

Introduction to working with multivariate data

Multiple responses



Introduction to working with multivariate data

Multiple responses

e.g.,

Do biotic communities differ between 2 site types?

area	JuvenileLunata	NotoBigNota	Luna	SmallBuena	BigBuena	NotoUndulata	rsNewts	Hydrophilids	PredDivingB eetlesSmall	PredDivingB eetlesLarge	PredDivingB eetleLarvae
restoration	0.666667		15	0	0	0	9.333333	0.666667	0	0	0
control		0	5.666667	0	0	0.666667	2.333333	0.333333	0	0	0
restoration		0	4.333333	0	0	0	2.666667	1.333333	0	0.333333	0
control		0	1.333333	0	0	0	1.333333	0.666667	0	0	0
restoration		0	11.33333	0	0	0	10.66667	1.666667	0	0	0
control		0	5.333333	0	0	0	3.333333	0	0	0	0
restoration		0	6	0	0	0	1.666667	3.666667	1	0.333333	0
control		0	2	0	0	0	0.666667	0	0	0	0
restoration	6.333333		16	0	0	0	4.333333	1	0.333333	0	0
control		0	2.333333	0	0	0	1.666667	0	0	0	0
restoration	1.333333		4.666667	0	0	0	1.333333	7.333333	0.333333	0	0
control		0	3.666667	0	0	0	0.333333	0.333333	0	0	0
restoration		1	9.333333	0	0	0.333333	6	0.666667	0	0	0
control	0.333333		3.333333	0	0	0	2.333333	0	0	0	0
restoration	2.333333		2.333333	0	0	0	0.333333	2.666667	0.666667	0.333333	0
control	0.666667		1.666667	0	0	0	1.333333	0.333333	0	0	0
restoration		2	4	0	0	0	17.66667	0.333333	0	0	0
control		0	3.333333	0	0	0	3.666667	0	0	0	0
restoration	1.666667		2	0	0	0	4	2.333333	0	0	0
control		0	0.333333	0	0	0	2.333333	1	0	0	0
restoration	6.333333		3.333333	0	0	0	11	0.333333	0	0.333333	0
control		0.5	2	0	0	0	1.333333	0	0	0	0

Introduction to working with multivariate data

Multiple responses

e.g.,

Do biotic communities differ between 2 site types?

area	JuvenileLunata	NotoBigNota	Luna	SmallBuena	BigBuena	NotoUndulata	rsNewts	Hydrophilids	PredDivingB eetlesSmall	PredDivingB eetlesLarge	PredDivingB eetleLarvae
restoration	0.666667		15	0	0	0	9.333333	0.666667	0	0	0
control	0	5.666667		0	0	0.666667	2.333333	0.333333	0	0	0
restoration	0	4.333333		0	0	0	2.666667	1.333333	0	0.333333	0
control	0	1.333333		0	0	0	1.333333	0.666667	0	0	0
restoration	0	11.33333		0	0	0	10.66667	1.666667	0	0	0
control	0	5.333333		0	0	0	3.333333	0	0	0	0
restoration	0	6		0	0	0	1.666667	3.666667	1	0.333333	0
control	0	2		0	0	0	0.666667	0	0	0	0
restoration	6.333333		16	0	0	0	4.333333	1	0.333333	0	0
control	0	2.333333		0	0	0	1.666667	0	0	0	0
restoration	1.333333	4.666667		0	0	0	1.333333	7.333333	0.333333	0	0
control	0	3.666667		0	0	0	0.333333	0.333333	0	0	0
restoration	1	9.333333		0	0	0.333333	6	0.666667	0	0	0
control	0.333333	3.333333		0	0	0	2.333333	0	0	0	0
restoration	2.333333	2.333333		0	0	0	0.333333	2.666667	0.666667	0.333333	0
control	0.666667	1.666667		0	0	0	1.333333	0.333333	0	0	0
restoration	2	4		0	0	0	17.66667	0.333333	0	0	0
control	0	3.333333		0	0	0	3.666667	0	0	0	0
restoration	1.666667	2		0	0	0	4	2.333333	0	0	0
control	0	0.333333		0	0	0	2.333333	1	0	0	0
restoration	6.333333	3.333333		0	0	0	11	0.333333	0	0.333333	0
control	0.5	2		0	0	0	1.333333	0	0	0	0

75 more
Taxa
counted
-
columns
not
shown...

