EXAM PRACTICE

- ***12 questions * 4 categories:***

Statistics Background

Multivariate Statistics

Interpret

True / False

Stats 1: What is a Hypothesis?

A testable assertion about how the world works

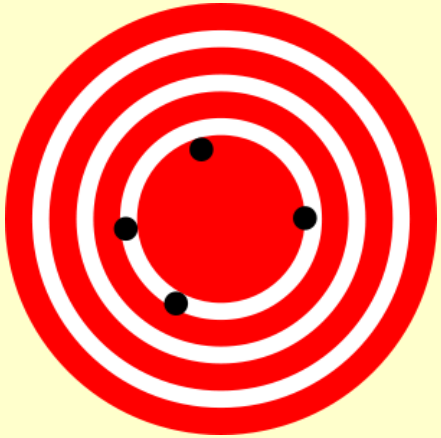
Hypothesis are formulated as the existence / absence of statistical associations between processes (variables).

Stats 2: Define Precision / Accuracy

- Precision:
  Different methods
  give same answer

- Accuracy:

Results approximate
real pattern



Stats 3: What does the Null Hypothesis State?

The null hypothesis (starting point) is that there is no pattern (e.g., no association or response); patterns are random

Stats 4: What does the Alternate Hypothesis State?

The Alternate Hypothesis (Ha) states that there is a significant pattern, a non-random association or response

Stats 5: Define the Type I Error

Type I error: rejection of a true null hypothesis

- occurs at rate chosen for rejection of Ho (alpha = 0.05; 1 in 20)
- rejection also occurs if assumptions of statistical tests violated

Stats 6: Define the Type II Error

Type II error: acceptance of a false null hypothesis

- occurs at rate 1 - βeta
- caused by low power (small sample size, measurement error)

Stats 7: Define the term: "variance"

Variance = sum of squared deviations from mean

---

degrees of freedom

$$Variance = \frac{\sum (Zi - \overline{Z})(Zi - \overline{Z})}{(n-1)}$$

Stats 8: Define the term: "covariance"

Amount of variability shared by two variables

$$Covariance = \frac{\sum (Xi - \overline{X})(Yi - \overline{Y})}{(n-1)}$$

Stats 9: Define the term: "correlation"

- Covariance of X and Y divided by SD in X and SD in Y
- Quantifies intensity of association between two variables

$$r = \frac{Covariance}{\sqrt{(Variance\ X)(Variance\ Y)}}$$

$$r = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum (Xi - \bar{X})^2 \sum (Yi - \bar{Y})^2}}$$

Stats 10: Define: Statistical Significance

The (arbitrary) amount of evidence required to accept that an event is unlikely to have arisen merely by chance is defined the **significance level** or **critical p value**.

The probability of observing data at least as extreme as that observed, *given that the null hypothesis is true*.

If the p-value is small enough, there are 2 possibilities:

- either the null hypothesis is false

  or

- an unusual event has occurred

Stats 11:Orthogonality

Define Orthogonality: (formula and range of values)

$$\text{Orthogonality},\% = 100(1-r^2)$$

Stats 12: Define F / pseudo-F

Define F and pseudo-F:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}},$$

$$F = \frac{\text{explained variance}}{\text{unexplained variance}},$$

MVS1: Explain how Randomization tests work

Multivariate methods allow non-parametric hypothesis testing by creating probability distributions of the resulting statistics

Monte Carlo methods allow the comparison observed statistic against randomized frequency distribution

For example:

1) Calculated ‘observed slope’ using real sequence of seasonal anomalies.
2) Randomly arranged each time series 1000 times, and calculated a distribution of ‘randomized’ slopes.
3) Estimated statistical significance of the ‘observed’ trends by calculating proportion of ‘randomized’ slopes larger in absolute value than the ‘observed’ slope.

MVS2: Explain how the Pearson Correlation works

Pearson correlation: r

measures the direction and strength of the linear relationship between two variables, describing the degree to which one variable is linearly related to another.

Values range from -1 to +1

MVS3: Explain how the partial correlation works

Partial correlation is the correlation of two variables while controlling for a third or more other variables.

- Partial correlation allows us to measure the region of three-way overlap and to remove it from the picture.
- This method determines the value of the correlation between any two of the variables (hypothetically) **if** they were not both correlated with the third variable.
- Mechanistically, this method allows us to determine what the correlation between any two variables would be (hypothetically) **if** the third variable were held constant.

MVS4: Explain how Pearson correlation can be used a distance index

Correlation Coefficient: Species in SU1 vs SU2

Problems: But… -1 < r < +1 (range not from 0 to 1)
And it needs to be flipped (larger r = more similar)

Solution: Correlation coefficient rescaled to distance measure of range 0 - 1 by:

$$r_{\text{distance}} = (1 - r)/2$$

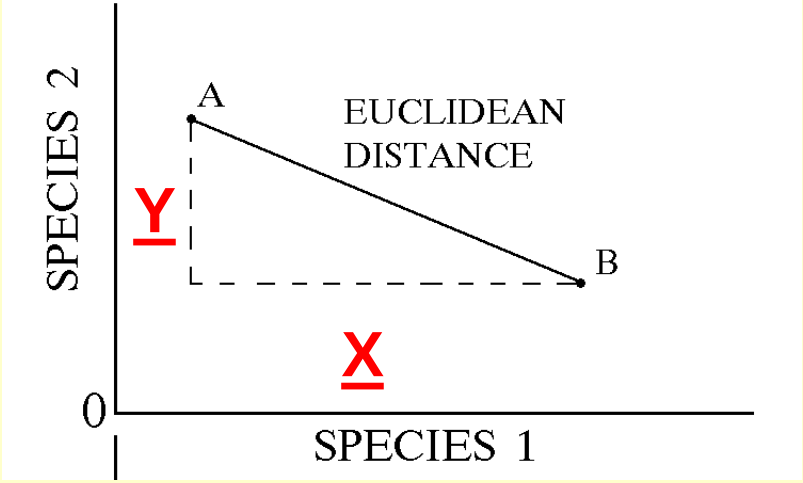
MVS5: Explain how the Euclidean Distance works

