

Biostatistics

Other topics

Nuno Sepúlveda, 13.01.2025

Syllabus

1. General review

- a. What is Biostatistics?
- b. Population/Sample/Sample size
- c. Type of Data – quantitative and qualitative variables
- d. Common probability distributions
- e. Work example – Malaria in Tanzania

2. Applications in Medicine

- a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve, Rogal-Gladen estimator
- b. Estimation of treatment effects - generalized linear models
- c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model

3. Applications in Genetics, Genomics, and other 'omics data

- a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
- b. Methylation association studies – M versus beta values, estimation of biological age
- c. ~~Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative Binomial~~

4. Other Topics

- a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
- b. Serological analysis – Gaussian (skew-normal) mixture models
- c. ~~Advanced sample size and power calculations~~

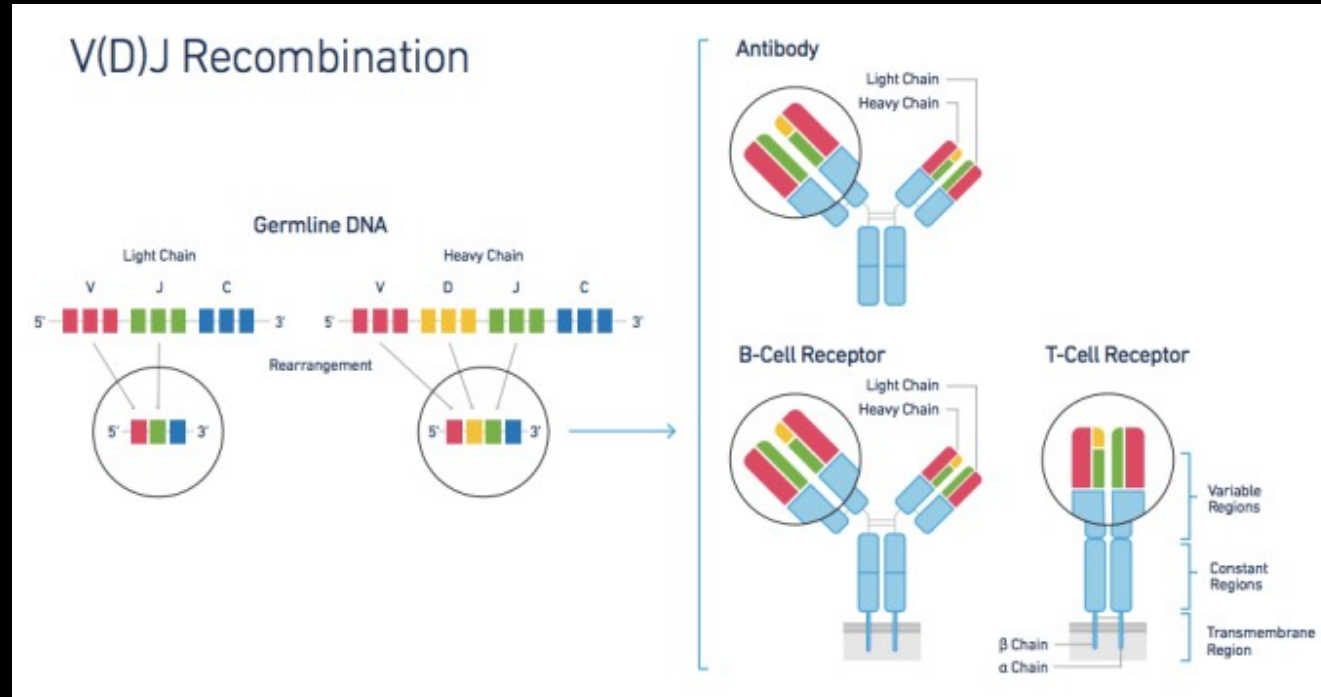
Species Biodiversity



Microbial diversity



B/T-cell diversity



Species richness

The number of different species present in a given population

Species diversity

The number of different species and their abundance present in a given population

