Biometry Review

Statistical Foundations for Multivariate Analyses

- ***Objectives:***

Review statistical principles: hypothesis, significance

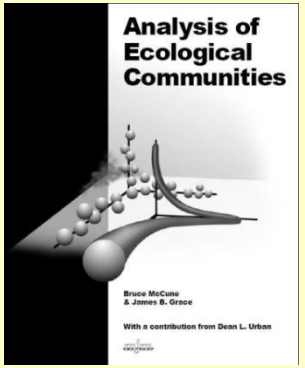
Review key statistics: Mean, Variance, Covariance, Correlation

Discuss Importance of Probability Distributions

Before Taking this Class

CHAPTER 3

Community Sampling and Measurements



What is sampling?

Sampling is the process of selecting objects of study from a larger number of those objects (the population). Each object is then subjected to one or more measurements or observations. Although the word "sampling" is often used in a broad sense to include a discussion of the measurements, the two concepts are distinct.

To be perfectly clear, we will use the word "sample" to refer to a collection of sampling units or sample units (SUs). In casual conversation, a "sample" is often used to mean a single sample unit or a collection of sample units.

What two steps are involved in sampling?

Defining the population

Write down the definition of your population. It is important that the population in the statistical sense be defined in writing during the planning stages. Revise the definition in the field as you encounter unexpected situations. Perhaps the clearest practical way of defining the population is to make a list of criteria used to reject SUs. This list should be reported in the resulting publication.

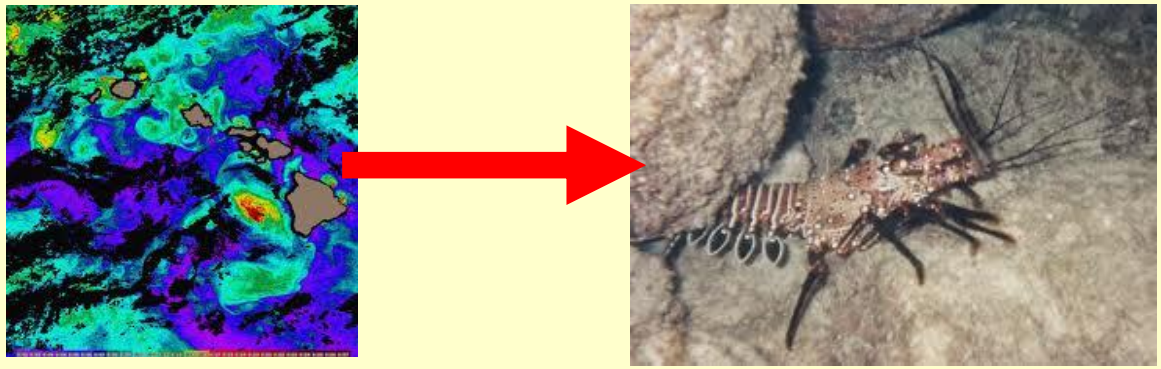
Placement of sample units

An anonymous early ecologist: "The most important decision an ecologist makes is where to stop the car."

What is a Hypothesis?

A testable assertion about how the world works:

e.g., “Spiny Lobster Recruitment driven by eddies”



Formulating Hypotheses

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

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

Working with Hypotheses

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

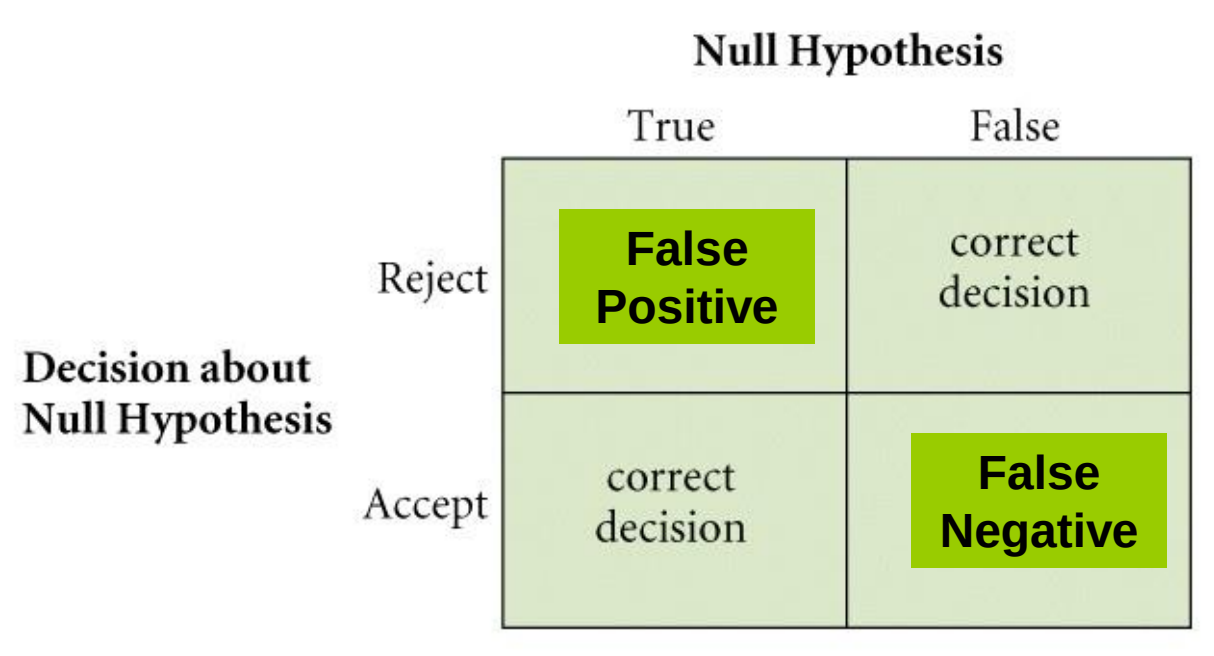
If the Null Hypothesis (Ho) is accepted, it may still not explain the phenomenon. There may be better alternative explanations.

