メタゲノムデータを用いた群集解析法

門脇 浩明
京都大学大学院 人間・環境学研究科
日本学術振興会特別研究員

第63回日本生態学会仙台大会 自由集会W31（企画：田辺晶史）

メタゲノム解析を通じ
生態学に対しどのような貢献ができるのか

1. メタゲノム解析によって得られるデータの特徴
2. データの取り扱い方法
3. データの統計解析法とその例

メタゲノム解析によって得られるデータ

群集行列データのExcel表 (Claident)

列：種、もしくはOTU (operational taxonomic unit)

行：サンプル

ID	OTU 1	OTU 2	OTU 3	OTU 4	・・・	OTU 99
1	21	132	0	56	・・・	2
2	0	16	0	1	・・・	56
3	0	56	10	44	・・・	0
4	59	71	0	28	・・・	0
5	0	0	199	0	・・・	24
6	0	2	87	2	・・・	75
7	33	1	0	0	・・・	0

次世代シーケンサーの配列データ

→配列をOTUごとにまとめる

→サンプルごとのOTUのリード数を集計

```
> comm <- read.table("/….txt", header=T)
```

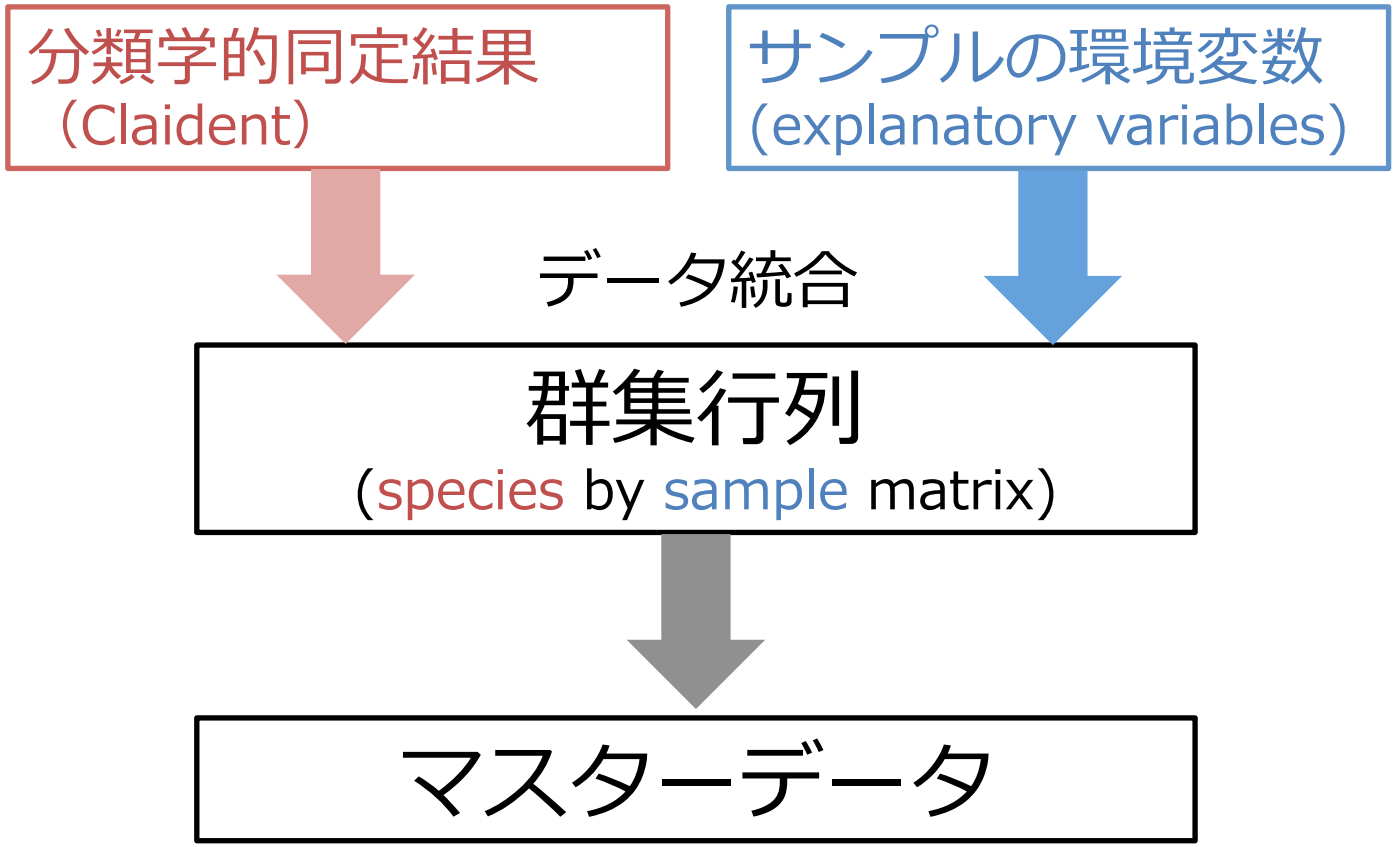
メタゲノム解析によって得られるデータ

分類学的同定結果のExcel表 (Claident)

query	Kingdom	Phylum	Class	Order	Family	Genus	Species
Sp1	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Amanitaceae	*Amanita*	
Sp2	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp3	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp4	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp5	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Amanitaceae	*Amanita*	*Amanita fuliginea*
Sp6	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Entolomataceae		
Sp7	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp8	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp9	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Tricholomataceae	*Laccaria*	*Laccaria bicolor*
Sp10	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Cortinariaceae	*Cortinarius*	
Sp11	Fungi	Basidiomycota	Agaricomycetes	Agaricales			
Sp12	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Amanitaceae	*Amanita*	
Sp13	Fungi	Basidiomycota	Agaricomycetes	Agaricales	Cortinariaceae	*Cortinarius*	

```
> idlist <- read.delim("/….txt")
```

データの取り扱い方法




```
> taxon <- idlist[match(colnames(comm),idlist$query),]
> masterdata<-cbind(t(comm),taxon)
```

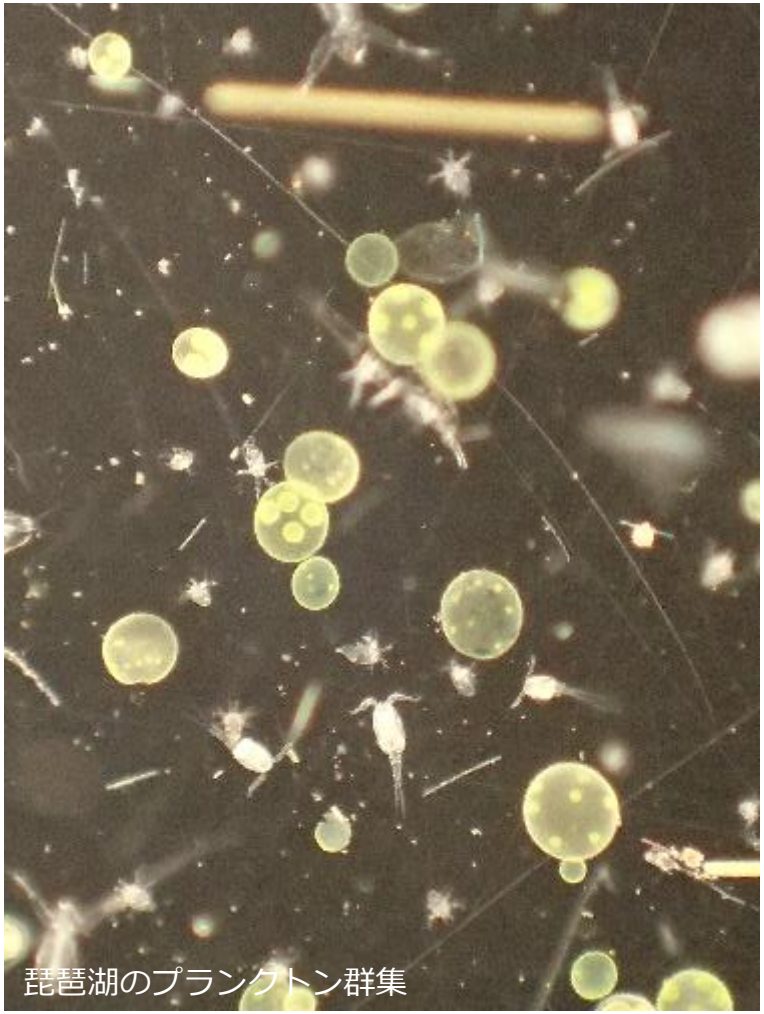
データの統計解析法

1. 希釈法 (rarefaction)

