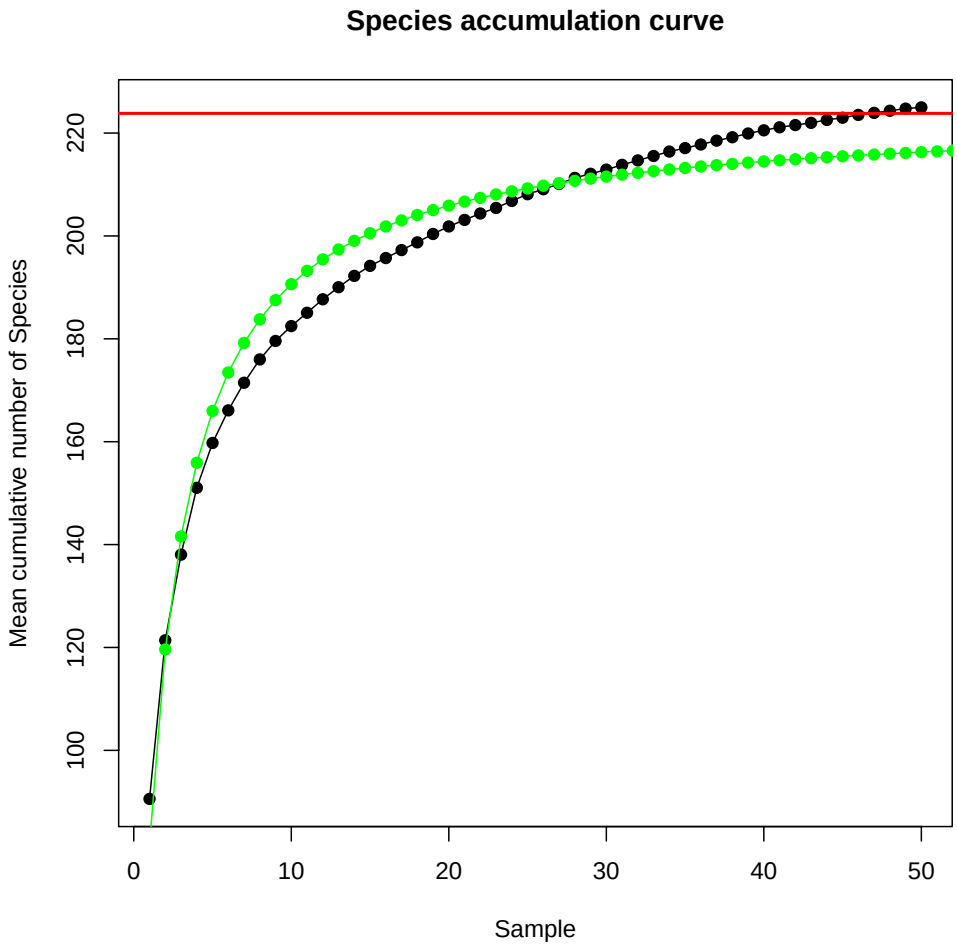
$$X_i = \frac{S_n}{n} \qquad Y_i = S_n$$

$$\hat{B} = \frac{\overline{X} S_{yy} - \overline{Y} S_{xy}}{\overline{Y} S_{xx} - \overline{X} S_{xy}}$$

$$\hat{S}_{max} = \overline{Y} + \hat{B}\overline{X}$$



Example: Barro Colorado Island Tree Counts



$$\hat{B} = 1.74 \quad \hat{S}_{max} = 223.8$$

Estimating species richness using the log-series model (parametric)

$$f(x) = \frac{-1}{\ln(1-p)}\frac{p^x}{x} \quad x = 1, 2, 3, \ldots \quad 0 < p < 1$$

x	1	2	3	4	$\cdots$	n
$f(x)$	kp	$k\frac{p^2}{2}$	$k\frac{p^3}{3}$	$k\frac{p^4}{4}$	$\cdots$	$k\frac{p^n}{n}$

$$k = \frac{-1}{\ln(1-p)}$$

We define
N = total number of individuals
S = total number of species

$$\alpha = Sk = \frac{-S}{\ln(1-p)} \text{ is a popular measure of diversity}$$

Note

$$\alpha = Sk \overset{p\approx 1}{\approx} Skp = \text{ Number of singletons}$$

α, S and N interrelated:

$$S = \alpha \ln\left(1 + \frac{N}{\alpha}\right)$$

Estimate α from your data, and substitute a total number of individuals N to obtain and estimate of S.
This is a **parametric** estimate