Main research question

**How many species existing based on
the information from a sample?**

Formulation of the problem

Data

Species	Abundance
Species_1	n_1
Species_2	n_2
Species_3	n_3
....	
Species_k	n_k

n = sample size

$$n = \sum_{i=1}^k n_i$$

n is the hypothetical maximum species richness due to sampling

n_i = frequency of individuals (abundance) from species i

k = number of different species represented in the sample

K = unknown number of different species in the population

$\hat{K} = ?$

Under the assumption of a large population

Formulation of the problem

Summarising the Data

Abundance	Number of Species
1	m_1
2	m_2
3	m_3
....	
l	m_l

Species-abundance distribution

n = sample size

$$n = \sum_{i=1}^k i \times m_i$$

m_i = frequency of species with abundance i

k = number of different species represented in the sample

Example

TABLE 1
DISTRIBUTION OF LEPIDOPTERA CAUGHT IN A LIGHT TRAP AT ROTHAMSTED IN 1934

Individuals per species (r)	Number of species (n _r)	Expected		
		Logarithmic	Grouped lognormal	Poisson lognormal
1	34	39.0	32.7	31.2
2	19	19.3	20.5	20.8
3	15	12.7	14.6	15.0
4	10	9.4	11.2	11.5
5	10	7.4	8.9	9.1
6-7	9	11.3	13.4	13.7
8-10	17	12.0	13.7	13.9
11-14	9	11.0	12.0	12.1
15-20	14	11.2	11.3	11.3
21-28	10	9.8	9.1	9.0
29-39	6	8.7	7.4	7.3
40-55	7	7.8	6.1	6.0
56-77	3	6.2	4.6	4.5
78-108	5	4.6	3.4	3.4
109-151	4	2.8	2.4	2.4
152-	4	2.9	4.5	4.7
Total	176	176.1	175.8	175.9
χ^2		9.2	6.9	7.4
Degrees of freedom		14	13	13

Simpson's Diversity index

$$D_s = 1 - \underbrace{\frac{\sum_{i=1}^k n_i(n_i - 1)}{n(n - 1)}} = 1 - \frac{\sum_{i=1}^l m_i i(i - 1)}{n(n - 1)}$$

The probability that two individuals taken at random from the sample (with replacement) are from the same species

$D_s = 0 \Rightarrow$ Individuals are from the same species ($k = 1, n_1 = n$)

$D_s = 1 \Rightarrow$ Every individual is from a different species ($n_i = 1, \forall_i$)

Shannon's Diversity index

$$H = - \sum_{i=1}^k p_i \log p_i \quad \text{where } p_i = n_i/n$$

It quantifies the uncertainty associated with the species prediction when one takes an individual from the sample randomly

Maximal entropy $n_i = 1, \forall_i \Rightarrow H_{\max} = \log n$

Exercise: Data_lecture_13_TCR_diversity.csv

Calculate the Simpson's and Shannon's diversity indexes for the species abundance distribution of DP CD3low

<i>i</i>	Thymus			Lymph nodes	
	DP CD3low	SP CD4 ⁺	SP CD8 ⁺	LN CD4 ⁺	LN CD8 ⁺
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

How to estimate species richness?

Abundance	Number of Species
0	$D-k$
1	m_1
2	m_2
3	m_3
...	
l	m_l
$> l$	0

Augmented Species-abundance distribution

If this is a contingency table, what is a possible sampling model?

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l
> l	0

Augmented Species-Abundance
distribution

Multinomial distribution

$$f(\{m_i\} | D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!} \theta_0^{D-M} \prod_{i=1}^k \theta_i^{m_i}$$

θ_i = probability of sampling i individuals from a given species

Unknown parameters $D, \theta_i, i = 0, 1, \dots, l$

Is it possible to estimate this model?