If the Null Hypothesis (Ho) is rejected by a test of the data, then we are left with the Alternate Hypothesis (Ha).

So, the Null Hypothesis **cannot be proved**; **only disproved**

And… **Hypotheses that cannot be refuted are strengthened**

Evaluation of Hypotheses



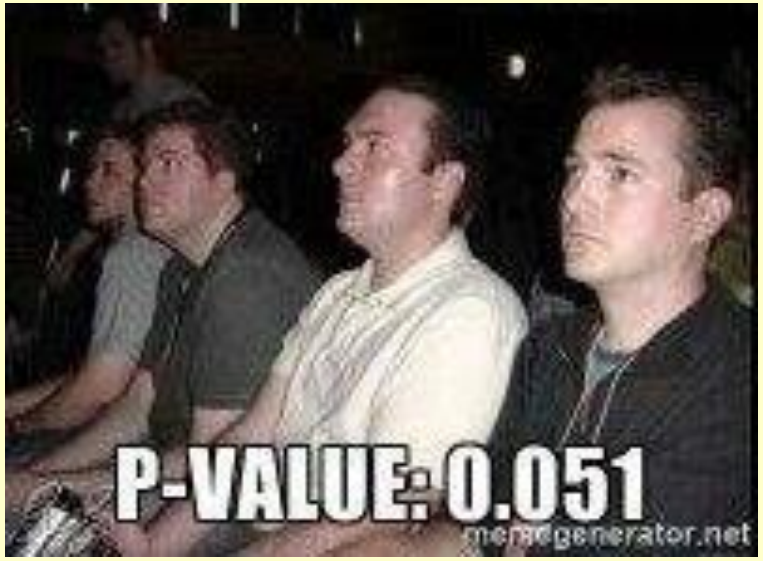
Type I error: rejection of a true null hypothesis

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

Type II error: acceptance of a false null hypothesis

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

What is / is not Statistically Significant ?



P value Definition:

Probability of a test yielding a result equal to or more extreme than the pattern observed in the data, if the null hypothesis is, in fact, true