Euclidean Distance: Just like hypotenuse of a triangle

$$D = \sqrt[k]{x^k + y^k}$$

$k = 2$ gives Euclidean distance

$k = 1$ gives city-block distance

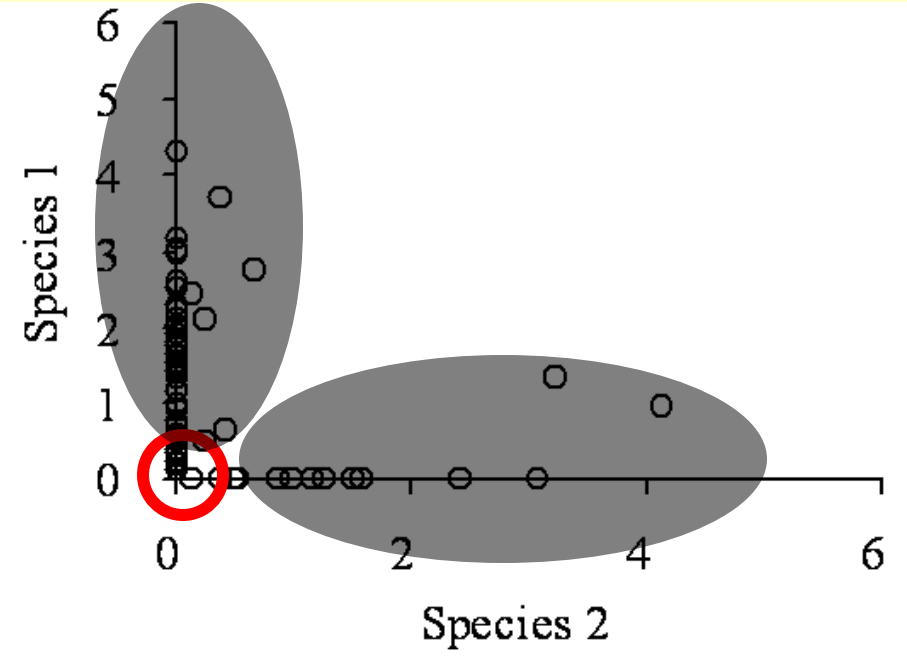


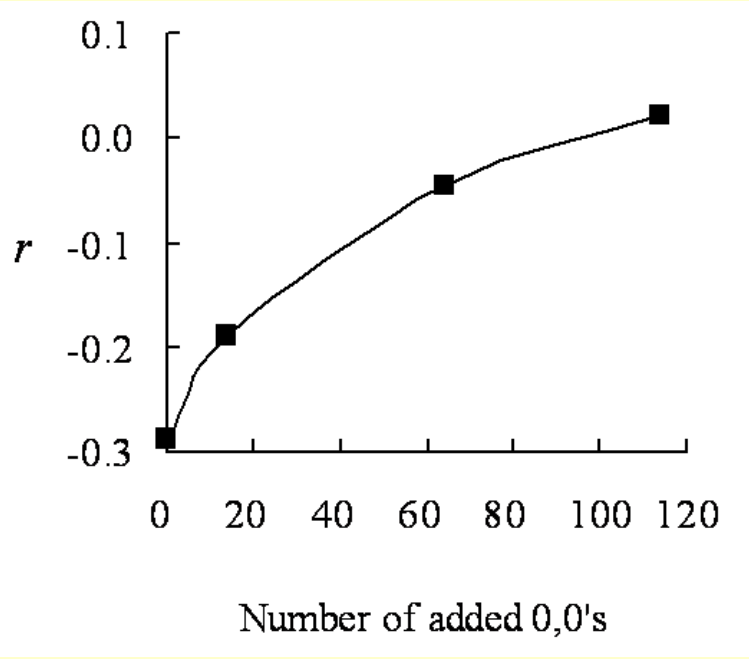
MVS6: What four traits make a good distance measure

- S = 0 if the two samples have no species in common.
- Of course, S = 100 if two samples are identical.

- A scale change in measurements does not change S.
(For example, biomass expressed in g rather than mg)
- "Joint absences" have no effect on S.
(Species not present in either sample, have no influence)

MVS7: What is the Joint Absence Problem?