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l
>l	0

Augmented Species-Abundance
distribution

Modelling θ_i

$$\theta_i = P[X = i | \lambda]$$

$$X | \lambda \rightsquigarrow \text{Poisson}(\lambda)$$

$$\theta_i = \frac{e^{-\lambda} \lambda^i}{i!}$$

$$f(\{m_i\} | D, \{\theta_i\}) = \frac{D!}{(D-k)! m_1! \cdots m_l!} e^{-\lambda(D-M)} \prod_{i=1}^k \frac{e^{-\lambda m_i} \lambda^{i m_i}}{i!}$$

How can we estimate this model?

D is an integer parameter while λ is a positive
continuous parameter

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

First solution (truncated Poisson)

$$k | D, \theta_0 \rightsquigarrow \text{Binomial}(D, 1 - \theta_0)$$

$$f(\{m_i\} | k, \{\theta_i\}) = \frac{k!}{m_1! \cdots m_l!} \prod_{i=1}^k \left(\frac{\theta_i}{1 - \theta_0} \right)^{m_i}$$

1. Estimate a Poisson truncated at zero using raw data only

2. Estimate D from the binomial using $\hat{D} = \frac{k}{1 - \hat{\theta}_0}$

$$\hat{\theta}_0 = e^{-\hat{\lambda}}$$

Exercise: Data_lecture_13_TCR_diversity.csv

Estimate the species richness D for the DP CD3low cells using the first solution. Use `vglm` from package `VGAM` to estimate the truncated Poisson model.

i	Thymus			Lymph nodes	
	DP CD3low	SP CD4 ⁺	SP CD8 ⁺	LN CD4 ⁺	LN CD8 ⁺
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

Second solution (profile likelihood)

$$f(\{m_i\} | D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!} \theta_0^{D-M} \prod_{i=1}^k \theta_i^{m_i}$$

1. Fix $\hat{D}=k$
2. Estimate the parameter of the Poisson distribution via maximum likelihood and calculate the respective maximized log-likelihood. (What is the MLE of λ ?)
3. Do $\hat{D} + 1$ in one unit and repeat previous step
4. Keep incrementing if the maximised log-likelihood is increasing
5. The estimate of D is the value immediately before when the maximized log-likelihood starts decreasing