(thus, the observed pattern caused by random sampling

Statistical Significance

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

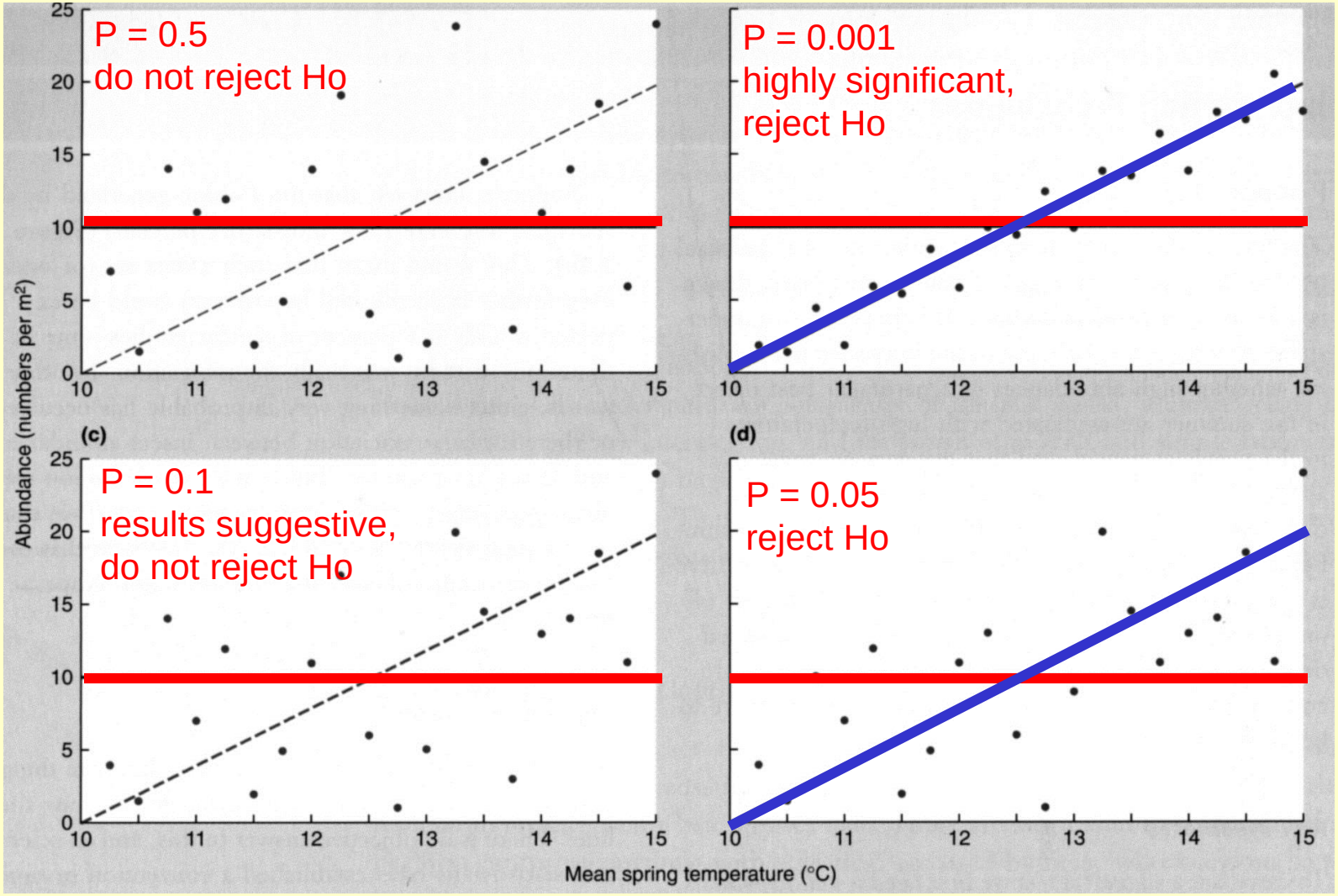
In traditional Fisherian statistical hypothesis testing, the p-value is the probability of observing data at least as extreme as that observed, *given that null hypothesis is true*.

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

- either the null hypothesis is false

  or

- an unusual event has occurred



Note: P-values greater than 0.05 could mean:

- There really is no significant pattern; the null is true
- We have little power to detect pattern (too few or noisy data)

Statistical vs Ecological Significance

df = n -2				
Level of Significance (p) for Two-Tailed Test	**.10**	**.05**	**.02**	**.01**
df				
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708

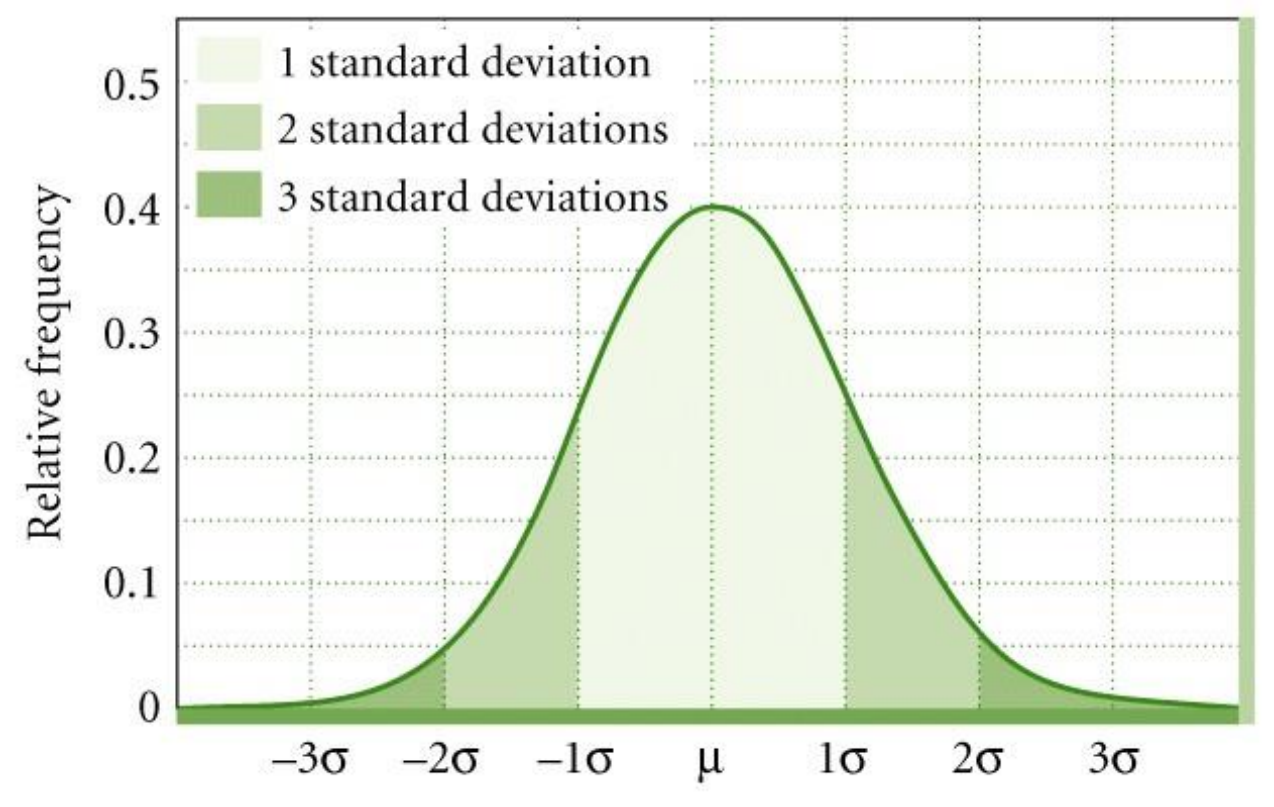
df = n -2				
Level of Significance (p) for Two-Tailed Test	**.10**	**.05**	**.02**	**.01**
df				
25	.323	.381	.445	.487
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.303
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