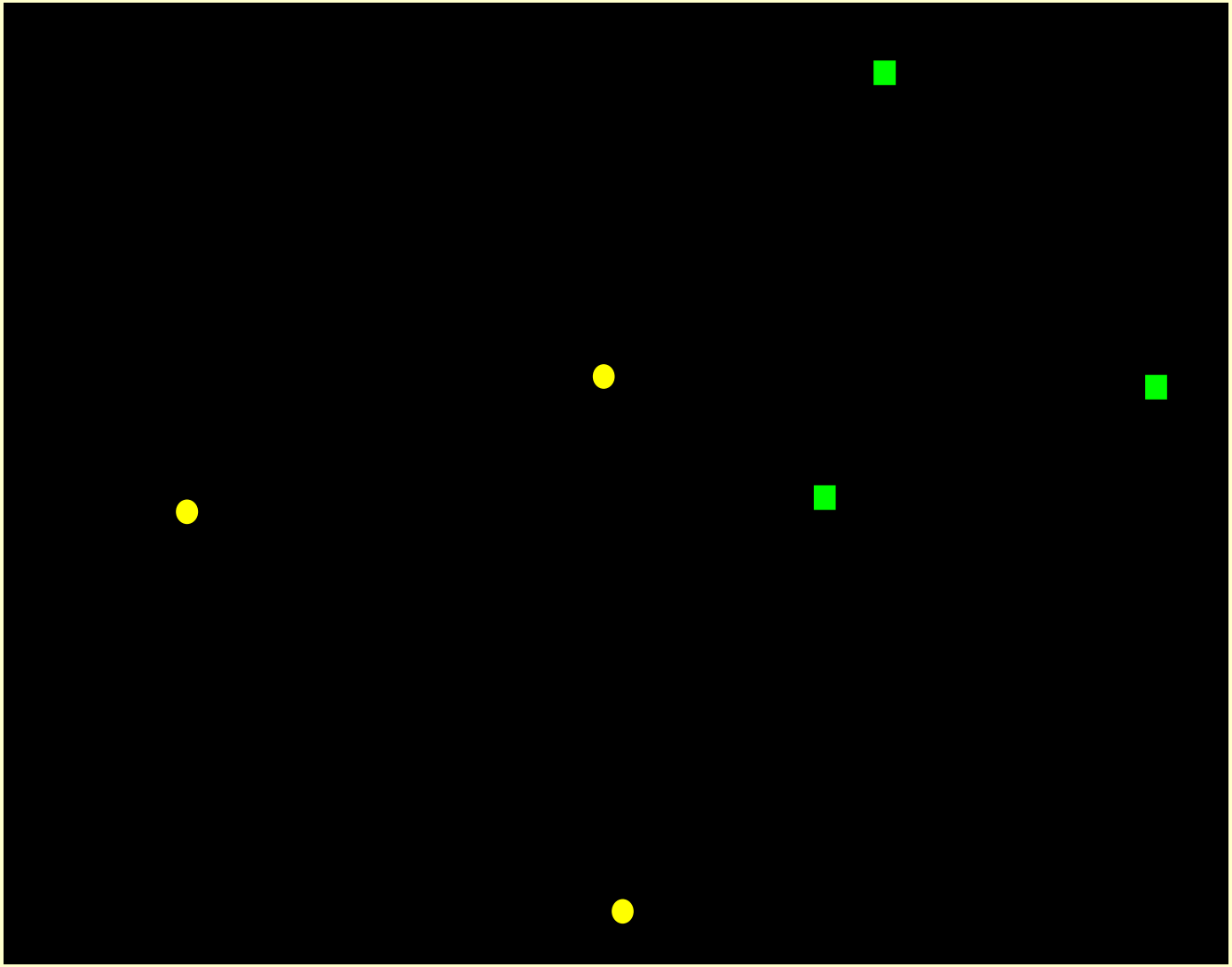




MVS8: What is the Clarke's Rule of Thumb?

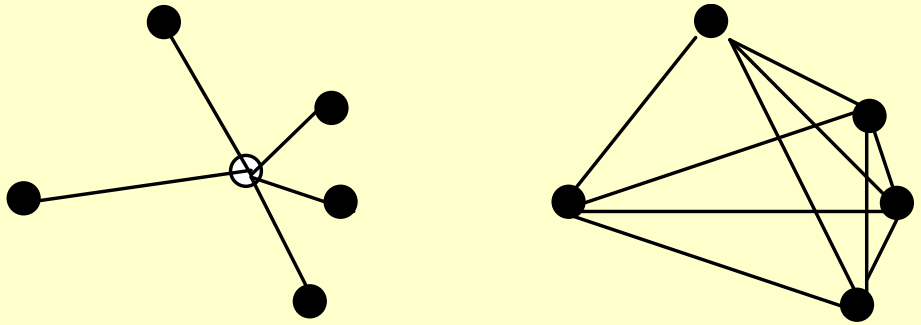
	Clarke's rules of thumb
< 5	An excellent representation with no prospect of misinterpretation. This is, however, rarely achieved.
5-10	A good ordination with no real risk of drawing false inferences
10-20	Can still correspond to a usable picture, although values at the upper end suggest a potential to mislead. Too much reliance should not be placed on the details of the plot.
> 20	Likely to yield a plot that is relatively dangerous to interpret. By the time stress is 35-40 the samples are placed essentially at random, with little relation to the original ranked distances.

MVS9: Explain how PerMANOVA works?



MVS10: What Conceptual Approach facilitated widespread use of MANOVA?

Anderson's (2001) recognition that sums of squares can be calculated directly from distances among data points, rather than distances from data points to the mean.



Sums of distances from points to centroid (left) calculated from average squared interpoint distance (right).

MVS11: How is the Indicator Species Value Calculated?

IndVal method proposed by Dufrêne and Legendre (1997):

IndValGroup *k, Species j* = *100 x A k,j x B k,j*

Where:

A *k,j* = *Specificity (Relative Abundance)*
Bk,j = *Fidelity (Relative Occurrence)*

Note: A species can only indicate one habitat / community

Thus, Individual Value Species *j = max [IndVal k,j]*

MVS12 - How does MRPP Work?

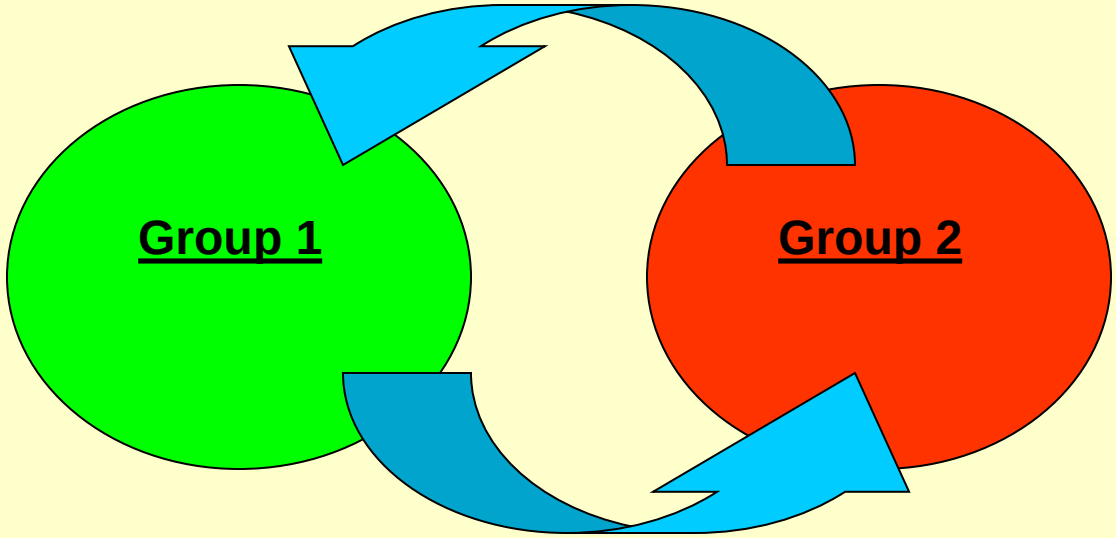
➢ Setting Up:

- Define a Grouping Variable:

  Species Presence / Absence - Main Matrix

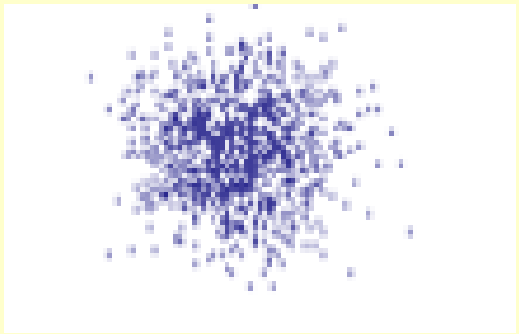
  Environmental Categorical Variable – Second Matrix