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

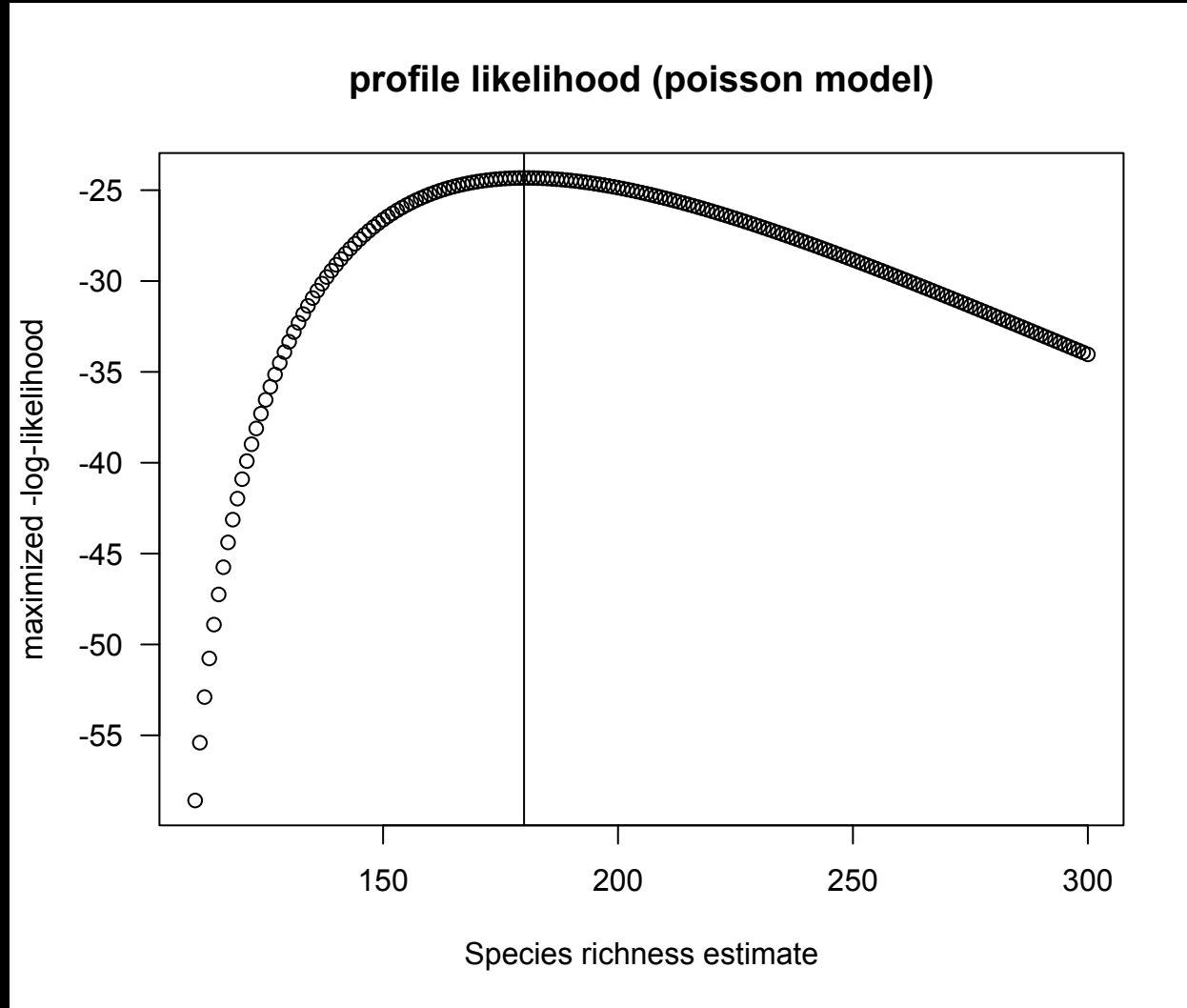
Augmented Species-Abundance
distribution

Second solution (profile likelihood)

$$f(\{m_i\} | D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!} \theta_0^{D-M} \prod_{i=1}^k \theta_i^{m_i}$$

1. Fix $\hat{D}=k$
2. Estimate the parameter of the Poisson distribution via maximum likelihood and calculate the respective maximized log-likelihood. (What is the MLE of λ ?)
3. Do $\hat{D} + 1$ in one unit and repeat previous step
4. Keep incrementing if the maximised log-likelihood is increasing
5. The estimate of D is the value immediately before when the maximized log-likelihood starts decreasing

How to estimate species richness?



Exercise: Data_lecture_13_TCR_diversity.csv

Estimate the species richness D for the DP CD3low cells using the second solution.

i	Thymus			Lymph nodes	
	DP CD3low	SP CD4 ⁺	SP CD8 ⁺	LN CD4 ⁺	LN CD8 ⁺
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

Calculation of a 95% confidence interval using the
profile likelihood

Use the critical value of the Wilks's ratio test

$$H_0 : D = D_0 \text{ versus } H_1 : D \neq D_0$$

$$\Lambda = -2(\log L_{D_0} - \log L_{\hat{D}}) | H_0 \rightsquigarrow \chi^2_{(1)}$$