Easier to Find Significance with a larger sample

Probability Distributions

Frequency of occurrence of “values” in a population

Many variables approximate a **normal distribution**



The average of this distribution is the parametric mean, μ

Shape of distribution determined by the way the observations spread about μ

Probability Distributions

Data Series X: 1,2,3,4,5

Data Series Y: 2,4,6,8,10

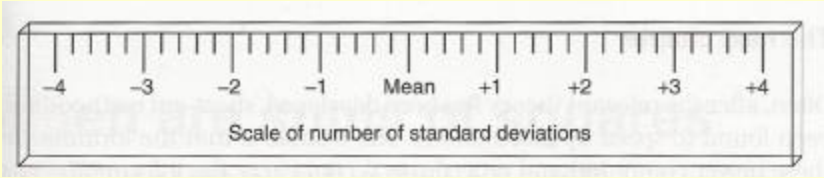
Variance X:	S.D. X:	Variance Y:	S.D. Y:
2.5	1.6	10	3.2

Standard Deviation: Square Root of Variance

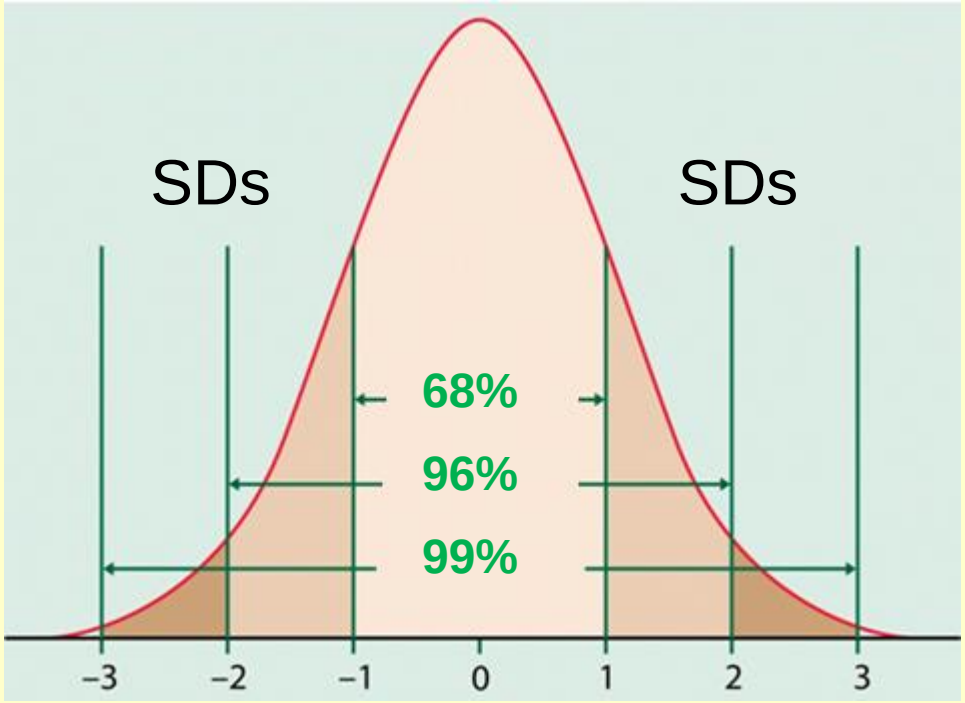
Why use the S.D.?

"Standardized" measure of dispersion about the metric used to describe central tendency (mean +/- SD) (CV)

Allows prediction of the distribution of observations in a measured sample



Parametric Statistics



"Parametric" statistical methods require that variables follow a **normal distribution**

They compare the **means and S.D.s**

In a normal distribution, distributions characterized by:

- ~ 68% within 1 standard deviation of mean
- ~ 96% within 2 standard deviations of mean
- ~ 99% within 3 standard deviations of mean

Non-parametric Statistics

"Non-parametric" statistical methods do not require that variables approximate a **normal distribution**.