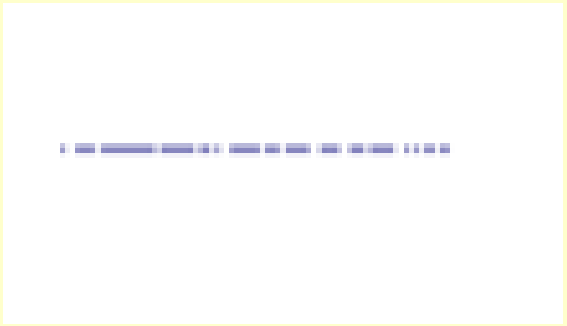
- Select a distance measure (Sorensen / Relative Sorensen) and calculate matrix of distances (D) between all pairs of points within each of the pre-defined groups we are testing



- Shuffle data and recalculate distances, for all possible arrangements of samples into groups

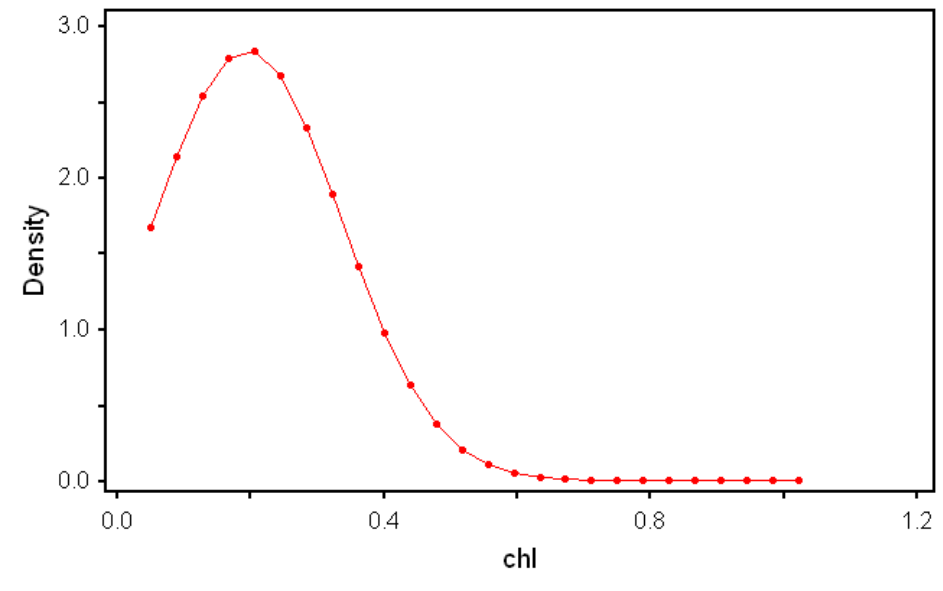
Interpret 1: What is the correlation coefficient?





r = 0

Interpret 2: How would you make these data normal?



y’ = log (y)

Note: no need to add a constant because there are no “zero” data

Interpret 3: Why would you use these data transformations?

ln(zooplankton)	Large counts with no “zero” data
Arcsin(murre production)	One chick per pair, values from 0 to 1
Cassin’s production	More than one chick per pair, values from 0 to 2

Interpret 4: How many Eigenvalues are Meaningful in this example?

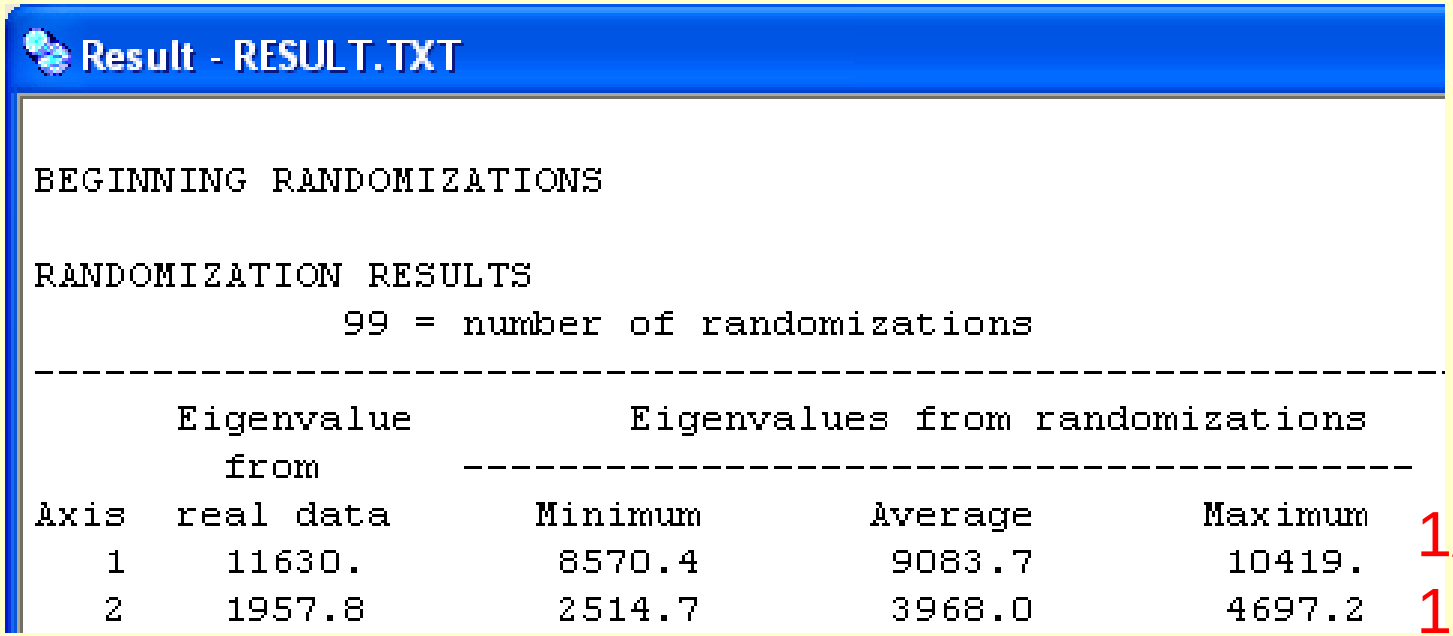
Result - RESULT.TXT

VARIANCE EXTRACTED, FIRST 10 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	11630.331	74.521	74.521	1404.362
2	1957.836	12.545	87.066	1092.227
3	914.362	5.859	92.925	936.159
4	580.305	3.718	96.643	832.114
5	292.397	1.874	98.516	754.080
6	123.200	0.789	99.306	691.653
7	68.209	0.437	99.743	639.630
8	19.814	0.127	99.870	595.039
9	9.192	0.059	99.929	556.022
10	4.004	0.026	99.954	521.341

2

Interpret 5: What is the p value for the first two Eigenvalues?