$$\text{critical value} = q_{95\%, \chi^2_{(1)}}$$

accept H_0 if $\Lambda < q_{95\%, \chi^2_{(1)}}$ reject H_0 , otherwise

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

Calculation of a 95% confidence interval using the
profile likelihood

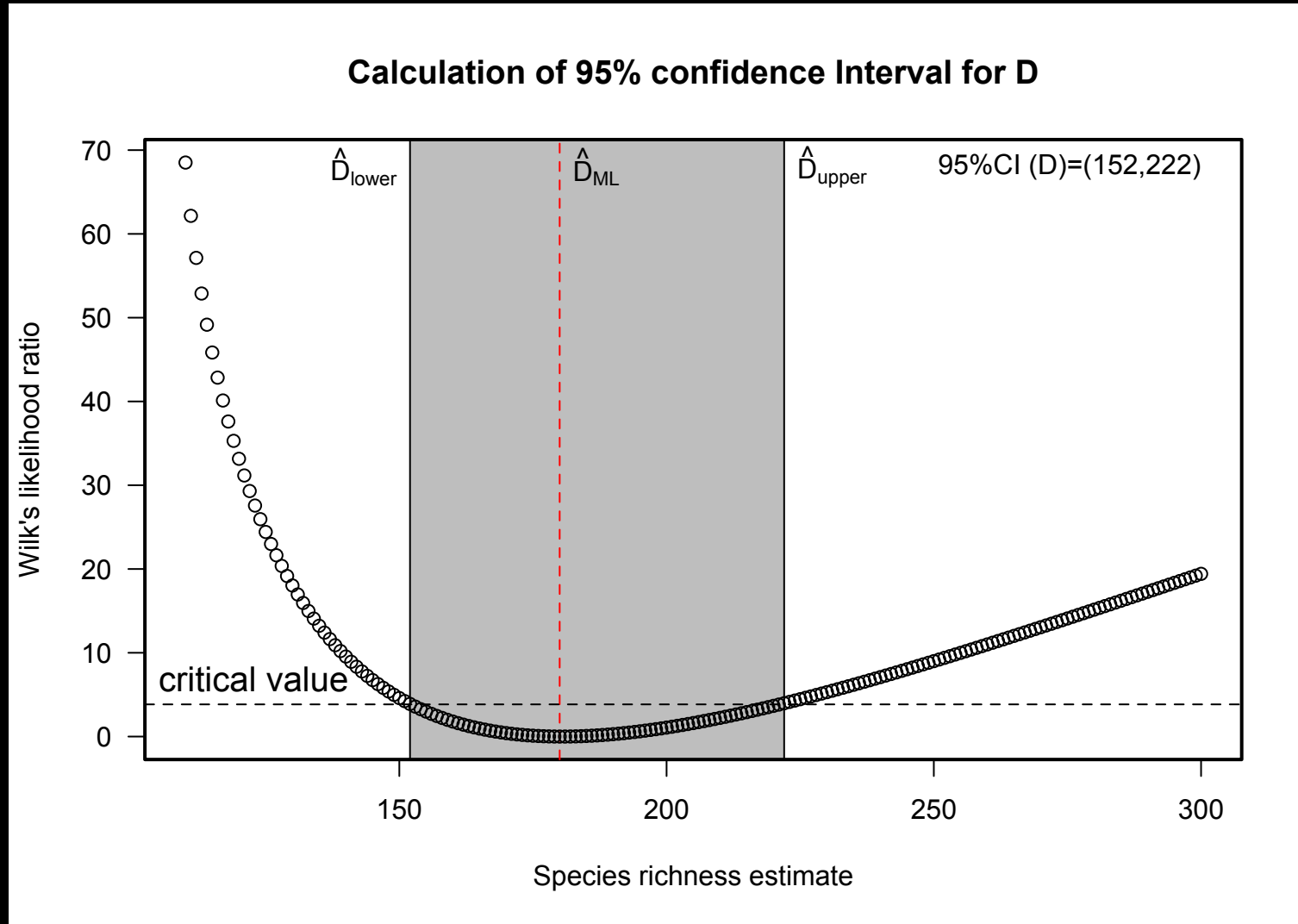
Use the critical value of the Wilks's ratio test

$$q_{95\%, \chi^2_{(1)}} = -2(\log L_{D_0} - \log L_{\hat{D}})$$

$$(\hat{D}_{lower}; \hat{D}_{upper})$$

\hat{D}_{lower} and \hat{D}_{upper} are the solutions of the above question

How to estimate species richness?