2. 探索的データ解析
   (exploratory analysis)

3. 仮説検定(hypothesis testing)

これら３つのステップを通じ
メタゲノムデータを生態学に活かす
方法について考える



琵琶湖のプランクトン群集

希釈法（rarefaction）とは

マーブルチョコのアナロジー

- チョコ1粒＝１個体の生物
- チョコの色ごとに異なる種

チョコをn個選んだ場合
何色（何種）が観察されるか？

リード数 ＝ シーケンサーの探索努力（例：観察個体数など）

リード数の異なるサンプルの種多様性は単純に比較できない

より公平な比較を行うためには、リード数を適切に希釈する必要がある

希釈法（rarefaction）

異なる探索努力（リード数）のもと得られたデータを比較するためには？

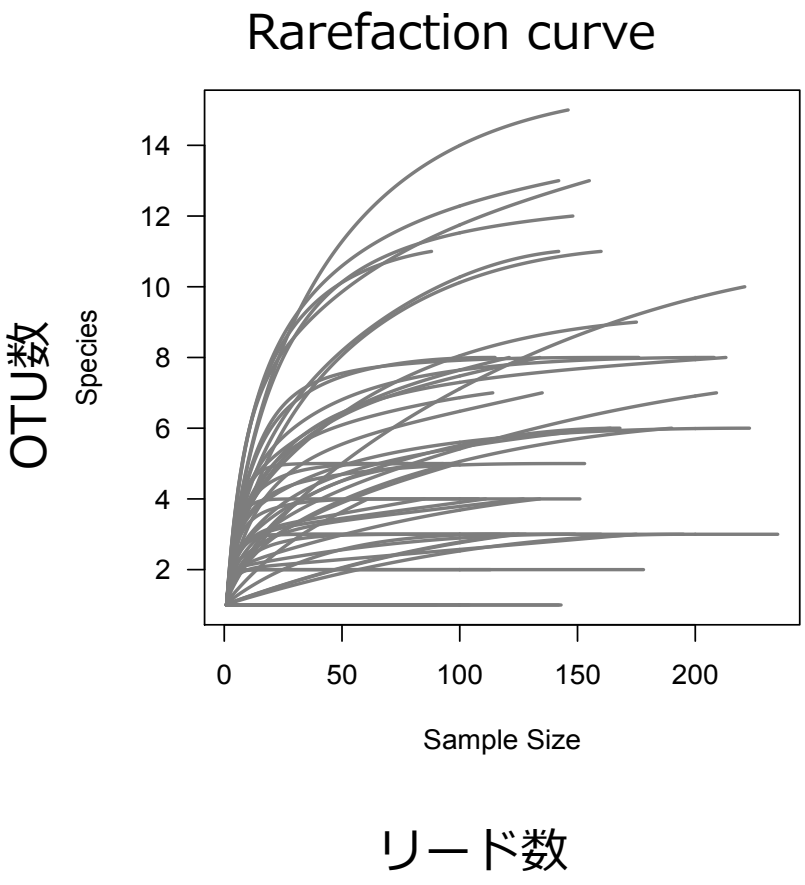
サンプルサイズ*n*のもとで期待される種数

$$E(S) = \sum_{i=1}^{s} \left(1 - \left[\frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right] \right)$$

N = total number of individuals

Ni = number of individuals in the *i* th species

n = size of the smaller sample (rarefied)



Hurlbert (1971)Ecology; Gotelli and Graves (1996)

希釈法（rarefaction）

最も素朴な方法：

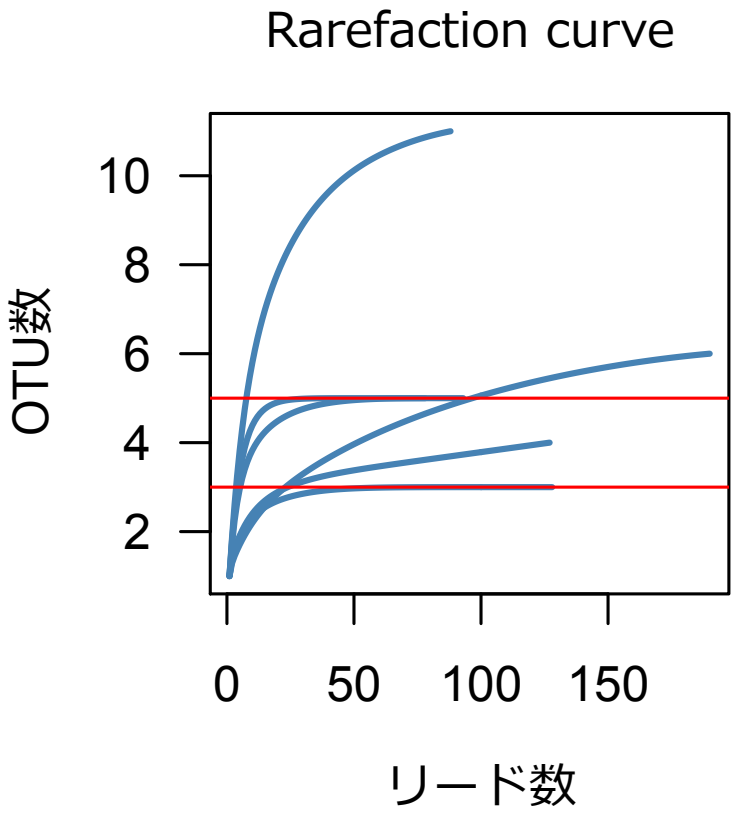
種数の増加が頭打ちになったサンプルのみを解析に用いる（十分なリード数が得られなかったサンプルを破棄）

↓

大量のデータが無駄となる可能性

↓

代替案：
カバレッジで揃え、データを希釈する



Hurlbert (1971)Ecology; Gotelli and Graves (1996)

Package Vegan: rarefy, rrarefy, rareslope, rarecurve

希釈法（rarefaction）

同じ傾きとなるリード数において希釈
例） カバレッジ98%(傾き0.02)で揃える

$$\mathrm{J} = \mathrm{E(S)} = \sum_{i=1}^{S}\left(1 - \left[\frac{\binom{N-N_i}{n}}{\binom{N}{n}}\right]\right)$$

nについて微分

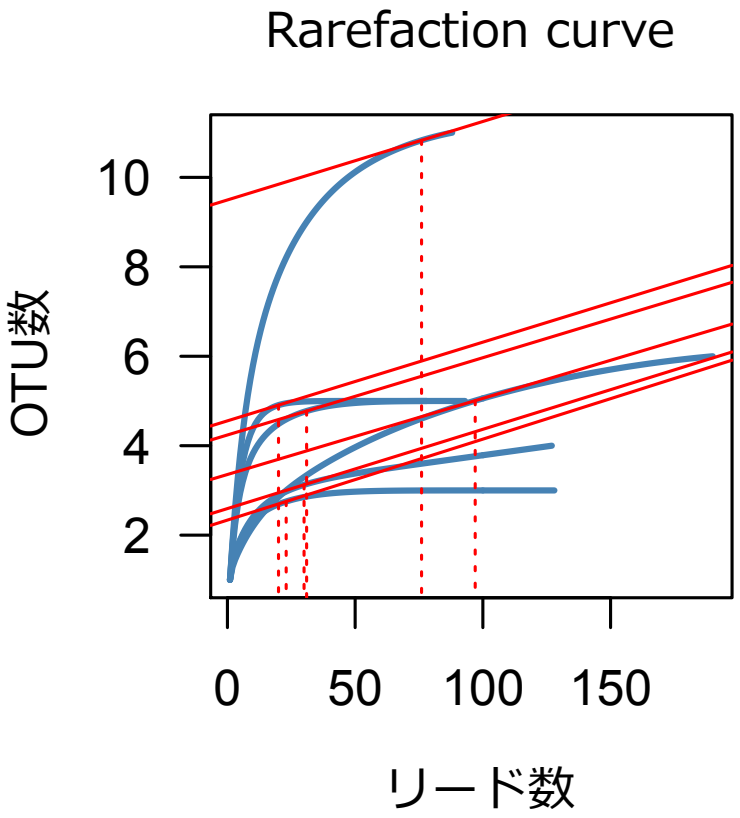
ガンマ関数の公式を活用

$$\Gamma(\mathrm{n}) = \int_0^{\infty} t^{n-1}e^{-t}dt$$

$$\Gamma(\mathrm{n}+1) = \mathrm{n}!$$

種数期待値の導関数（傾き）

$$\frac{d}{dn}(\mathrm{J}) = \left\{\frac{\Gamma'(N-n+1)}{\Gamma(N-n+1)} - \frac{\Gamma'(N-\mathrm{N}_i-n+1)}{\Gamma(N-\mathrm{N}_i-n+1)}\right\} e^{\{\log(\Gamma(N-Ni+1)+\log(\Gamma(N-n+1))-\log(\Gamma(N-N_i-n+1))-\log(\Gamma(\mathrm{N}+1))\}}$$