Result - RESULT.TXT

```
BEGINNING RANDOMIZATIONS

RANDOMIZATION RESULTS
              99 = number of randomizations
-------------------------------------------------------------
      Eigenvalue          Eigenvalues from randomizations
        from       ---------------------------------------
Axis  real data       Minimum        Average        Maximum
   1    11630.         8570.4         9083.7         10419.
   2    1957.8         2514.7         3968.0         4697.2
```

1/100

100/100

Interpret 6: What is the strongest driving variable for each PC axis?

Environmental variable	Eigenvector loading	
	PC1	PC2
Front A	–0.49	–0.29
Front B	–0.55	–0.20
Sea-surface salinity	–0.09	0.58
Chlorophyll maximum	–0.08	0.48
38 kHz backscatter	0.01	–0.52
120 kHz backscatter	0.39	–0.06
200 kHz backscatter	0.43	0.17
420 kHz backscatter	0.41	0.05

Interpret 7: Were these data relativized?

Summary of: 5 vars N = 240 samples

Num.	Name	Mean	Stand.Dev.	Sum	Minimum	Maximum
1	time	-0.1490E-07	1.000	-0.3576E-05	-1.721	1.721
2	MEI	-0.3204E-07	1.000	-0.7689E-05	-3.999	3.251
3	PDO	0.4470E-07	1.000	0.1073E-04	-2.954	2.199
4	upwell36	-0.1490E-07	1.000	-0.3576E-05	-2.756	3.933
5	upwell39	0.3104E-08	1.000	0.7451E-06	-4.906	3.385
AVERAGES:		-0.2806E-08	1.000	-0.6735E-06	-3.267	2.898

NO: maximums different from 1 – no relativization by maximum

NO: sums different from 1 – no general relativization (p = 1)

Interpret 8: How many PCA axes are meaningful? (Use 3 rules)

VARIANCE EXTRACTED, FIRST 5 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	426.649	35.703	35.703	545.717
2	355.865	29.780	65.482	306.717
3	211.665	17.713	83.195	187.217
4	112.108	9.381	92.576	107.550
5	88.713	7.424	100.000	47.800

None : rule 1

BEGINNING RANDOMIZATIONS

RANDOMIZATION RESULTS

999 = number of randomizations

Axis	Eigenvalue from real data	Eigenvalues from randomizations: Minimum	Average	Maximum	p *
1	426.65	258.04	282.69	330.16	0.001000
2	355.87	236.94	257.14	282.83	0.001000
3	211.66	213.22	237.77	260.81	1.000000
4	112.11	197.14	219.96	240.84	1.000000
5	88.713	152.27	197.45	224.34	1.000000

* p-value for an axis is (n+1)/(N+1), where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

Two: rule 2

Two: rule 3

Interpret 9: General Relativization (p = 1) will yield this result:

By columns – generalized: (p = 1):

5	Stands			
5	Species			
	Q	Q	Q	Q
	A	B	C	D
s1	1	10	0.1	100
s2	2	20	0.2	200
s3	3	30	0.3	300
s4	4	40	0.4	400
s5	5	50	0.5	500

5	Stands			
5	Species			
	Q	Q	Q	Q
	A	B	C	D
s1	0.06666667	0.06666667	0.06666667	0.06666667
s2	0.1333333	0.1333333	0.1333333	0.1333333
s3	0.2	0.2	0.2	0.2
s4	0.2666667	0.2666667	0.2666667	0.2666667
s5	0.3333333	0.3333333	0.3333333	0.3333333

Done by rows: totals add up to 1

Interpret 10: How many NMDS axes are meaningful? (Use 2 rules)

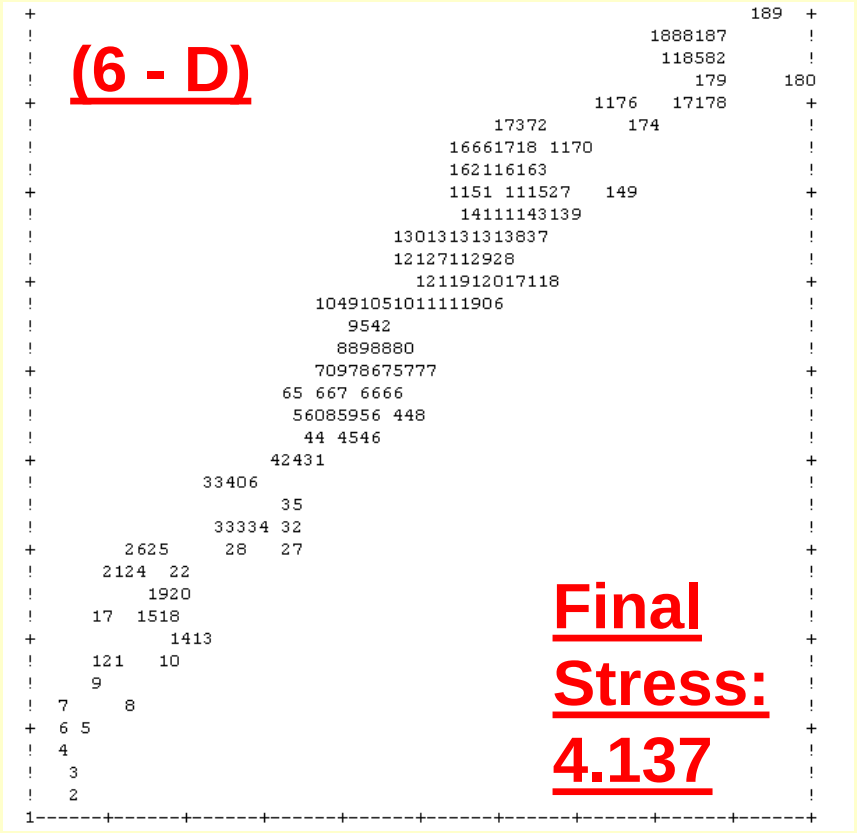
STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)