Non-parametric estimators

Many non-parametric estimators for S have been proposed

$$S_{chao} = S_{obs} + \frac{F_1^2}{2F_2}$$

is an estimator of the lower bound for the asymptotic richness. Its variance is given by:

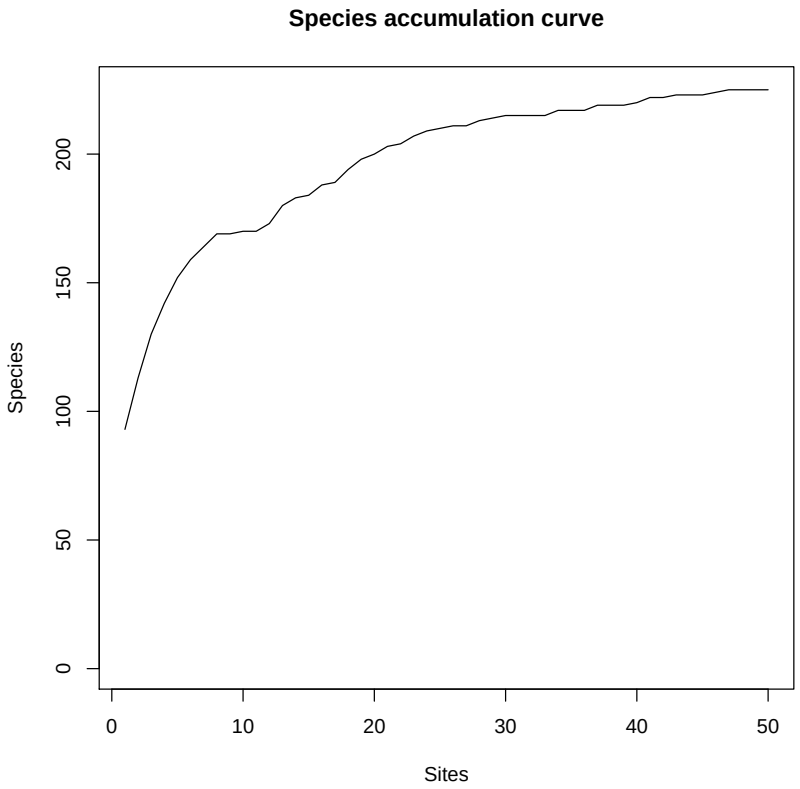
$$V(S_{chao}) = F_2 \left(\frac{1}{2} \left(\frac{F_1}{F_2} \right)^2 + \left(\frac{F_1}{F_2} \right)^3 + \frac{1}{4} \left(\frac{F_1}{F_2} \right)^4 \right)$$

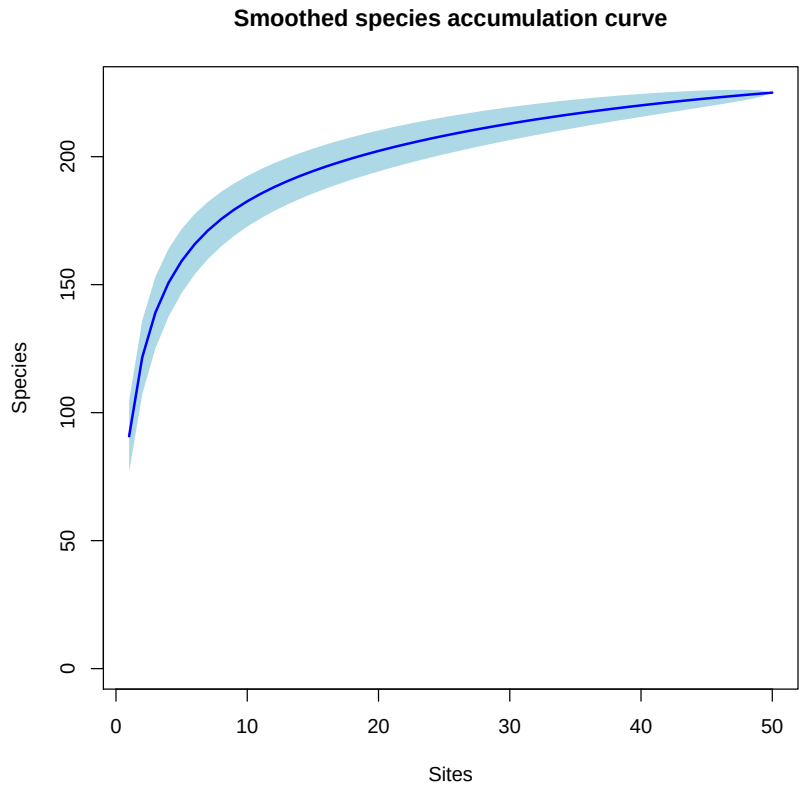
- S_{obs} the number of observed species.
- $F_1 =$ the number of singletons.
- $F_2 =$ the number of doubletons.
- $\frac{F_1^2}{2F_2}$ represents the number of species not seen, and this number is thought to be related to the number of rare species.
- S_{chao} can not be calculated if $F_2 = 0$.