They rely on ranked data or compare the **medians** (50% percentile) of the distributions (not the means).

However, every significance test requires information on the actual probability distribution of the statistic.

Critical Value of Statistic

df = n -2				
Level of Significance (p) for Two-Tailed Test	.10	.05	.02	.01
df				
1	.988	.997	.9995	.9999

Observed Statistic < Critical Value

Not Significant Pattern

Observed Statistic > Critical Value

Significant Pattern

Non-parametric Statistics

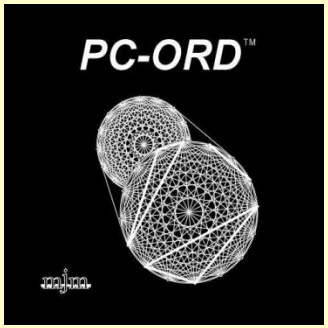
Monte Carlo methods allow the comparison observed statistic against randomized frequency distribution

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

William Sealy Gosset
(1910s)

Ronald Fisher
(1920s)

Maurice Kendall
(1940s)



YOU

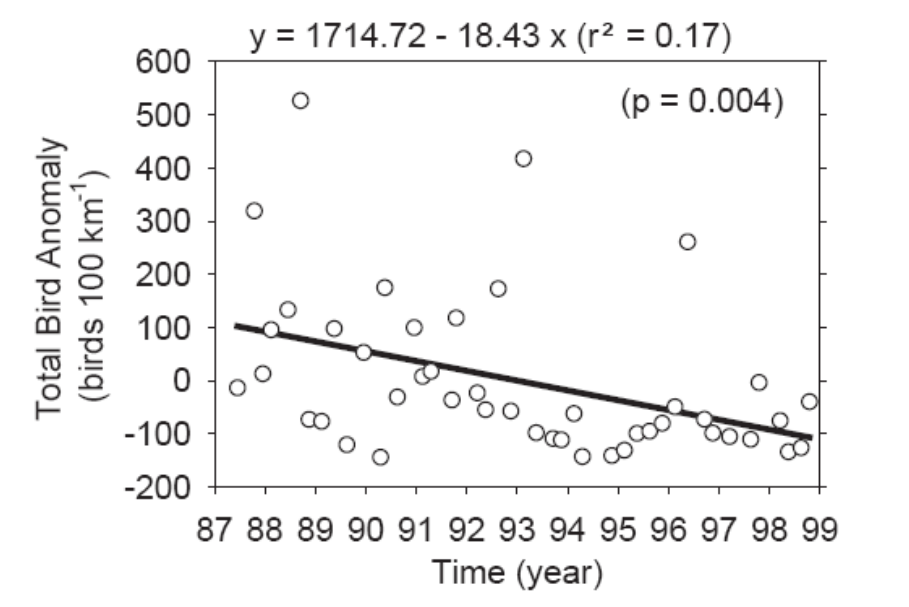
Randomizations

Looking for trends in highly non-normally distributed seabird abundance data (Hyrenbach & Veit 2003).

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.

For instance, if 50 randomizations using the shuffled data yielded a slope with a larger absolute value than the ‘observed’ slope, the p value for that test was 0.05 (50/1000).

Randomizations



Documented a significant decline in bird anomalies

But what is the probability of committing a type I error?

Randomization tests revealed that probability of committing a type I error when performing the regression analyses was: 0.87% (157/18000), 5.50% (991/18000), and 17.22% (3099/18000), at the 0.01, 0.05, and 0.10 probability levels.

Since we performed 18 randomization tests, we expect less than one of these comparisons to yield a significant result merely by chance at the alpha = 0.05 significance level. Nine regressions were significant.

Correlations in PC-ORD

Correlation is a bivariate analysis that measures the strength of association between two variables.

Example of Output

PEARSON AND KENDALL CORRELATIONS WITH ORDINATION AXES N= 54

AXIS:	1			2			3		
	r	r-sq	tau	r	r-sq	tau	r	r-sq	tau
SST	.510	.260	.435	-.106	.011	-.076	.481	.231	.348
SSS	.836	.700	.754	-.199	.040	-.251	-.410	.168	-.286

… etc.

Parametric Correlations in PC-ORD

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

Assumptions: both variable (Y and X) are interval or ratio variables and are well approximated by a normal distribution.