200+ more samples taken – rows not shown...

Introduction to working with multivariate data

Options...

- Focus on one variable
- Use multivariate ANOVA (MANOVA)

area	JuvenileNotoLunata	BigNotoLunata	SmallBuena	BigBuena	NotoUndulata	rsNewts	Hydrophilids	PredDivingB eetlesSmall	PredDivingB eetlesLarge	PredDivingB eetleLarvae
restoration	0.666667	15	0	0	0	9.333333	0.666667	0	0	0
control	0	5.666667	0	0	0.666667	2.333333	0.333333	0	0	0
restoration	0	4.333333	0	0	0	2.666667	1.333333	0	0.333333	0
control	0	1.333333	0	0	0	1.333333	0.666667	0	0	0
restoration	0	11.33333	0	0	0	10.66667	1.666667	0	0	0
control	0	5.333333	0	0	0	3.333333	0	0	0	0
restoration	0	6	0	0	0	1.666667	3.666667	1	0.333333	0
control	0	2	0	0	0	0.666667	0	0	0	0
restoration	6.333333	16	0	0	0	4.333333	1	0.333333	0	0
control	0	2.333333	0	0	0	1.666667	0	0	0	0
restoration	1.333333	4.666667	0	0	0	1.333333	7.333333	0.333333	0	0
control	0	3.666667	0	0	0	0.333333	0.333333	0	0	0
restoration	1	9.333333	0	0	0.333333	6	0.666667	0	0	0
control	0.333333	3.333333	0	0	0	2.333333	0	0	0	0
restoration	2.333333	2.333333	0	0	0	0.333333	2.666667	0.666667	0.333333	0
control	0.666667	1.666667	0	0	0	1.333333	0.333333	0	0	0
restoration	2	4	0	0	0	17.66667	0.333333	0	0	0
control	0	3.333333	0	0	0	3.666667	0	0	0	0
restoration	1.666667	2	0	0	0	4	2.333333	0	0	0
control	0	0.333333	0	0	0	2.333333	1	0	0	0
restoration	6.333333	3.333333	0	0	0	11	0.333333	0	0.333333	0
control	0.5	2	0	0	0	1.333333	0	0	0	0

75 more
Taxa
counted
-
columns
not
shown...

200+ more samples taken – rows not shown...

Introduction to working with multivariate data

Options...

- Focus on one variable
- Use multivariate ANOVA (MANOVA)

MANOVA

Assumptions of multivariate normality, homoscedasticity, & equal covariance (correlation b/w 2 variables is same in all groups)

Introduction to working with multivariate data

Options...

- Focus on one variable
- Use multivariate ANOVA (MANOVA)

MANOVA

Assumptions of multivariate normality, homoscedasticity, & equal covariance (correlation b/w 2 variables is same in all groups)

Needs far more samples than responses in order to retain power

Introduction to working with multivariate data

Options...

- Focus on one variable
- Use multivariate ANOVA (MANOVA)

MANOVA

Assumptions of multivariate normality, homoscedasticity, & equal covariance (correlation b/w 2 variables is same in all groups)

Needs far more samples than responses in order to retain power

Most standard statistical packages will do MANOVA & texts have good descriptions

Gotelli & Ellison, Zar, Sokal & Rohlf

Introduction to working with multivariate data

Options...