希釈法（rarefaction）

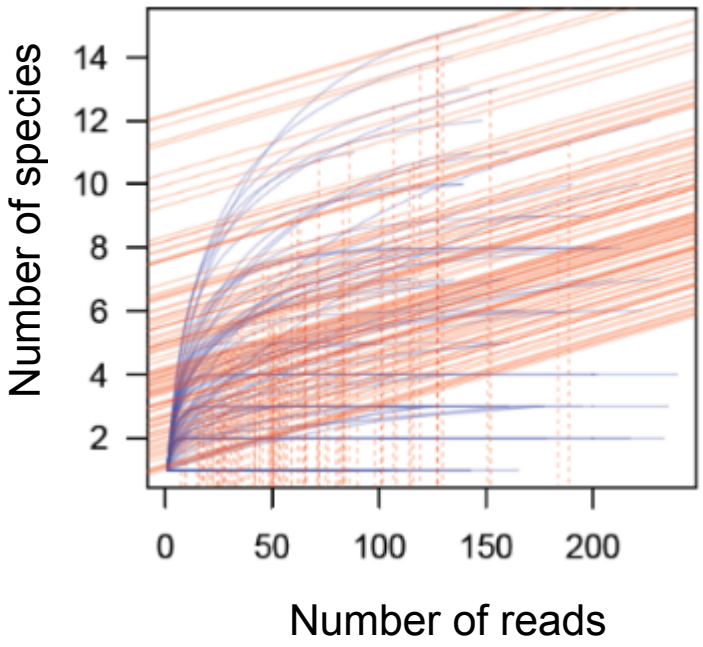
指定されたリード数の群集行列データを希釈する（rrarefy関数を用いる）

→　二値データ（在／不在データ）に変換する

生成された二値データの例

ID	Sp 1	Sp 2	Sp 3	Sp4	・・・	Sp99
1	1	1	0	1	・・・	0
2	0	0	0	0	・・・	1
3	0	1	1	1	・・・	0
4	1	1	0	1	・・・	0
5	0	0	1	0	・・・	0
6	0	0	1	0	・・・	1
7	1	0	0	0	・・・	0

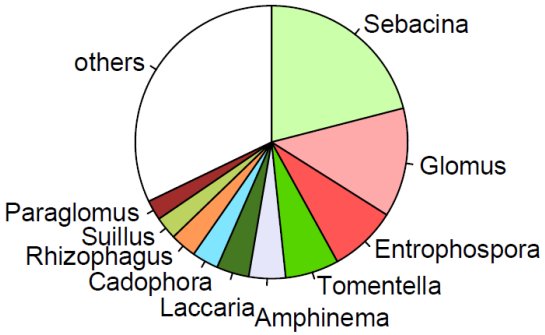
Rarefaction curveの描画



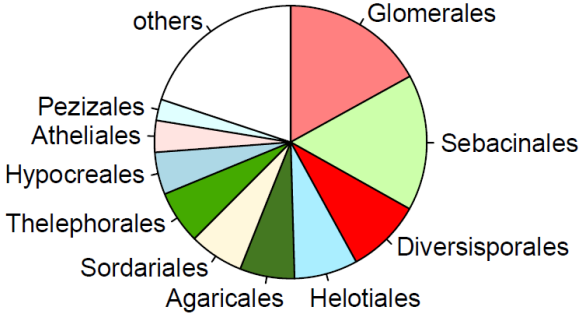
以上をもって群集解析の準備が完了

探索的データ解析ー種構成・分類群構成の図

Genusレベルでの群集組成

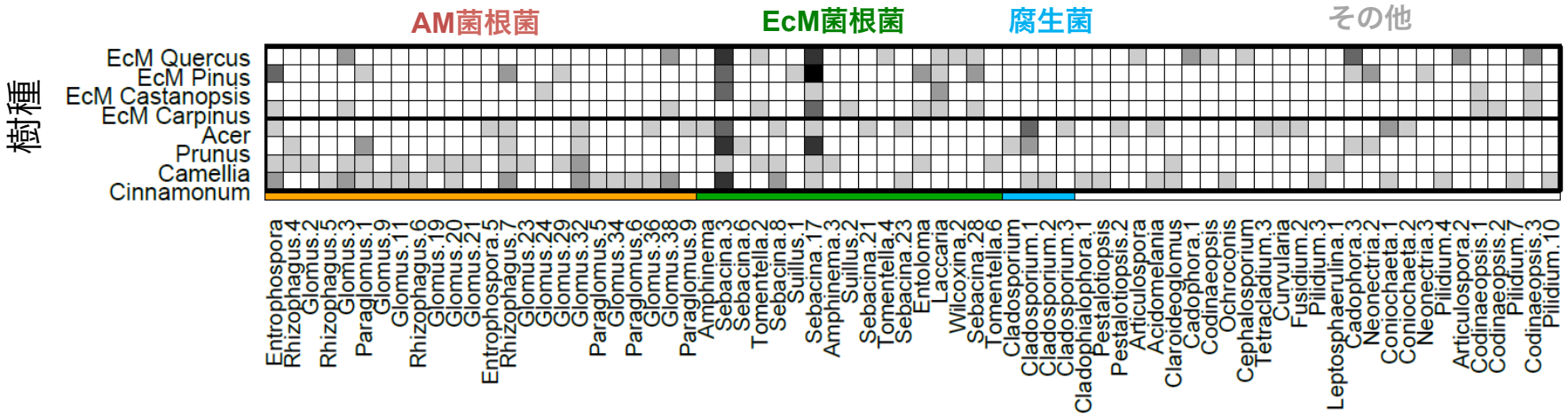


Orderレベルでの群集組成



Claidentの分類情報を活用した様々なグラフィックスが利用可能

樹木と根圏真菌群集の相互作用行列



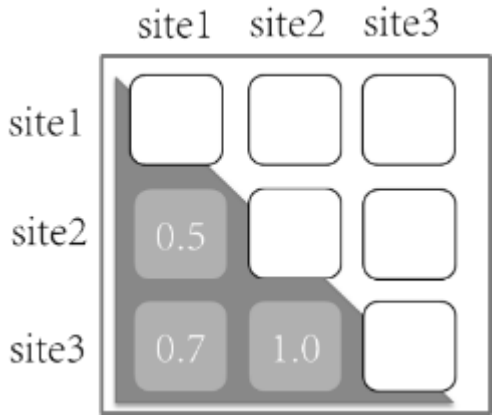
Package Graphics: piechart ; Bipartite : visweb

α、β、γ多様性の定量化

サンプル×種の群集行列

	sp1	sp2	sp3	
site1	1	1	0	site1 α diversity = 2 species
site2	0	1	0	site2 α diversity = 1 species
site3	1	0	1	site3 α diversity = 2 species
	2 sites	2sites	1site	γ diversity = 3 species

群集の非類似度行列



Jaccardの非類似度指数

$$\frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- β多様性の計算手法は多様である
- β多様性に関する論文の全てが役立つわけではない

Anderson et al. (2011) Ecol Lett

```
> vegdist(..., method="bray", binary=FALSE)
```

仮説検定 (hypothesis testing)