Example: Fisher's butterfly data

```
> library("SPECIES")
> data("butterfly")
> colnames(butterfly)
[1] "j"   "n_j"
> butterfly
    j n_j
1   1 118
2   2  74
3   3  44
.   .   .
.   .   .
.   .   .
23 23   3
24 24   3
25 25 119
> S <- sum(butterfly[,2])
> S
[1] 620
> chao1984(butterfly)
$Nhat
[1] 714

$SE
[1] 22.66572

$CI
      lb  ub
[1,] 679 770

>
```

Example: Barro Colorado Island Tree Counts





Uncertainty in diversity measures

- The statistical distribution of many diversity measures is not known, though in some cases approximations exist.
- For the estimation of uncertainty, **replicate measurements** are essential.
- The bootstrap and the jackknife can be used to obtain estimates of uncertainty (standard errors).

The Jackknife

- S_t the original estimate
- $S_{t(-i)}$ estimate leaving out i
- $\phi_i = nS_t - (n-1)S_{t(-i)}$ the pseudovalue
- $\overline{\phi} = \frac{\sum_{i=1}^{n} \phi_i}{n}$ the jackknife estimate of diversity
- $se(\overline{\phi}) = \sqrt{\frac{\sum_{i=1}^{n}(\phi_i - \overline{\phi})^2}{n(n-1)}}$
- Confidence interval: $\overline{\phi} \pm t_{\alpha/2,n-1} se(\overline{\phi})$

Comparing diversity measures

To sensibly compare estimates of a certain diversity measure, a set of assumptions must be met.

- Samples must be collected by using the same sampling technique.
- Similar biological communities are being compared.
- Evenness of the compared communities is similar.
- Individuals are sampled at random.
- The sampling effort is similar, or sampling effort is adjusted for.
- ...

Comparing the Shannon index for two sites (Diversity t-test)

$$p_i = \frac{n_i}{N} \qquad H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

$$V(H') = \frac{\sum_{i=1}^{S} p_i (\ln(p_i))^2 - \left(\sum_{i=1}^{S} p_i \ln(p_i)\right)^2}{N} + \frac{S-1}{N^2}$$

$$T = \frac{|H_1 - H_2|}{\sqrt{V(H_1) + V(H_2)}}$$

$$\nu = \frac{(V(H_1) + V(H_2))^2}{V(H_1)^2/N_1 + V(H_2)^2/N_2}$$

Under H_0 of equality of the indexes:

$$T \sim t_\nu \qquad \text{p-value } = P(t_\nu > T)$$

Notes:

- Assumes equal sampling conditions.
- Relies on asymptotic normality of the Shannon index.

Example Diversity t-test

Species	Derrycunniy.oakwood	Muckross.yew.wood
Chaffinch	35	9
Robin	26	20
Blue tit	25	10
Goldcrest	21	21
Wren	16	5
Coal tit	11	14
Spotted flycatcher	6	0
Tree creeper	5	3
Siskin	3	2
Blackbird	3	6
Great tit	3	9
Long-tailed tit	3	2
Woodpigeon	3	0
Hooded crow	2	0
Woodcock	2	0
Song thrush	2	6
Redstart	1	0
Mistle thrush	1	0
Dunnock	1	0
Sparrow hawk	1	1
Long-eared owl	0	1
Jay	0	1
Chiff chaff	0	0

```
N1 = 170
H1 = 2.408
N2 = 110
H2 = 2.346
T = 0.585
DF = 263.741
p-value = 0.560
```

α and β diversity

- α diversity refers to the diversity of a well-defined community or assemblage, and also refers to a defined spatial unit.
- The term β diversity is used to refer to biotic change and species replacement as we move to a second spatial unit.
- β diversity measures the extent to which the diversity of two or more spatial units differ.

How many species are there on the earth?

- Estimates range between 5 and 10 million.
- Only less than 2 million have been formally recorded.
- Insects make a large contribution.

Bibliography

- Magurran, A. E. (2004) Measuring biological diversity, Blackwell Publishing.
- Lowe, A., Harris, S. & Ahston, P. (2004) Ecological genetics: design, analysis and application, Blackwell Publishing.