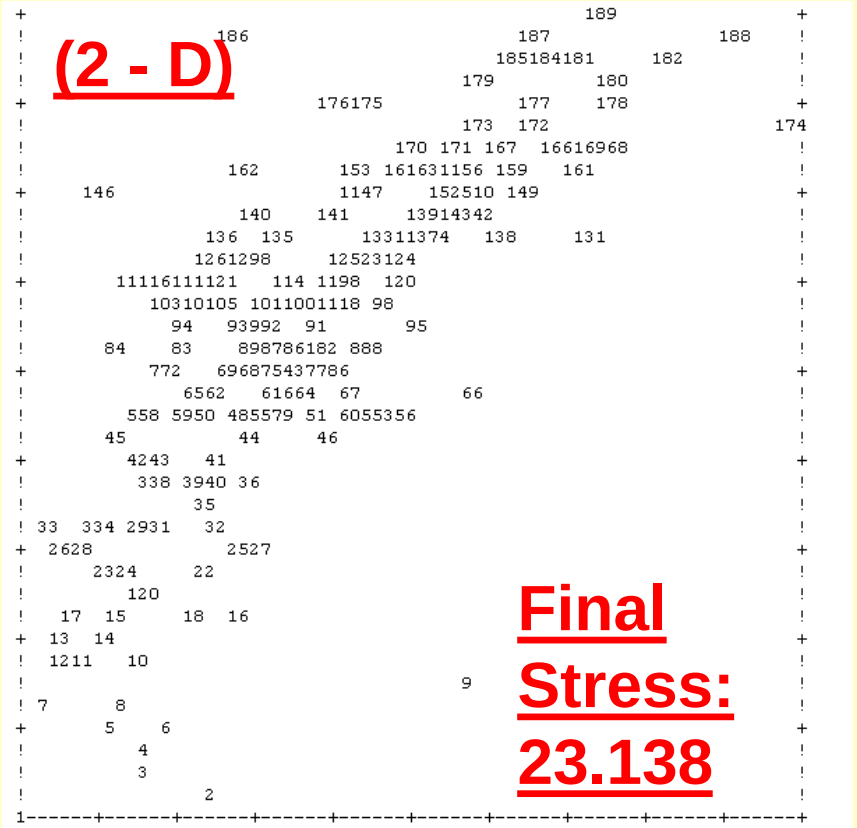
	Stress in real data, 10 run(s)			Stress in randomized data, Monte Carlo test, 20 runs			
Axes	Minimum	Mean	Maximum	Minimum	Mean	Maximum	p
1	38.376	46.541	54.222	41.561	48.626	54.483	0.0476
2	20.366	22.469	25.766	21.752	24.574	28.997	0.0476
3	13.418	13.670	14.855	13.809	15.954	17.877	0.0476
4	8.919	8.954	9.268	8.579	10.807	12.085	0.0952
5	6.078	6.288	6.587	6.662	7.863	9.987	0.0476
6	4.138	4.217	4.499	4.635	5.716	7.708	0.0476

p = proportion of randomized runs with stress < or = observed stress
i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations)

Both rules yield same result: 3 axes

Interpret 11: Which dimension provides the best NMS solution ?





Interpret 12: What is the weighted average of plot B?

	sp 1	sp 2	sp 3	sum
plot A	4	1	0	5
plot B	0	2	4	6
sum	4	3	4	

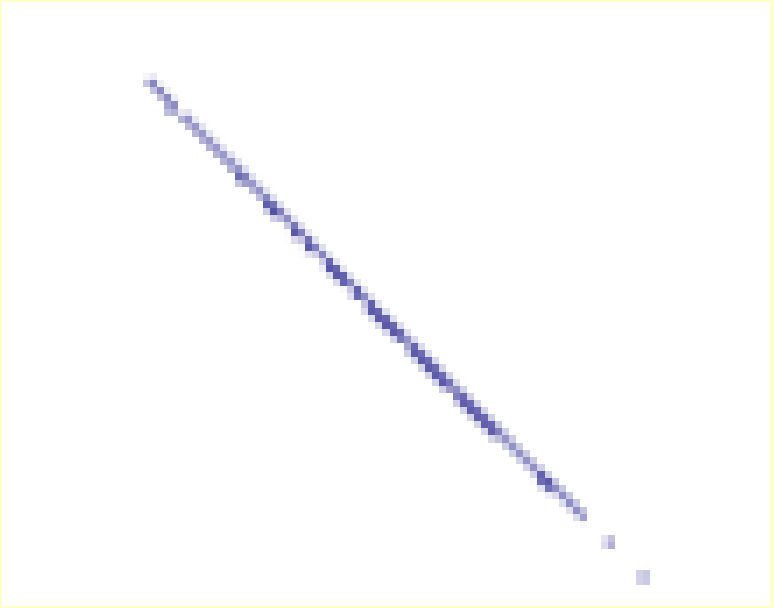
Sp 1	Sp 2	Sp 3
Dry site indicator	Medium site indicator	Wet site indicator
weight = 0	weight = 50	weight = 100

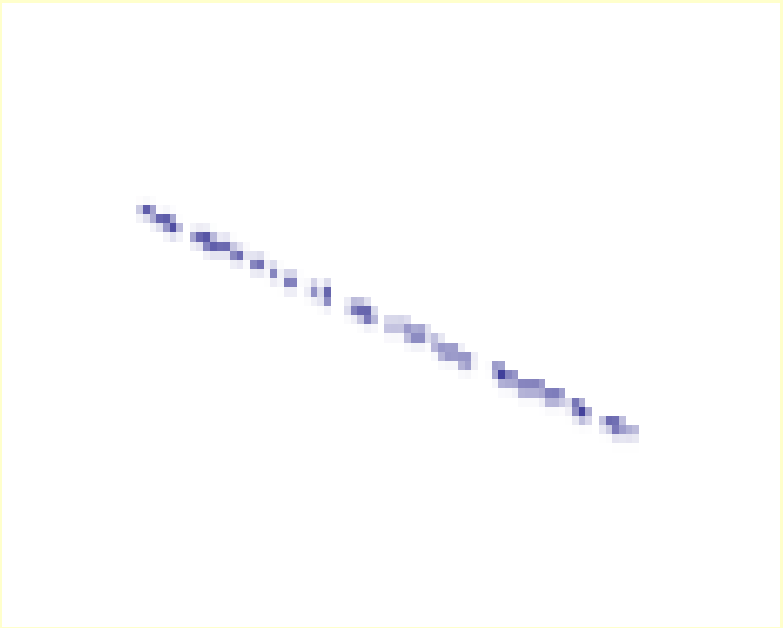
$$v_B = \frac{0(0) + 2(50) + 4(100)}{6} = 500/6 = 83.3$$