- サルの食性幅とその変異性の解析
- 魚類DNAの時空間出現パターン（メタ個体群動態）の解析
- 微生物種間の排他的分布パターンの解析
- 真菌群集が寄主植物の成長率に与える影響の解析

サルの食性幅とその変異性の解析（PERMANOVAとPERMDISP）

サルの採餌行動は集団間でどのように異なるか？

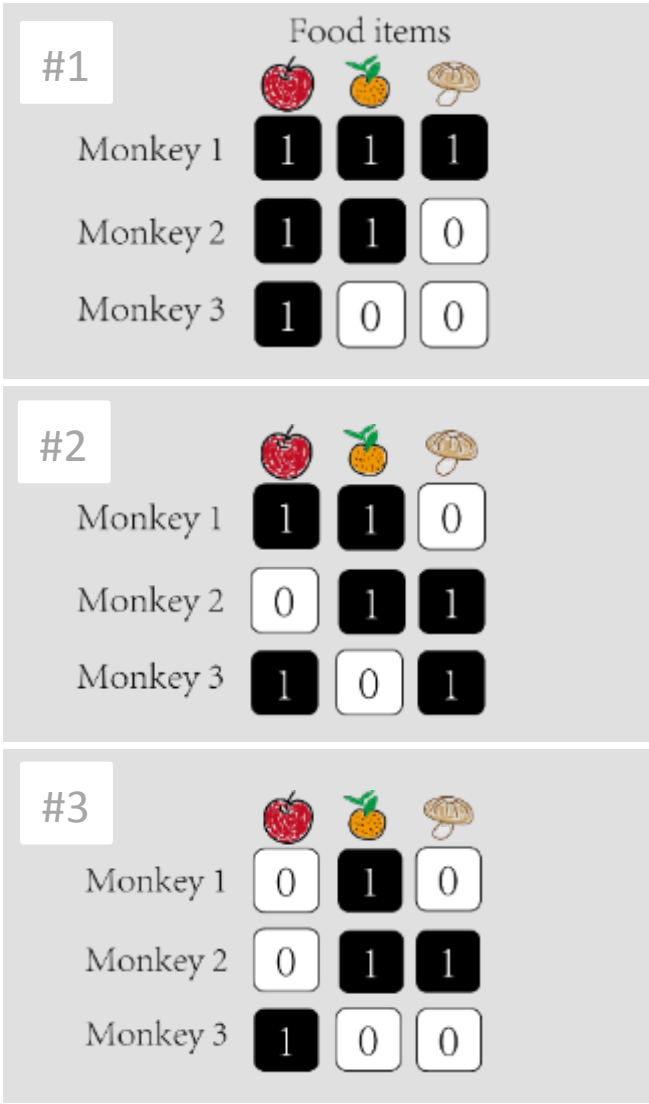
サル糞のDNA分析により、サル集団の餌組成について、３つの集団間で比較する場合

複数のβ多様性指数で解析する利点：

- ◆ 餌種数＋餌組成の違い → Jaccard非類似度指数
- ◆ 餌組成の違いのみ → Raup-Crick非類似度指数

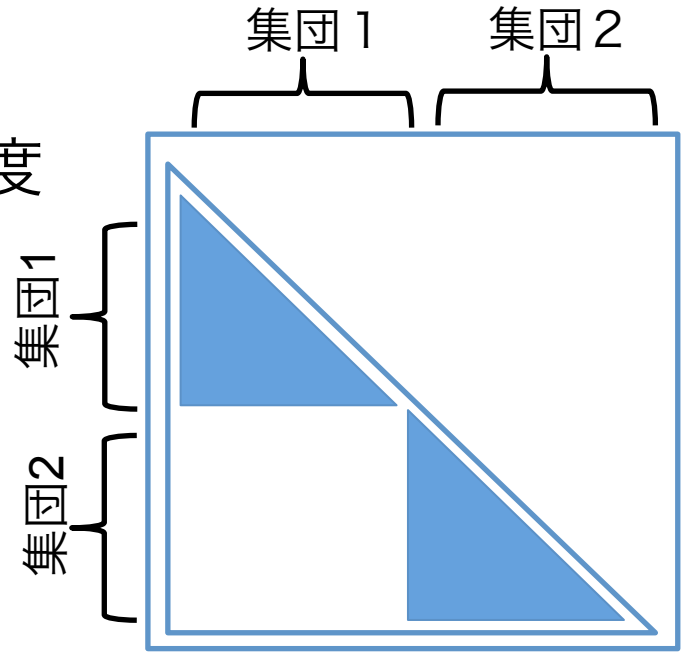
Jaccardの非類似度指数

$$\frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$



サルの食性幅とその変異性の解析（PERMANOVAとPERMDISP）

非類似度行列



全サンプル数 N
グループ内のデータ数 n
サンプルiとjの間のβ多様性 d_{ij}
サンプルiとjが同グループか否か ε_{ij}

Total sum of squares (SS)

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2$$

Within-group SS

$$SS_W = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \varepsilon_{ij}$$

Among-group SS

$$SS_A = SS_T - SS_W$$

Anderson MJ (2001) Austral Ecology

サルの食性幅とその変異性の解析（PERMANOVAとPERMDISP）

分散分析表

	Df	SS	MS	F	R^2	P
Group	2	1.13	0.57	3.01	0.30	0.005
Residual	14	2.64	0.19		0.70	
Total	16	3.78			1.00	

$$SS_A = SS_T - SS_W$$

$$SS_W = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \varepsilon_{ij}$$

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2$$

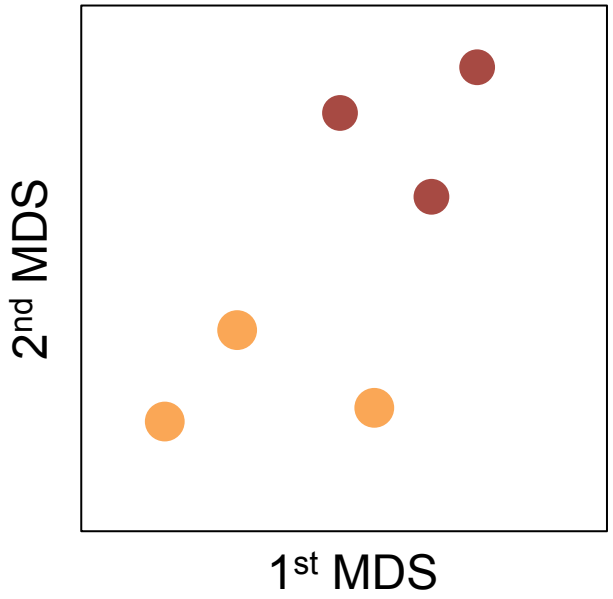
$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