In the Beginning of Statistics

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

The Mean & The Variance

Data Series X: 1,2,3,4,5

Data Series Y: 2,4,6,8,10

Mean X:
3

Mean Y:
6

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Variance X:
2.5

Variance Y:
10

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

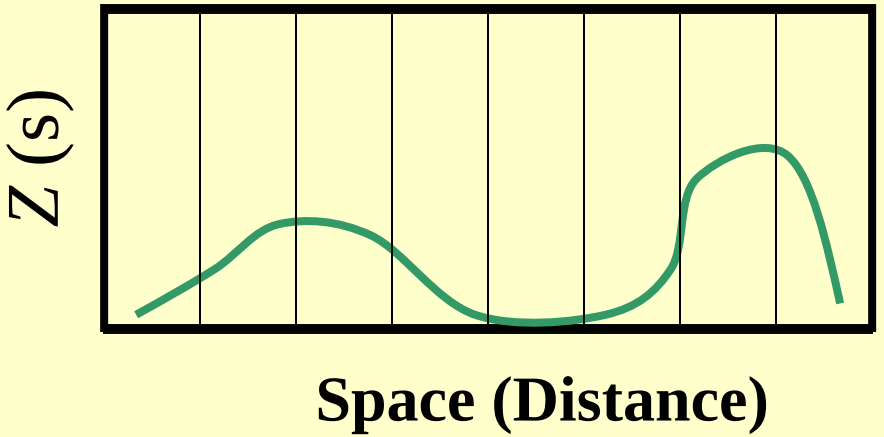
Variance = sum of squared deviations from mean

degrees of freedom

Co-variance = Variance in Two Variables

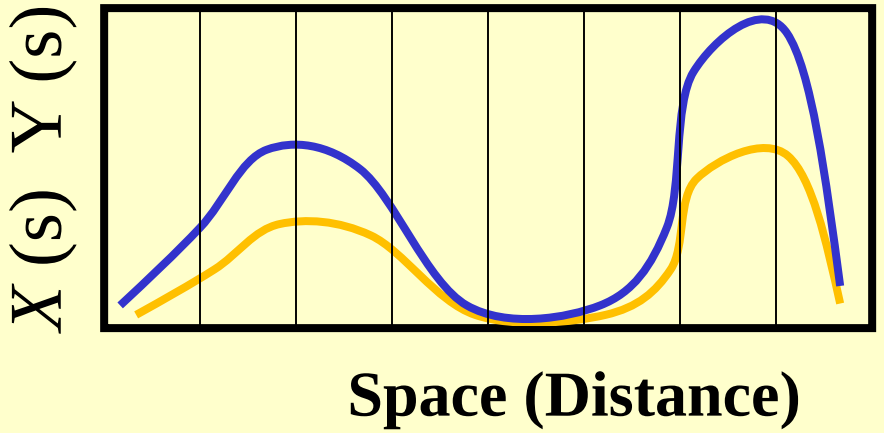
The variance used to assess variability in one variable

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



How about quantifying variability shared by two variables?

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



Covariance in Regression

Covariance is a widely used principle in statistics

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

For Example: **The Regression Coefficient**

Covariance between X and Y divided by the variance in X

Quantifies best-fit slope of line expressing relationship of X and Y

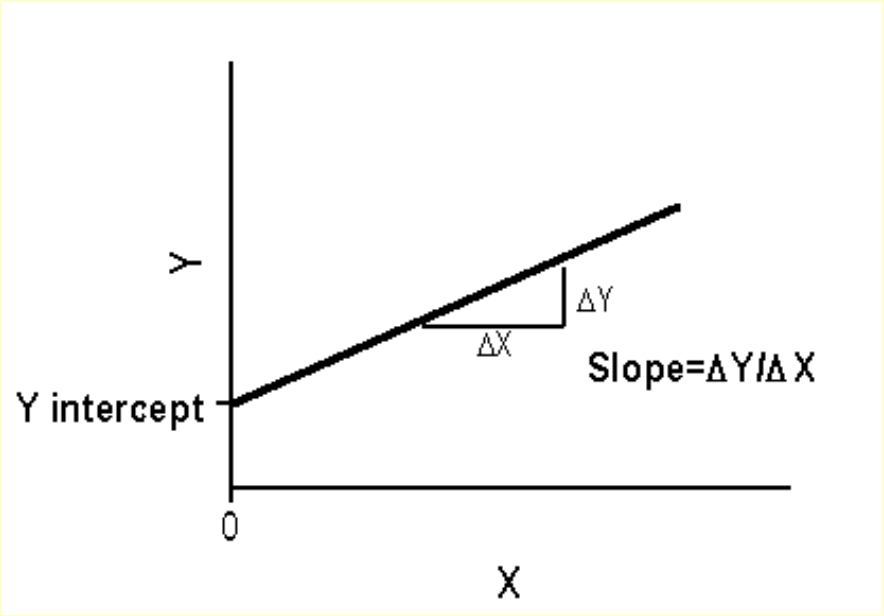
$$b = \frac{Covariance = \frac{\sum (Xi - \overline{X})(Yi - \overline{Y})}{(n-1)}}{Variance = \frac{\sum (Xi - \overline{X})(Xi - \overline{X})}{(n-1)}}$$

Mechanics of Linear Regression

Cancel out the term (n-1) from numerator / denominator:

$$b = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sum (Xi - \bar{X})(Xi - \bar{X})}$$