- More multivariate approaches that allow you to consider many responses at once

area	JuvenileNoto Lunata	BigNotoLuna ta	SmallBuena	BigBuena	NotoUndulat a	rsNewts	Hydrophilids	PredDivingB eetlesSmall	PredDivingB eetlesLarge	PredDivingB eetleLarvae
restoration	0.666667	15	0	0	0	9.333333	0.666667	0	0	0
control	0	5.666667	0	0	0.666667	2.333333	0.333333	0	0	0
restoration	0	4.333333	0	0	0	2.666667	1.333333	0	0.333333	0
control	0	1.333333	0	0	0	1.333333	0.666667	0	0	0
restoration	0	11.33333	0	0	0	10.66667	1.666667	0	0	0
control	0	5.333333	0	0	0	3.333333	0	0	0	0
restoration	0	6	0	0	0	1.666667	3.666667	1	0.333333	0
control	0	2	0	0	0	0.666667	0	0	0	0
restoration	6.333333	16	0	0	0	4.333333	1	0.333333	0	0
control	0	2.333333	0	0	0	1.666667	0	0	0	0
restoration	1.333333	4.666667	0	0	0	1.333333	7.333333	0.333333	0	0
control	0	3.666667	0	0	0	0.333333	0.333333	0	0	0
restoration	1	9.333333	0	0	0.333333	6	0.666667	0	0	0
control	0.333333	3.333333	0	0	0	2.333333	0	0	0	0
restoration	2.333333	2.333333	0	0	0	0.333333	2.666667	0.666667	0.333333	0
control	0.666667	1.666667	0	0	0	1.333333	0.333333	0	0	0
restoration	2	4	0	0	0	17.66667	0.333333	0	0	0
control	0	3.333333	0	0	0	3.666667	0	0	0	0
restoration	1.666667	2	0	0	0	4	2.333333	0	0	0
control	0	0.333333	0	0	0	2.333333	1	0	0	0
restoration	6.333333	3.333333	0	0	0	11	0.333333	0	0.333333	0
control	0.5	2	0	0	0	1.333333	0	0	0	0

Introduction to working with multivariate data

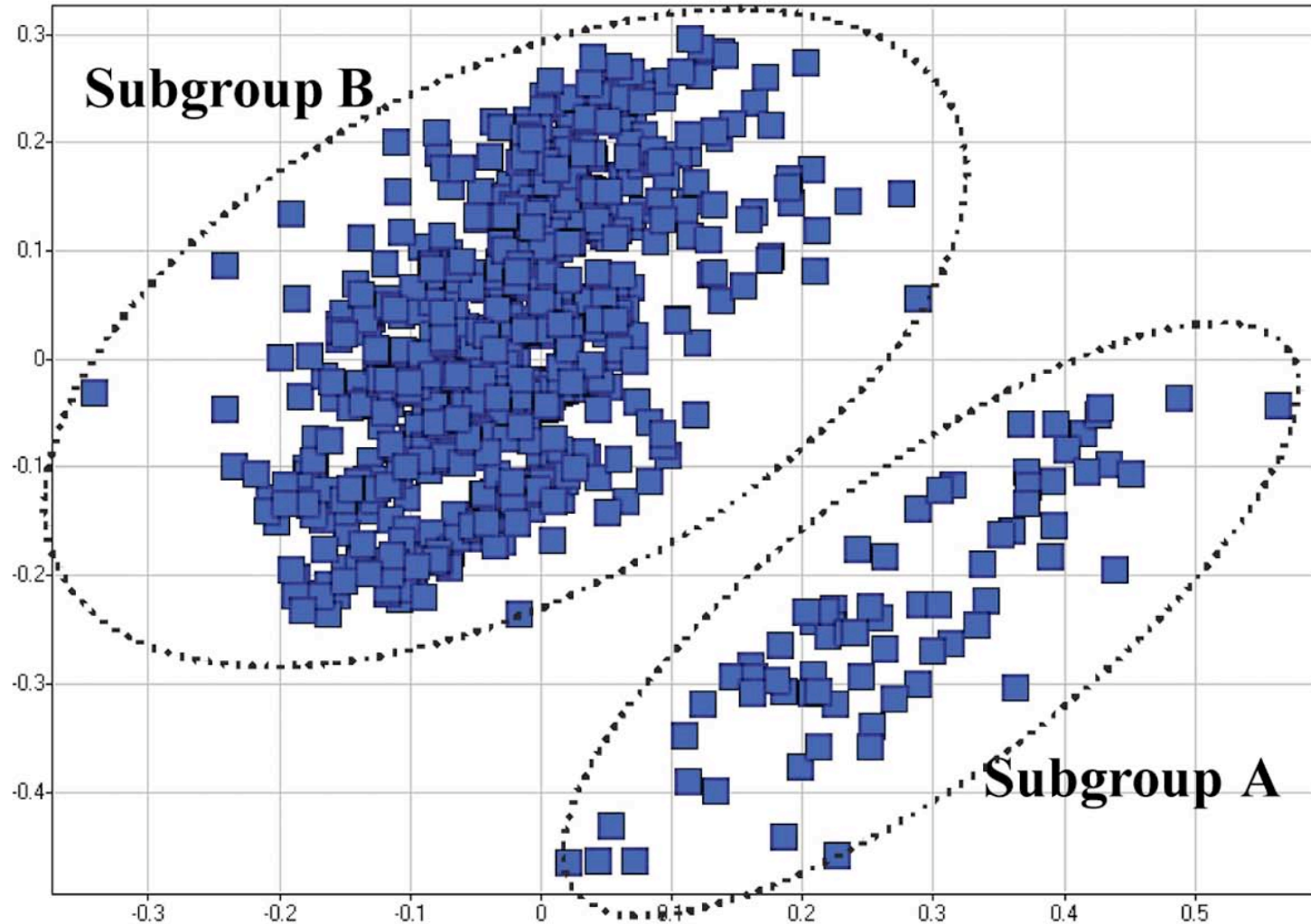
Options...