グループのラベルをシャッフルして
F値を繰り返し計算
→　実測値と比較し、並べ替え検定

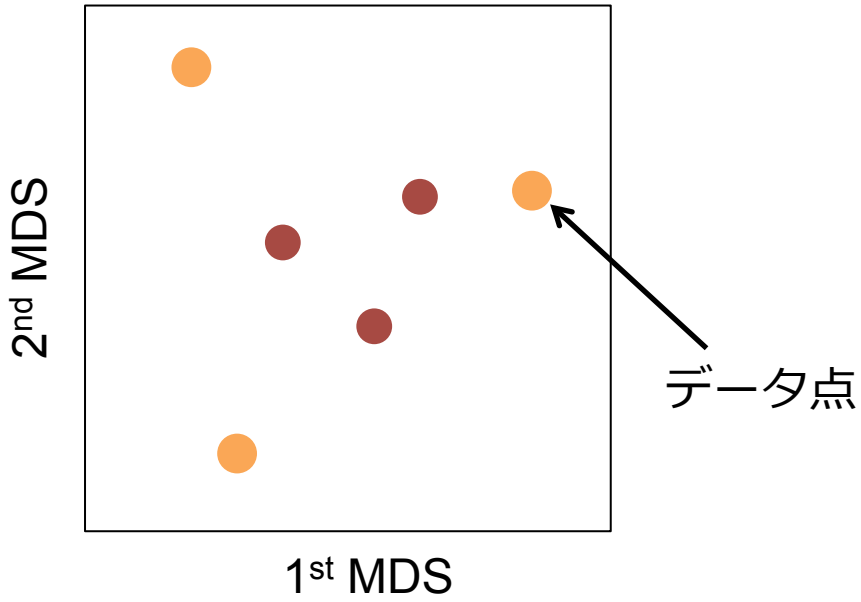
Anderson MJ (2001) Austral Ecology

PERMANOVAの結果解釈における注意点：平均 vs 分散の効果

(i)処理区間で平均的な群集構造が異なる場合



(ii) 処理区ごとに群集構造の分散が異なる場合

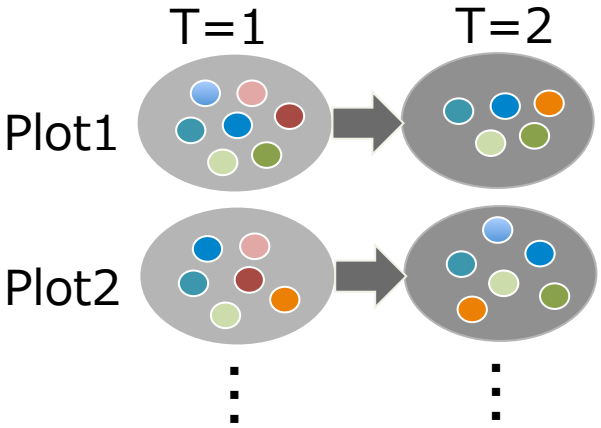


	シナリオ(i)	シナリオ(ii)
PERMANOVA	有意	有意
PERMDISP	有意でない	有意

Anderson MJ (2001) Austral Ecology

魚類のメタ個体群動態パターンの解析

環境DNAの出現パターン（局所移入・絶滅率に影響する要因を特定する方法



オッズ比

$$\psi = \frac{ad}{bc}$$

生息地の水中浮遊DNAから魚種特定を行う

	在 (T_2=1)	不在 (T_2=0)
在 (T_1=1)	a	b
不在 (T_1 =0)	c	d

← 各々のOTUについてこの表を得る
同時にT=1, 2における環境要因を測定

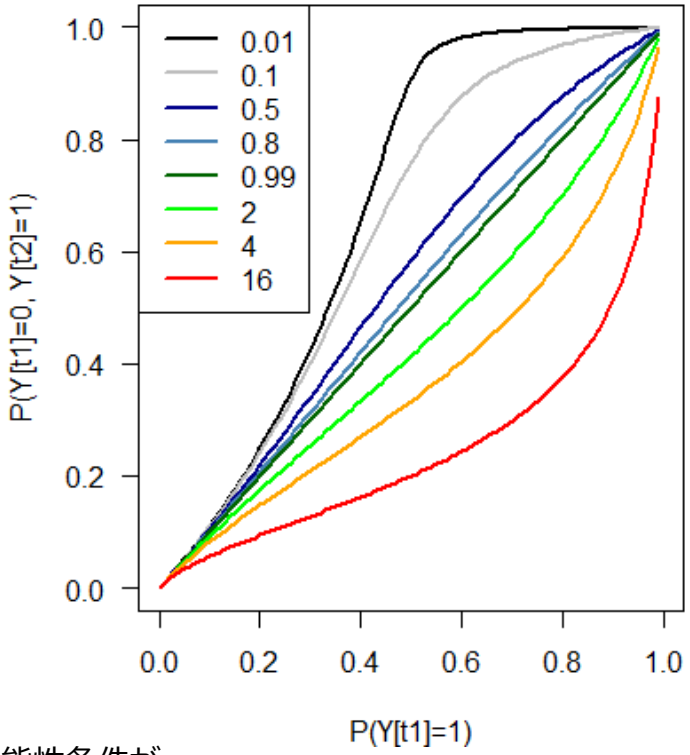
McCullagh and Nelder (1989) ; Yee & Dirnböck (2009) Journal of Theoretical Biology

魚類のメタ個体群動態パターンの解析

オッズ比 ＝ 局所的移入率・絶滅率の形を特徴づける変数

局所的**移入**率

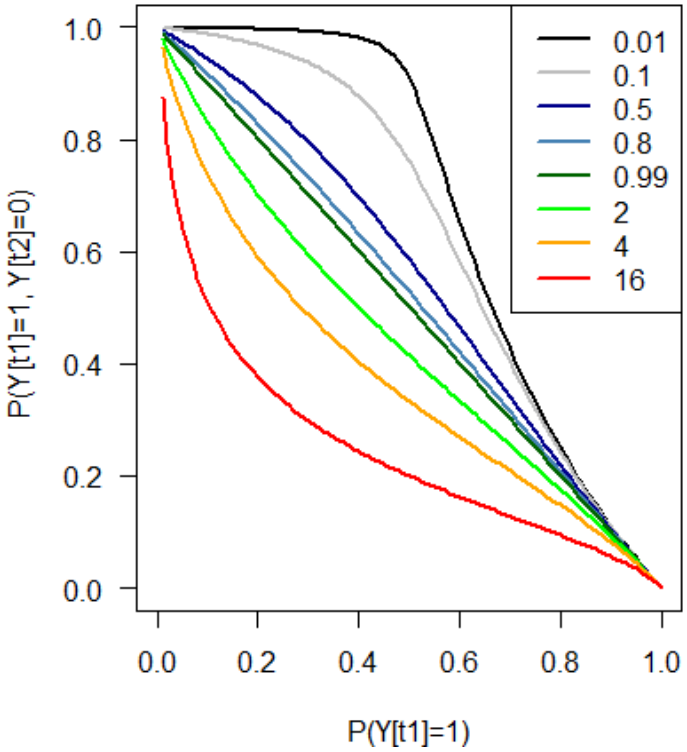
0→1になる確率



周辺確率

局所的**絶滅**率

1→0 になる確率



周辺確率

*交換可能性条件が満たされる場合

Yee & Dirnböck (2009)

```
> vglm(..., family = binom2.or("cloglog", exchangeable = TRUE))
```

魚類のメタ個体群動態パターンの解析

二変量回帰モデル

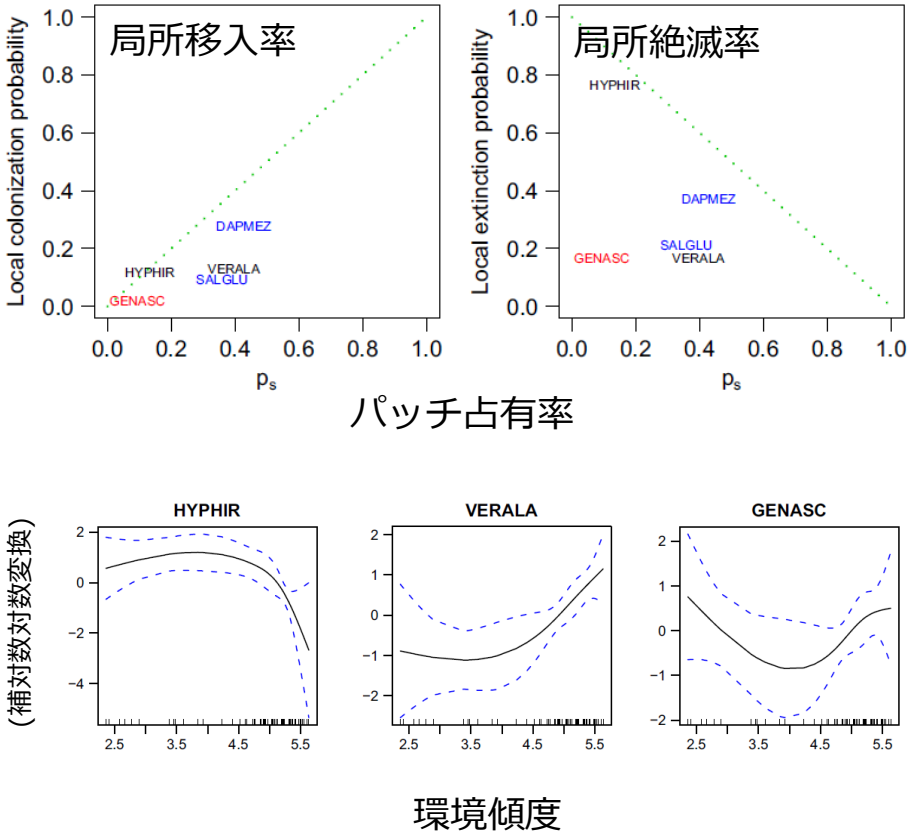
$$cloglog\ \mathrm{P}(Y_1 = 1|x) = \eta_1$$
$$cloglog \mathrm{P}(Y_2 = 1|x) = \eta_2$$
$$log\ \psi(x) = \eta_3$$