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



$$Y = a + bX + e$$

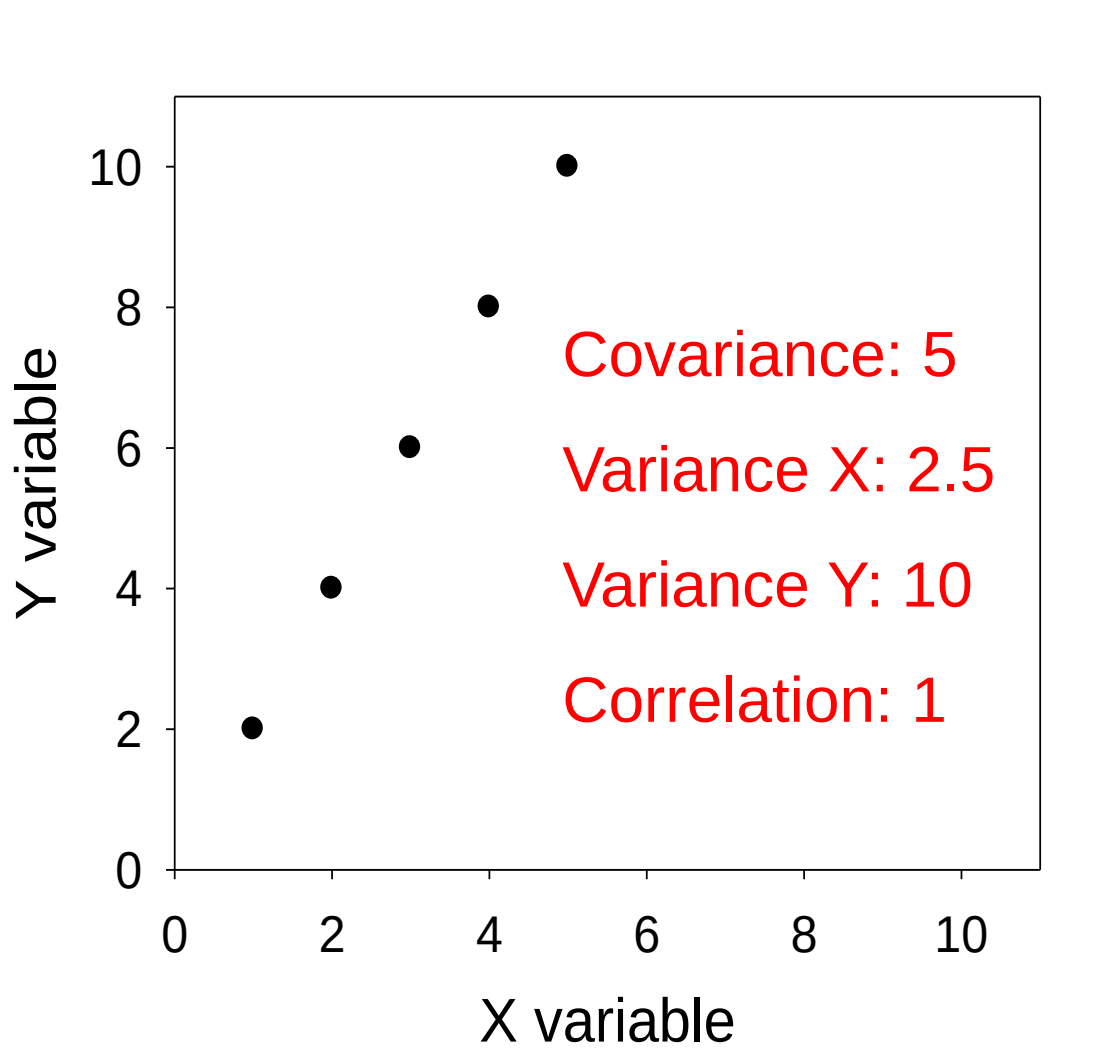
Covariance in Correlation

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

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

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

Mechanics of Pearson Correlation



X	Y
1	2
2	4
3	6
4	8
5	10

Pearson correlation:

5 / [sqrt(2.5)*sqrt(10)]