How to estimate species richness?

Abundance	Number of Species
1	m_1
2	m_2
3	m_3
....	
l	m_l
$>l$	0

Augmented Species-Abundance
distribution

Pearson's goodness of fit test to check whether the model fits the data well

Use only the observed data

$$f(\{m_i\} | k, \{\theta_i\}) = \frac{k!}{m_1! \cdots m_l!} \prod_{i=1}^k \left(\frac{\theta_i}{1 - \theta_0} \right)^{m_i}$$

$$\hat{\theta}_i = \frac{e^{-\hat{\lambda}} \hat{\lambda}^i}{i!}$$

Exercise: Data_lecture_13_TCR_diversity.csv

Calculate the confidence interval for the species richness D for the DP CD3low cells using the profile likelihood plot. Check whether the Poisson model fits the data well using the Pearson's goodness of fit test.

i	Thymus			Lymph nodes	
	DP CD3low	SP CD4 ⁺	SP CD8 ⁺	LN CD4 ⁺	LN CD8 ⁺
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

Poisson-Gamma mixture model for estimating diversity richness

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
I	m _I

Augmented Species-Abundance
distribution

Modelling θ_i

$$\theta_i = P[X = i | \lambda]$$

$$X | \lambda \rightsquigarrow \text{Poisson}(\lambda)$$

$$\lambda | \alpha, \beta \rightsquigarrow \text{Gamma}(\alpha, \beta)$$

$$\begin{aligned}
 P[X = x] &= \int_0^\infty P[X = x | \lambda] P[\lambda] d\lambda = \int_0^\infty \frac{e^{-\lambda} \lambda^x}{x!} \times \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} d\lambda \\
 &= \frac{\Gamma(i + \alpha)}{\Gamma(i + 1)\Gamma(\alpha)} \left(\frac{\beta}{\beta + 1} \right)^\alpha \left(\frac{1}{\beta + 1} \right)^i
 \end{aligned}$$

Negative Binomial

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

First solution (truncated Negative Binomial)

$$k | D, \theta_0 \rightsquigarrow \text{Binomial}(D, 1 - \theta_0)$$

$$f(\{m_i\} | k, \{\theta_i\}) = \frac{k!}{m_1! \cdots m_l!} \prod_{i=1}^k \left(\frac{\theta_i}{1 - \theta_0} \right)^{m_i}$$

1. Estimate a Poisson truncated at zero using raw data only

2. Estimate D from the binomial using $\hat{D} = \frac{k}{1 - \hat{\theta}_0}$

$$\hat{\theta}_0 = e^{-\hat{\lambda}}$$

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m ₁
2	m ₂
3	m ₃
....	
l	m _l

Augmented Species-Abundance
distribution

Second solution (profile likelihood)

$$f(\{m_i\} | D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!} \theta_0^{D-M} \prod_{i=1}^k \theta_i^{m_i}$$

1. Fix $\hat{D}=k$
2. Estimate the parameters of the Negative distribution via maximum likelihood and calculate the respective maximized log-likelihood. (What is the MLE of λ ?)
3. Do $\hat{D} + 1$ in one unit and repeat previous step
4. Keep incrementing if the maximised log-likelihood is increasing
5. The estimate of D is the value immediately before when the maximized log-likelihood starts decreasing

Exercise: Data_lecture_13_TCR_diversity.csv

Estimate the species richness D for the DP CD3low cells using the Negative Binomial distribution. Estimate via the second solution.

i	Thymus			Lymph nodes	
	DP CD3low	SP CD4 ⁺	SP CD8 ⁺	LN CD4 ⁺	LN CD8 ⁺
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1