←ポアソン分布を仮定

ηをそれぞれ柔軟にモデリングすることで生物のニッチと環境変化に伴うメタ個体群過程を評価

例）
リンク関数を式ごとに変える
回帰項を式ごとに変える (平行性の仮定)
縮小ランク回帰やゼロ過多モデル
ベクトル一般化線形モデル、など

推定されたメタ個体群動態と種のニッチ



Yee & Dirnböck (2009)

```
> vglm(..., family = binom2.or("cloglog", exchangeable = TRUE))
```

オッズ比に関する注意点：Simpson‘s Paradox

重要な未知の要因 Z が絶滅・移入過程に関与する場合

(i) 交絡因子Zを考慮した場合

	Z+		Z-	
	在T_2	不在T_2	在T_2	不在T_2
在 T_1	80	16	50	452
不在 T_1	160	160	10	452
		ψ= 5		ψ= 5

(ii) 交絡因子Zを考慮しない場合

	在T_2	不在T_2
在 T_1	130	468
不在T_1	170	612
	ψ= 1	

母集団全体で見たときのオッズ比
≠分割して見たときのオッズ比

未知の要因Zの例：

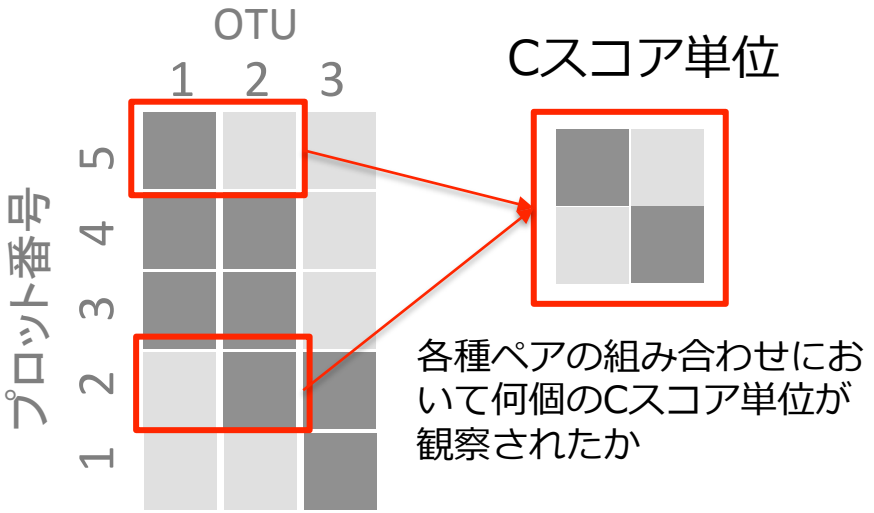
生息地分断化、隠蔽種(OTU)の存在など

柳川 (1985) 離散多変量データの解析

微生物種間の排他的分布の検出方法


木材腐朽菌群集の競争が見られる倒木


化学合成細菌群集の競争が見られる淀み



種*i*と*j*のCスコア $C_{ij} = \frac{(r_i - S_{ij})(r_j - S_{ij})}{(r_i + r_j - S_{ij})}$

群集のCスコア $C = \sum_{j=1}^{M} \sum_{i<j} \frac{C_{ij}}{M(M-1)/2}$

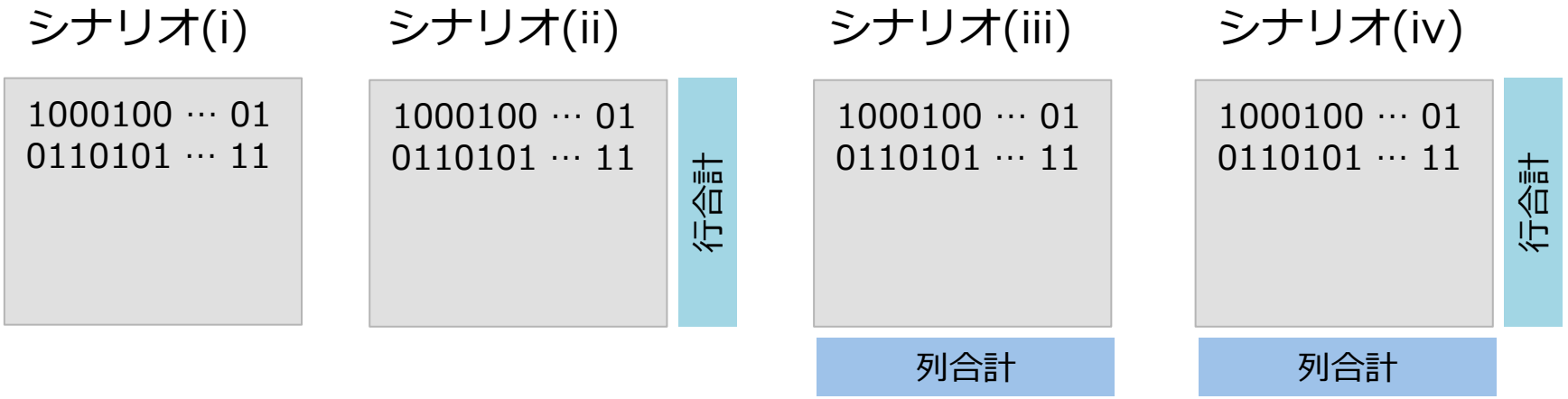
r_i 種iの出現回数
S_{ij} 種*i*と*j*の同時出現回数
M 全体の種の数

Stone and Roberts (1990) Oecologia

排他的分布の検出ー 帰無モデルの選定

帰無モデルをカスタマイズすることが可能

→ Monte Carlo simulation の方法を変更



合計在数を固定し0/1を生成

行合計を固定し無作為に0/1を生成

列合計を固定し無作為に0/1を生成

行と列の合計両方を固定し無作為に0/1を生成

Goteli and Graves (1993) Null model ecology

```
> oecosimu(comm, method ="swap", nsimul = 9999, ...)
```