- More multivariate approaches that allow you to consider many responses at once

And that are more robust to the “messy” data that are common from observational studies

area	JuvenileNoto Lunata	BigNotoLuna ta	SmallBuena BigBuena	BigBuena	NotoUndulat a	rsNewts	Hydrophilids	PredDivingB eetlesSmall	PredDivingB eetlesLarge	PredDivingB eetleLarvae	
restoration	0.666667	15	0	0	0	9.333333	0.666667	0	0	0	0
control	0	5.666667	0	0	0.666667	2.333333	0.333333	0	0	0	0
restoration	0	4.333333	0	0	0	2.666667	1.333333	0	0.333333	0	0
control	0	1.333333	0	0	0	1.333333	0.666667	0	0	0	0
restoration	0	11.33333	0	0	0	10.66667	1.666667	0	0	0	0
control	0	5.333333	0	0	0	3.333333	0	0	0	0	0
restoration	0	6	0	0	0	1.666667	3.666667	1	0.333333	0	0
control	0	2	0	0	0	0.666667	0	0	0	0	0
restoration	6.333333	16	0	0	0	4.333333	1	0.333333	0	0	0
control	0	2.333333	0	0	0	1.666667	0	0	0	0	0
restoration	1.333333	4.666667	0	0	0	1.333333	7.333333	0.333333	0	0	0
control	0	3.666667	0	0	0	0.333333	0.333333	0	0	0	0
restoration	1	9.333333	0	0	0.333333	6	0.666667	0	0	0	0
control	0.333333	3.333333	0	0	0	2.333333	0	0	0	0	0
restoration	2.333333	2.333333	0	0	0	0.333333	2.666667	0.666667	0.333333	0	0
control	0.666667	1.666667	0	0	0	1.333333	0.333333	0	0	0	0
restoration	2	4	0	0	0	17.66667	0.333333	0	0	0	0
control	0	3.333333	0	0	0	3.666667	0	0	0	0	0
restoration	1.666667	2	0	0	0	4	2.333333	0	0	0	0
control	0	0.333333	0	0	0	2.333333	1	0	0	0	0
restoration	6.333333	3.333333	0	0	0	11	0.333333	0	0.333333	0	0
control	0.5	2	0	0	0	1.333333	0	0	0	0	0

1. Similarity & dissimilarity metrics: the foundation
2. Basic analysis of multivariate data
3. Visualizing multivariate data



1. Similarity & dissimilarity metrics: the foundation

Many multivariate methods quantify the difference among individual samples, observations, or groups – frequently expressed as “**distance**” between observations in multivariate space. The inverse of distance or **dissimilarity** is **similarity**.

How similar are samples based on not only who/what is there, but also how many or how much?