True / False 1:

The correlation coefficients for these two scatterplots are the same:

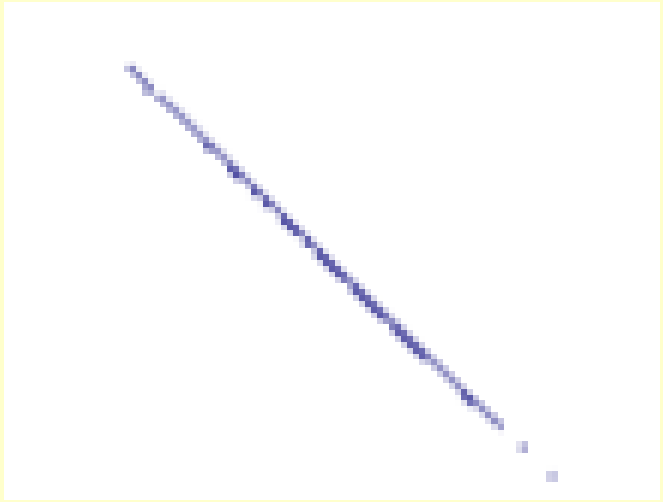
YES: r = -1





True / False 2:

The correlation coefficient and the coefficient of determination for this scatterplot is the same:



NO

r = -1, r_squared = 1

True / False 3:

Both of these relationships are always true:

$r_{12} = r_{21}$ & $r_{12.3} = r_{13.2} = r_{23.1}$

YES

NOT NECESSARILY

True / False 4:

This is a monotonic transformation:

A: Y = (2X + 1) YES

B: Y = sqrt (X) YES

C: Y = (-1) * X NO

D: Y = X ^ 0 NO

True / False 5:

Which one of these is true:

A: Detrended Correspondence Analysis is a disreputed multivariate method

B: Correspondence Analysis is a disreputed multivariate method

C: Canonical Correspondence Analysis is a disreputed multivariate method

D: A & B

True / False 6:

These two statements are true:

4 ^ 0 = 1 YES

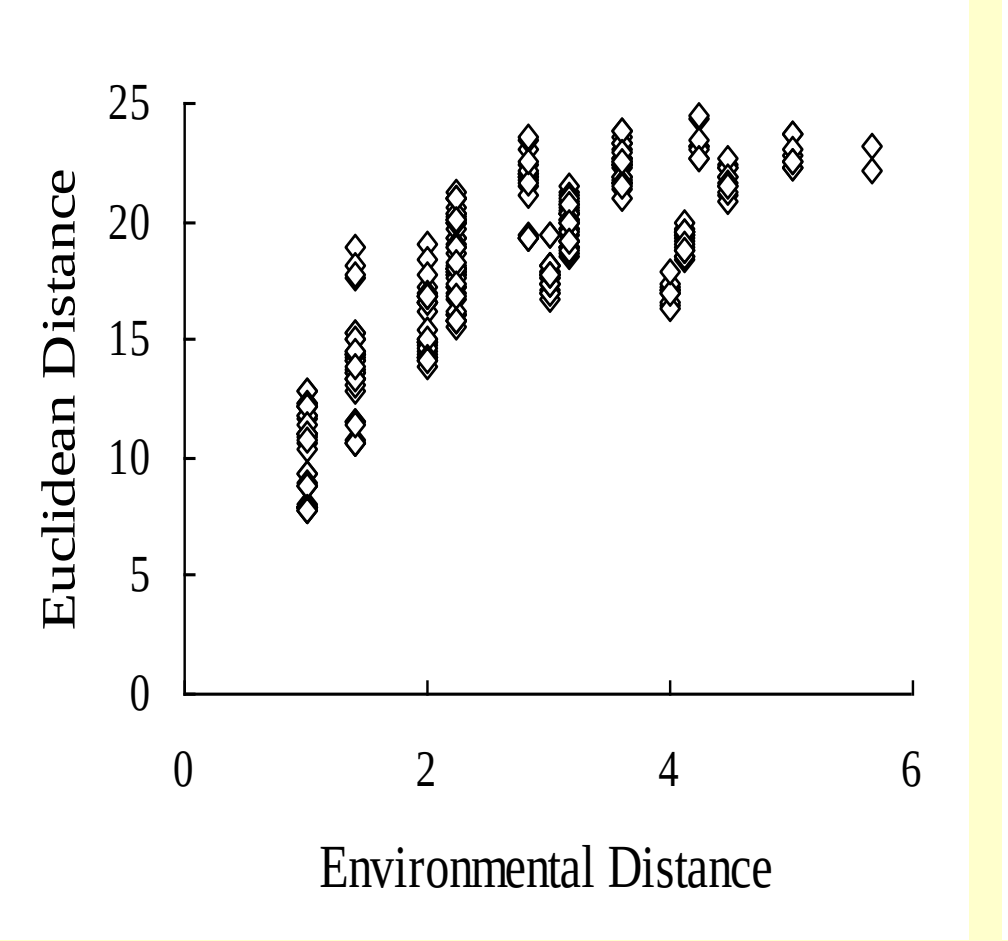
0 ^ 0 = 0 YES

Note:

THIS IS THE FOUNDATION OF
THE “POWER-ZERO”
(PRESENCE / ABSENCE)
TRANSFORMATION

True / False 7:

The Euclidean Distance measure is bounded: NO



True / False 8:

These four criteria define a distance semimetric:

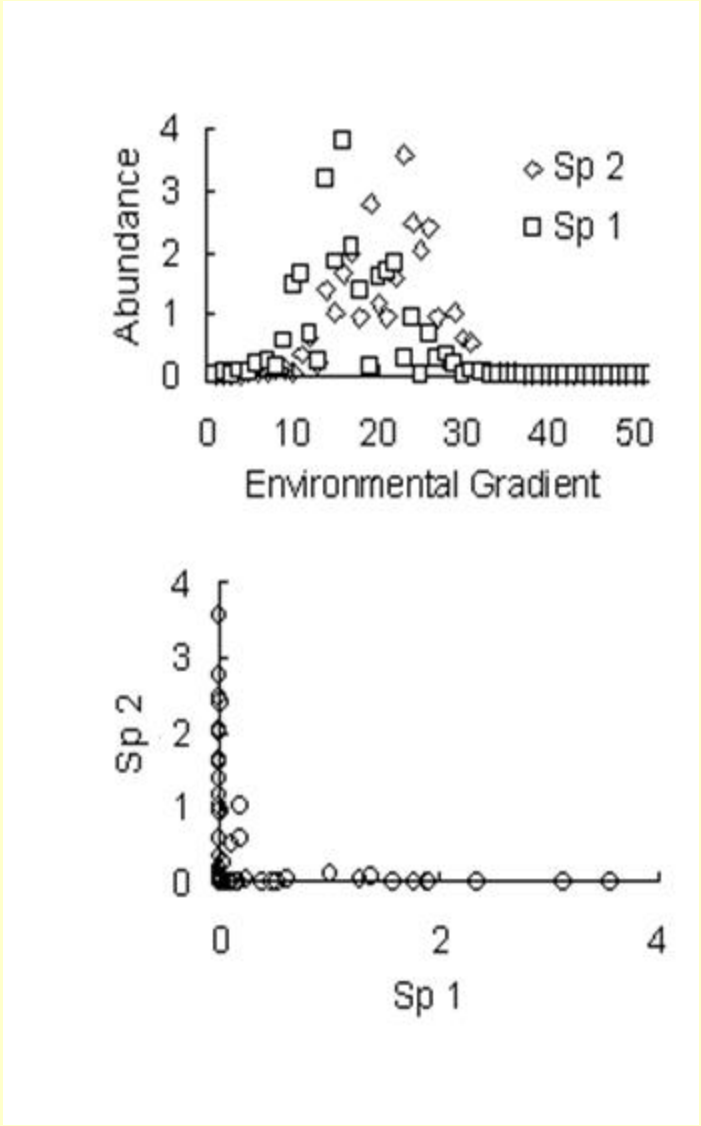
1. The minimum value is zero when two items are identical.
2. When two items differ, the distance is positive (negative distances are not allowed).
3. Symmetry: the distance from objects A to object B is the same as the distance from B to A.
4. Triangle inequality axiom: With three objects, the distance between two of these objects cannot be larger than the sum of the two other distances.

NO, THESE ARE THE CRITERIA DEFINING A METRIC.

A SEMIMETRIC MEETS 1.2 AND 3. (4 NOT NECESSARY)

True / False 9:

These two plots show the same data distributions:



NO.

TOP PLOT SHOWS A “NORMAL CLOUD” SHOWING OVERLAP IN SPECIES DISTRIBUTION

BOTTOM PLOT SHOWS “DUST BUNNY”, WITH NO OVERLAP IN THE TWO SPECIES DISTRIBUTIONS

True / False 10:

This general relativization by standard deviate cannot be performed for the columns in this dataset:

5	Stands				
5	Species				
	Q	Q	Q	Q	Q
	A	B	C	D	E
s1	1	10	0.1	0	1
s2	2	20	0.2	0	1
s3	3	30	0.3	0	1
s4	4	40	0.4	0	1
s5	5	50	0.5	0	1

NO. BECAUSE THE STD OF ROWS D AND E ARE “ZERO”.

STANDARD DEVIATE RELATIVIZATION SUBTRACTS THE MEAN FROM EACH VALUE AND DIVIDES BY THE STD. DIVING BY “ZERO” YIELDS INFINITE.

True / False 11:

General Relativization: (by totals) when p = 2,

ALWAYS makes the area under each species

distribution response curve = 1

NO. GENERAL RELATIVIZATION (BY TOTALS) WHEN P = 1 YIELDS AREAS UNDER EACH SPECIES DISTRIBUTION RESPONSE CURVE = 1

True / False 12:

Which one of these is true:

A: PO is ideal for looking at community structure gradients

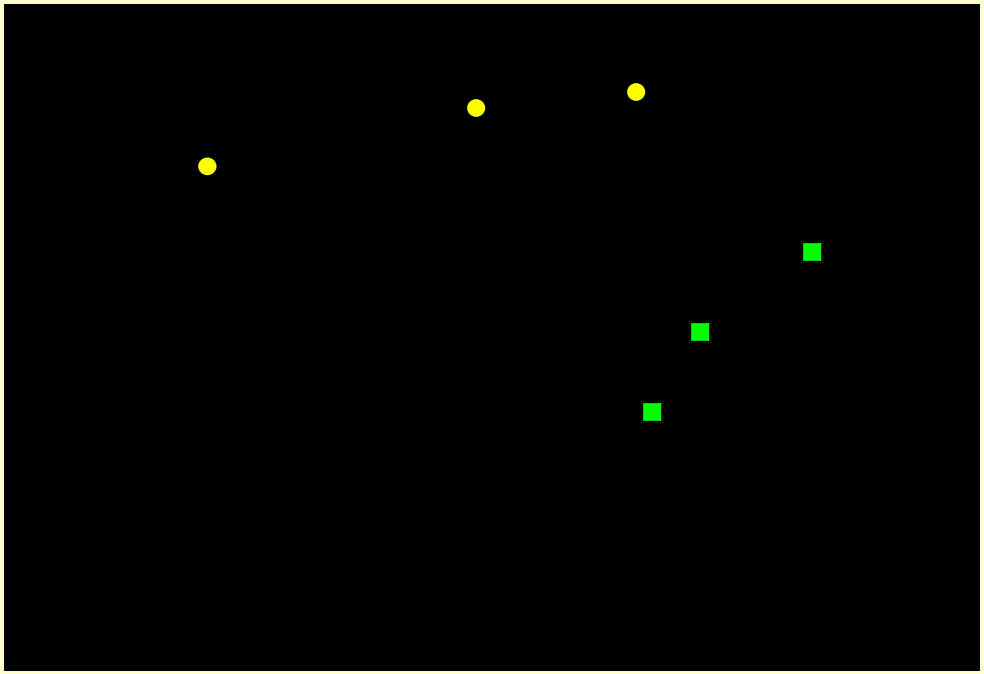
B: CCA is ideal for looking at community structure gradients

C: both A and B

D: none of the above

Tie Breaker:

How many different ways can I distribute these six samples into two groups?



Combinations:
6 samples into 2:

Numerator: 6!

Denominator: 4! * 2!

= (6 * 5) / (2) = 15