排他的分布の検出ー 帰無モデル検定

帰無モデルをカスタマイズすることが可能

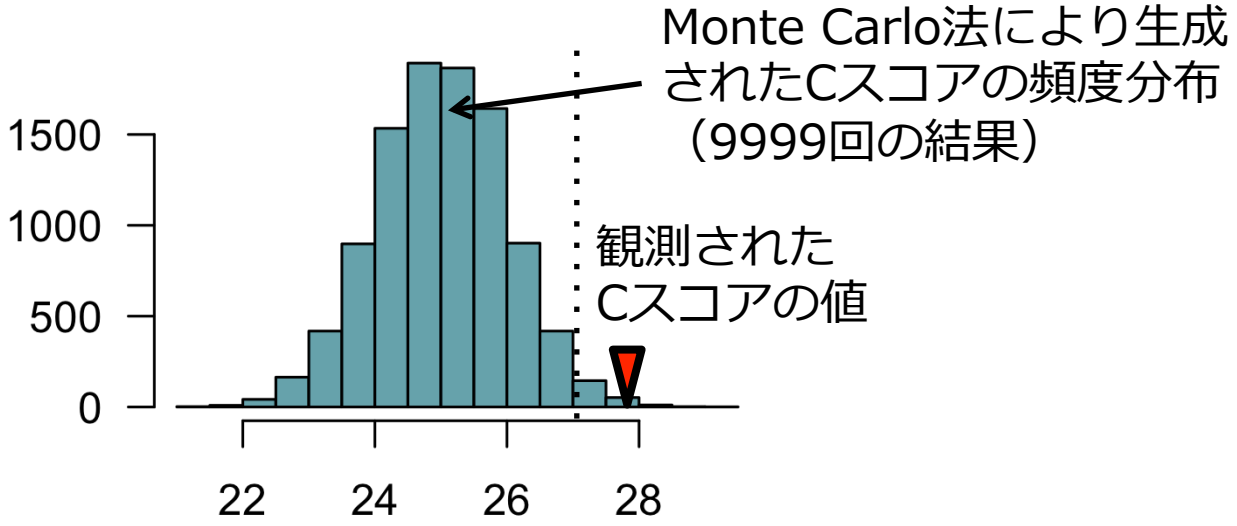
→ Monte Carlo simulation の方法を変更

データセット

1000100 … 01
0110101 … 11

行合計

列合計



$$Effect\ size = \frac{C_{obs} - E(C_{sim})}{SD(C_{sim})}$$

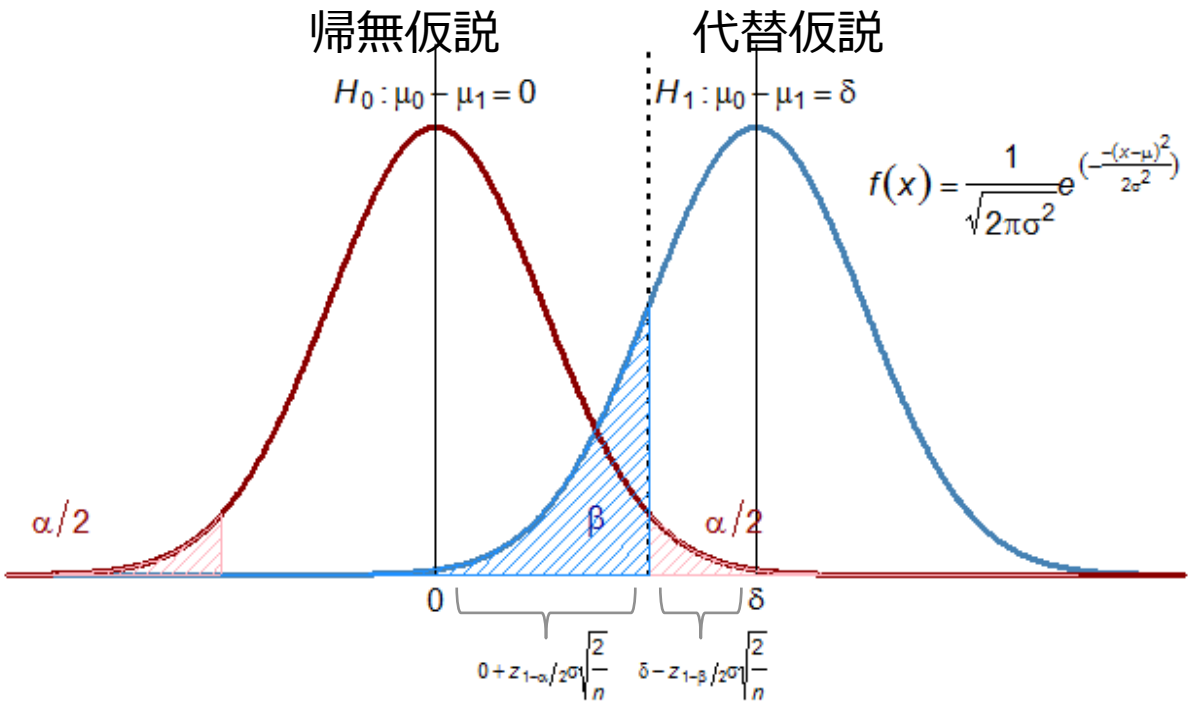
```
> oecosimu(comm, method ="swap", nsimul = 9999, ...)
```

補足：Cスコア検定における2つの過誤

Type I error：　正しい帰無仮説を誤って棄却する確率(α)

Type II error ：誤っている帰無仮説を棄却しない確率(β)

検出力＝１－β ：誤っている帰無仮説を棄却する確率



Van Belle (2008) Statistical rules of thumb

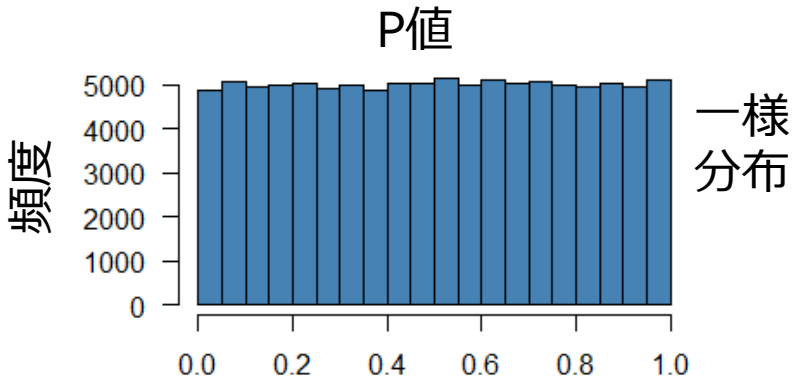
```
> oecosimu(comm, method ="swap", nsimul = 9999, ...)
```

排他的分布の検出ー 多重検定とp値補正

OTUペアごとにCスコア解析 → Type I errorの増大

- Familywise error rate (FWER)　→　Bonferroni補正
  少なくとも1つの帰無仮説を誤って棄却する確率を制御

- False discovery rate (FDR) → Benjamini-Hochberg補正, q値法
  多くの帰無仮説群のうち、誤って棄却される帰無仮説の割合を制御



← ランダムに生成した変数ペア(~N(0,1))を用いて t 検定を繰り返し、その結果得られたp値の分布

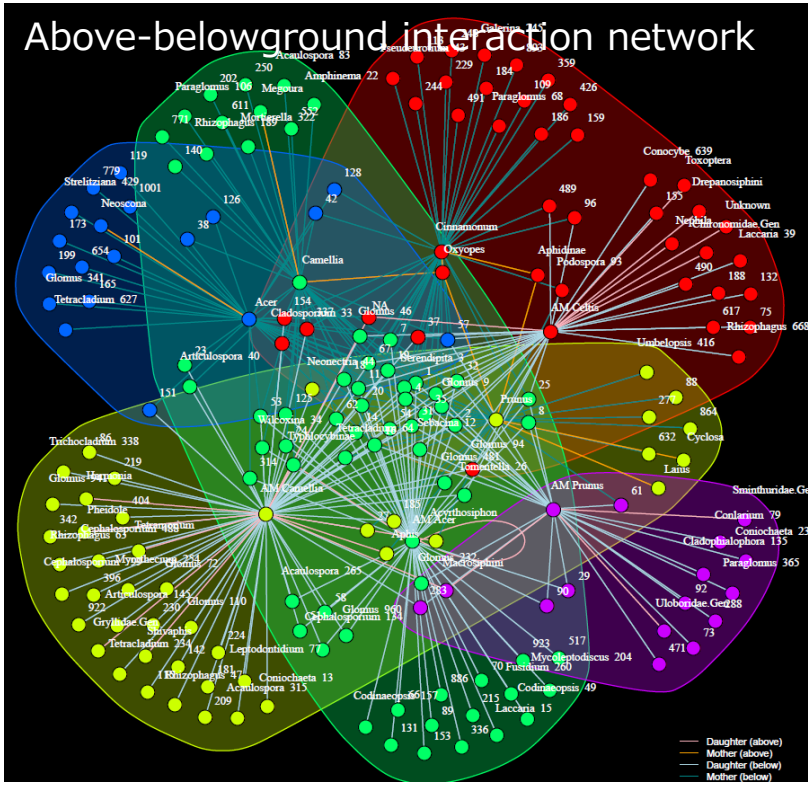
Benjamini and Hochberg (1995) J R Stat Soc; Storey and Tibshirani (2003) PNAS

```
> p.adjust(p, method = "BH")  #see also Package: qvalue
```

高次元データの解析

高次元データ解析の課題

- 高い予測精度を保証すること
  →将来予測が可能となる
- 現象と真に関連する変数を選択すること
  →推定量の漸近正規性が保証されず検定不可能になることを避ける