1. Similarity & dissimilarity metrics: the foundation
Consider first a simple case – the multivariate
distance between 2 individual plants based on
2 characteristics

Plant	Variable 1- height	Variable 2 – leaf spread
Plant A	654	55
Plant B	413	60

1. Similarity & dissimilarity metrics: the foundation
Consider first a simple case – the multivariate
distance between 2 individual plants based on
2 characteristics

Use the Pythagorean theorem to calculate

Euclidean distance

$$d_{ij} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2}$$

Plant	Variable 1- height	Variable 2 – leaf spread
Plant A	654	55
Plant B	413	60

1. Similarity & dissimilarity metrics: the foundation
Consider first a simple case – the multivariate
distance between 2 individual plants based on
2 characteristics

Use the Pythagorean theorem to calculate
Euclidean distance

$$d_{ij} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2}$$

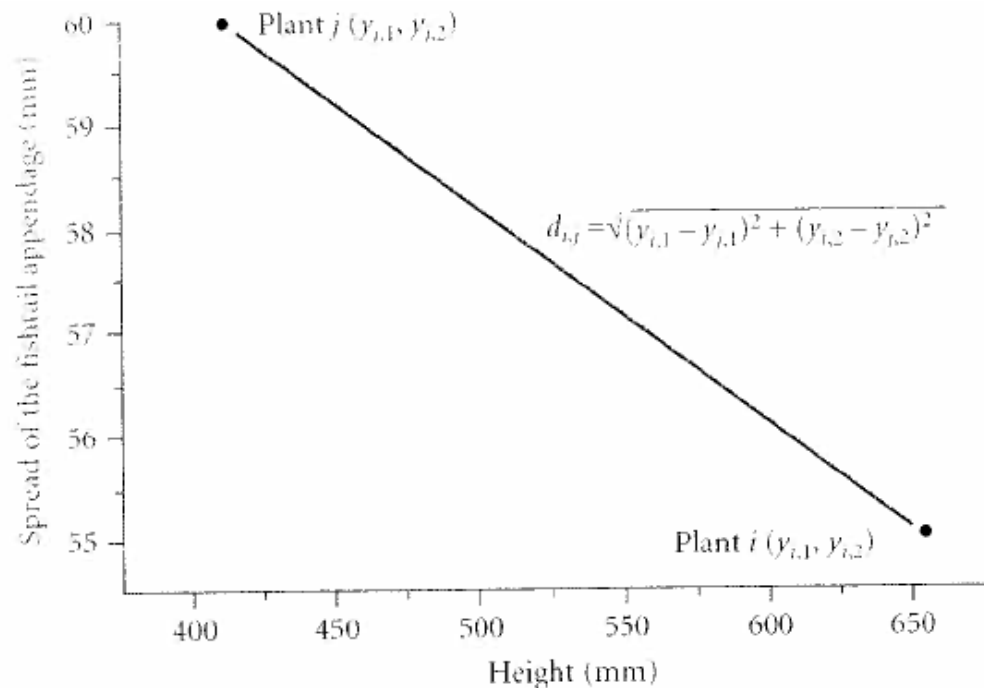
Plant	Variable 1- height	Variable 2 – leaf spread
Plant A	654	55
Plant B	413	60

$$D_{\text{between A \& B}} = \sqrt{(654 - 413)^2 + (55 - 60)^2} = 241.05$$

Euclidean distance

$$d_{ij} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2}$$

Plant	Variable 1- height	Variable 2 – leaf spread
Plant A	654	55
Plant B	413	60



$$D_{\text{between A \& B}} = \sqrt{(654 - 413)^2 + (55 - 60)^2} = 241.05$$

1. Similarity & dissimilarity metrics: the foundation



Plant	Variable 1- height	Variable 2 – leaf spread
Plant A	654	55
Plant B	413	60

$$D_{\text{between A \& B}} = \sqrt{(654 - 413)^2 + (55 - 60)^2} = 241.05$$

1. Similarity & dissimilarity metrics: the foundation



Site	Variable 1- Sage	Variable 2 – Brome
Restored	654	55
Control	413	60

$$D_{\text{between R \& C}} = \sqrt{(654 - 413)^2 + (55 - 60)^2} = 241.05$$

1. Similarity & dissimilarity metrics: the foundation

The difference between two samples with multiple responses can be expressed as a single number

Site	Variable 1- Sage	Variable 2 – Brome
Restored	654	55
Control	413	60

$$D_{\text{between R \& C}} = \sqrt{(654 - 413)^2 + (55 - 60)^2} = \mathbf{241.05}$$