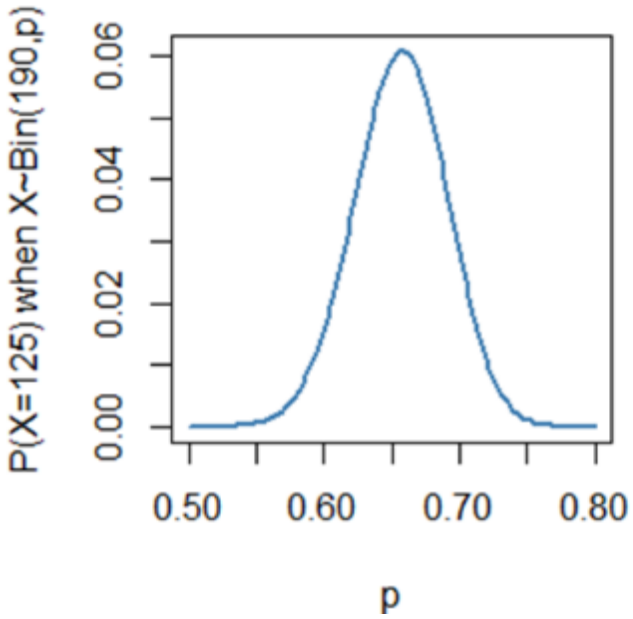
安道 (2014) 高次元データ分析の方法

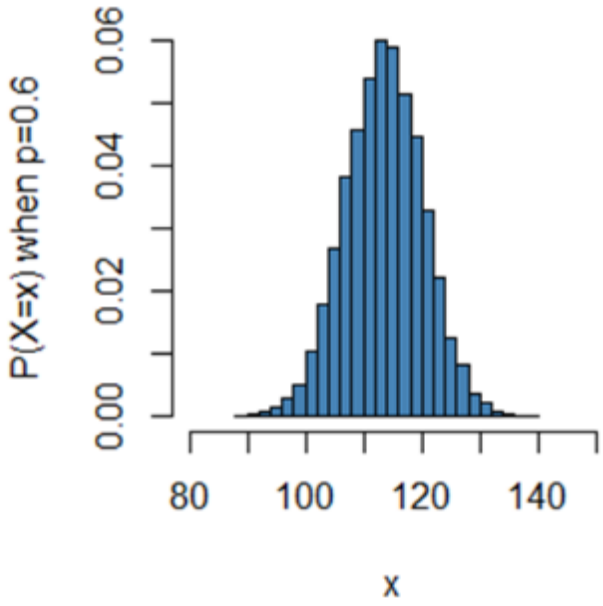

土壌に張り巡らされた菌糸のネットワーク

補足： 尤度関数と確率関数の違い

尤度関数：
あらゆるpにおいて特定のxの値が観察される確率 (xを固定)

確率関数：
特定のpにおいて観察されるあらゆる異なるxの値がとりうる確率(pを固定)





←二項分布の例

$$\mathbf{E}(X) = np$$

$$\mathbf{Var}(X) = np(1-p)$$

$$L(p;125)={}_{190}\mathrm{C}_{125}\, p^{125}(1-p)^{65}$$

lasso推定による高次元データの解析

least absolute shrinkage and selection operator

植物の葉のC：N比に影響を与える根圏微生物OTUを選択

y =
葉のC:N比
(元素分析)

X=
根圏微生物
OTUデータ
(メタゲノム)

アカマツの実生

$$l(\beta) = \frac{1}{n}\sum_{\alpha=1}^{n} \log f(y_\alpha | x_\alpha, \beta) - \lambda \sum_{j=1}^{p} |\beta_j|$$

尤度関数　　罰則項

正則化パラメータ

罰則付き尤度関数を最大とするβを
見つけることで、説明変数選択と
回帰係数βの推定を同時に実行

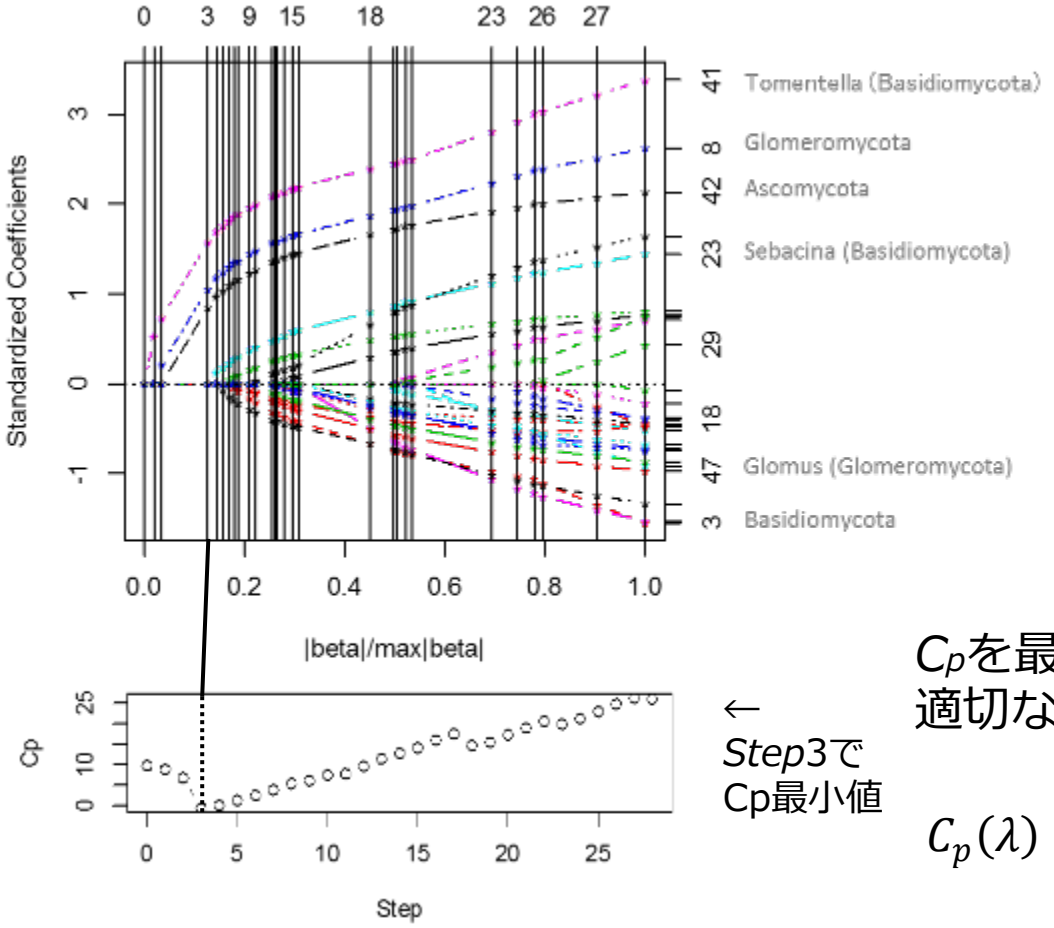
Tibshirani 1996 J Royal Stat Soc; 安道 (2014) 高次元データ分析の方法

Package lars:　lars

lasso推定による高次元データの解析

least absolute shrinkage and selection operator

植物の葉のC：N比に影響を与える根圏微生物OTUを選択



←
*Step*3で
Cp最小値

一般線形モデルにより
植物に影響を与える真菌OTU
候補を絞り込む

例．ラシャタケ類(Tomentella)が
寄主植物の葉のC:N比を上昇
させる可能性を示唆

C_pを最小化する正則化パラメータを
適切な値として解釈する

$$C_p(\lambda) = \frac{1}{n}\sum_{\alpha=1}^{n}\{y_k - \hat{\mu}_\alpha(\lambda)\}^2 + \frac{2}{n}\hat{p}\sigma^2$$

Package lars:　lars

メタゲノム生態学の展開

- メタゲノム・環境DNAに対する高い関心とその応用可能性
- メタゲノム解析は、これまでの生態学を置き換えるものではなく、フィールドワークに基づく知見のさらなる強化を可能とする
- 生態学者が初めて直面するデータ解析の問題（高次元・スパースデータ）
- サンプリングの時空間スケール、反復の採り方、ランダム化と層別化などに関する実験生態学的な知見と結びけることでさらなる力を発揮する

メタゲノム生態学の展開

- 今後は、生態学的なプロセス（相互作用や共存機構）の解明、および、その予測を視野に入れた、時系列データの蓄積が最も重要となる

- 環境DNAを用いたレジリエンス(resilience)や時間的安定性(stability)の定量化、早期警告シグナルの確立などが求められる