1. Similarity & dissimilarity metrics: the foundation

Easy to expand to more variables

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
Restored	654	55	38
Control	413	60	22

$$D = \sqrt{(654 - 413)^2 + (55 - 60)^2 + (38 - 22)^2} = \mathbf{241.58}$$

1. Similarity & dissimilarity metrics: the foundation

Notice adding the 3rd variable doesn't change Euclidean distance very much

3rd variable has smaller mean and variance, compared to 1st variable – for this reason, standardizing variables can be desirable

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
Restored	654	55	38
Control	413	60	22

$$D = \sqrt{(654 - 413)^2 + (55 - 60)^2 + (38 - 22)^2} = \mathbf{241.58}$$

1. Similarity & dissimilarity metrics: the foundation

Standardizing using the Z-score

$$\text{Z-score} = (Y_i - \bar{Y}) / \text{sd}$$

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
RestoredA	654	55	38
ControlA	413	60	22
<i>more...</i>
Mean(sd)	415(90)	51(8)	37(6)

1. Similarity & dissimilarity metrics: the foundation

Standardizing using the Z-score

$$Z\text{-score} = (Y_i - \bar{Y}) / sd$$

$$Z_{\text{restoredA,var1}} = (654 - 415) / 90 = 2.66$$

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
RestoredA	654	55	38
ControlA	413	60	22
<i>more...</i>
Mean(sd)	415(90)	51(8)	37(6)

1. Similarity & dissimilarity metrics: the foundation

Standardizing using the Z-score

$$Z\text{-score} = (Y_i - \bar{Y}) / \text{sd}$$

$$Z_{\text{restoredA, var1}} = (654 - 415) / 90 = 2.66$$

Site	Variable 1 - Sage	Variable 2 - Brome	Variable 3 - Scrub Oak
RestoredA	2.66	0.5	0.17
ControlA	-0.02	1.13	-2.5
<i>more...</i>

1. Similarity & dissimilarity metrics: the foundation

Standardizing using the Z-score

$$Z\text{-score} = (Y_i - \bar{Y}) / \text{sd}$$

$$Z_{\text{restoredA, var1}} = (654 - 415) / 90 = 2.66$$

Site	Variable 1 - Sage	Variable 2 - Brome	Variable 3 - Scrub Oak
RestoredA	2.66	0.5	0.17
ControlA	-0.02	1.13	-2.5
<i>more...</i>

$$D_{\text{restA, conA}} = \sqrt{(2.66 + 0.02)^2 + (0.5 - 1.13)^2 + (0.17 + 2.5)^2} = \mathbf{3.88}$$

1. Similarity & dissimilarity metrics: the foundation

Euclidean distance may not always be the best

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

1. Similarity & dissimilarity metrics: the foundation
Euclidean distance may not always be the best
With Euclidean distance, 2 sites with no species in
common may have smaller Euclidean distance
than 2 sites that share some species

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

1. Similarity & dissimilarity metrics: the foundation
Euclidean distance may not always be the best
With Euclidean distance, 2 sites with no species in
common may have smaller Euclidean distance
than 2 sites that share some species

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

$$D_{\text{siteA,siteB}} = \sqrt{(0-1)^2 + (1-0)^2 + (1-0)^2} = \mathbf{1.73}$$

1. Similarity & dissimilarity metrics: the foundation
Euclidean distance may not always be the best
With Euclidean distance, 2 sites with no species in
common may have smaller Euclidean distance
than 2 sites that share some species

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

$$D_{\text{siteA,siteB}} = \sqrt{(0-1)^2 + (1-0)^2 + (1-0)^2} = \mathbf{1.73}$$

$$D_{\text{siteA,siteC}} = \sqrt{(0-0)^2 + (1-4)^2 + (1-4)^2} = \mathbf{4.24}$$

Euclidean distance

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

e.g., $D_{\text{siteA,siteC}} = \sqrt{(0-0)^2 + (1-4)^2 + (1-4)^2} = \mathbf{4.24}$

Site	Site A	Site B	Site C
Site A	0	1.73	4.24
Site B	1.73	0	5.75
Site C	4.24	5.75	0

Other metrics may be better at times

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

Bray-Curtis dissimilarity

$$D_{ij} = \frac{\sum |y_{i,1} - y_{j,1}|^2}{\sum (y_{i,1} + y_{j,1})^2}$$

Bray-Curtis similarity

$$D_{ij} = 1 - \left(\frac{\sum |y_{i,1} - y_{j,1}|^2}{\sum (y_{i,1} + y_{j,1})^2} \right)$$

Site	Variable 1 - Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

Bray-Curtis similarity

$$D_{i,j} = 1 - \left(\frac{\sum |y_{i,1} - y_{j,1}|^2}{\sum (y_{i,1} + y_{j,1})^2} \right) \quad D_{A,B} = 1 - \left(\frac{1^2 + 1^2 + 1^2}{1^2 + 1^2 + 1^2} \right)$$

Site	Site A	Site B	Site C
Site A	-	0	
Site B	0	-	
Site C			-

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

Bray-Curtis similarity

$$D_{i,j} = 1 - \left(\frac{\sum |y_{i,1} - y_{j,1}|^2}{\sum (y_{i,1} + y_{j,1})^2} \right) \quad D_{B,C} = 1 - \left(\frac{1^2 + 4^2 + 4^2}{1^2 + 4^2 + 4^2} \right)$$

Site	Site A	Site B	Site C
Site A	-	0	0.64
Site B	0	-	0
Site C	0.64	0	-

Site	Variable 1- Sage	Variable 2 – Brome	Variable 3 - Scrub Oak
A	0	1	1
B	1	0	0
C	0	4	4

Now multivariate similarities & dissimilarities between sites, groups, samples, or individuals are expressed as pairwise comparisons in a single matrix

Site	Site A	Site B	Site C
Site A	-		
Site B	0	-	
Site C	0.64	0	-

1. Similarity & dissimilarity metrics: the foundation

Many different metrics may be used, depending on your data & your questions

TABLE 12.6 Some common measures of distance or dissimilarity used by ecologists

Name	Formula	Property
Euclidean	$d_{i,j} = \sqrt{\sum_{k=1}^n (y_{i,k} - y_{j,k})^2}$	Metric
Manhattan (aka City Block)	$d_{i,j} = \sum_{k=1}^n y_{i,k} - y_{j,k} $	Metric
Chord	$d_{i,j} = \sqrt{2 \times \left(1 - \frac{\sum_{k=1}^n y_{i,k} y_{j,k}}{\sqrt{\sum_{k=1}^n y_{i,k}^2 \sum_{k=1}^n y_{j,k}^2}} \right)}$	Metric
Mahalanobis	$d_{y_i, y_j} = \mathbf{d}_{i,j} \mathbf{V}^{-1} \mathbf{d}_{i,j}^T$ $\mathbf{V} = \frac{1}{m_i + m_j - 2} [(m_i - 1)\mathbf{C}_i + (m_j - 1)\mathbf{C}_j]$	Metric
Chi-square	$d_{i,j} = \sqrt{\sum_{i=1}^m \sum_{j=1}^m y_{ij} \times \left[\sum_{k=1}^n \frac{1}{\sum_{k=1}^n y_{jk}} \times \left(\frac{y_{ik}}{\sum_{k=1}^n y_{ik}} - \frac{y_{jk}}{\sum_{k=1}^n y_{jk}} \right)^2 \right]}$	Metric
Bray-Curtis	$d_{i,j} = \frac{\sum_{k=1}^n y_{i,k} - y_{j,k} }{\sum_{k=1}^n (y_{i,k} + y_{j,k})}$	Semi-metric
Jaccard	$d_{i,j} = \frac{a+b}{a+b+c}$	Metric
Sørensen's	$d_{i,j} = \frac{a+b}{a+b+2c}$	Semi-metric

1. Similarity & dissimilarity metrics: the foundation
Many different metrics may be used, depending
on your data & your questions

Transformations (such as Z transforms) used prior
to calculating the similarity/dissimilarity matrix
can **strongly** influence your results

e.g., You may choose to allow a rare species to
affect your results more strongly by using a 4th
root or log transform

1. Similarity & dissimilarity metrics: the foundation
Many different metrics may be used, depending
on your data & your questions

Transformations (such as Z transforms) used prior
to calculating the similarity/dissimilarity matrix
can **strongly** influence your results

e.g., You may choose to allow a rare species to
affect your results more strongly by using a 4th
root or log transform

Transform responsibly – if unsure, transform to give
equal weight across responses (e.g, Z scores)
(Manly p. 60)