$r = +1$

$r^2 = 1$

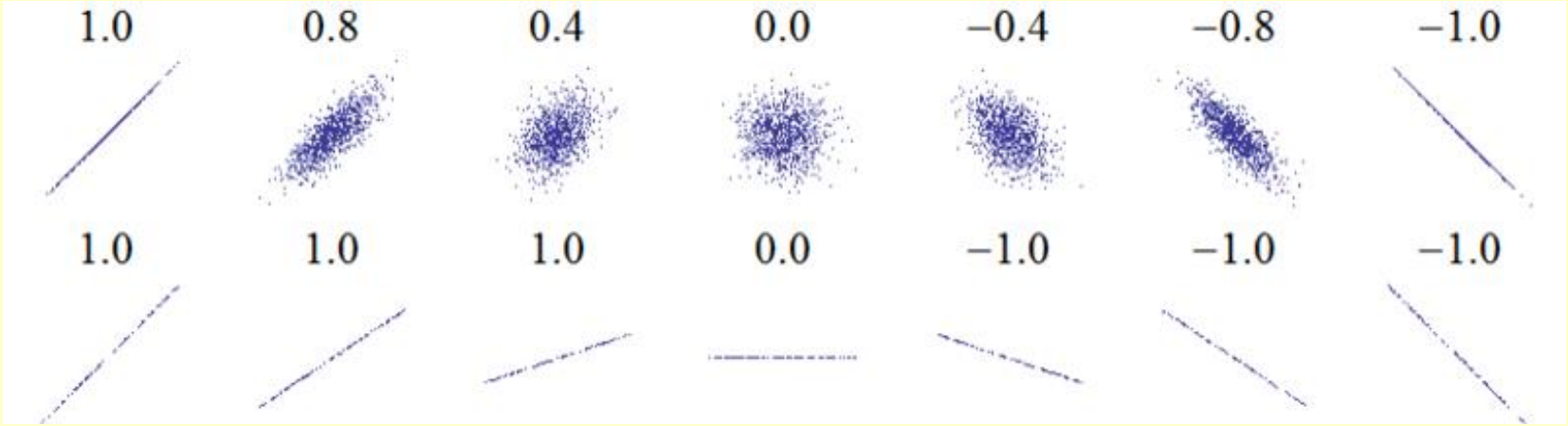
Interpretation of Pearson Correlation

The correlation coefficient (r) indicates the strength and the direction of a linear relationship between two random variables

$-1 \leq r \leq 1$

The coefficient of determination (r^2) indicates the % of the variance in one variable that is explained by the other (bidirectional)

$0 \leq r^2 \leq 1$



Correlations in PC-ORD

Correlation is a bivariate analysis that measures the strength of association between two variables.

Example of Output

PEARSON AND KENDALL CORRELATIONS WITH ORDINATION AXES N= 54

AXIS:	1			2			3		
	r	r-sq	tau	r	r-sq	tau	r	r-sq	tau
SST	.510	.260	.435	-.106	.011	-.076	.481	.231	.348
SSS	.836	.700	.754	-.199	.040	-.251	-.410	.168	-.286

… etc.

Non-Parametric Correlations in PC-ORD

Kendall rank correlation: A non-parametric test that does not make any assumptions about the distributions - unlike the Pearson's correlation.

Assumptions: both variable (Y and X) are either interval, ratio or categorical variables; no assumptions concerning the data distributions (this is a distribution-free statistical test).

Kendall Correlation

Kendall rank index, Tau:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Where:

- concordant pairs have the same relative rankings
- discordant pairs have different relative rankings
- n is the number of observations
- and the total number of pairs compared = **½ n (n-1)**

Let's say we have three pairs of X and Y data (n = 3)

How many pair-to-pair comparisons can we make ?

Mechanics of Kendall Correlation

Data pair	X	Y	Rank X	Rank Y
A	3	6	3	3
B	2	4	2	2
C	1	2	1	1

Xa, Ya vs Xb, Yb: Xa > Xb & Ya > Yb Concordant Pairs

Xb, Yb vs Xc, Yc: Xc < Xb & Yc < Yb Concordant Pairs

Xa, Ya vs Xc, Yc: Xa > Xc & Ya > Yc Concordant Pairs

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

Tau = (3 – 0) / 3 = 1

What happens with tied pairs?

Interpretation of Kendall Correlation

Coefficient must be in the range $-1 \leq Tau \leq 1$

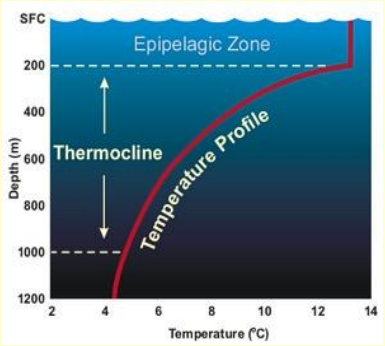
- If agreement between the two rankings is perfect (i.e., two rankings equal) Tau = +1
- If disagreement between two rankings is perfect (i.e., one ranking is the reverse of the other) Tau = −1
- If *X* and *Y* are independent, then expect the Tau = 0

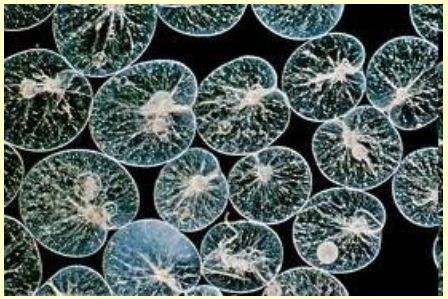
Important Differences with Pearson Correlation:

Relationship between variables NOT linear

Only one coefficient: sign and strength of association (NOTE: never use Tau squared)

Direct / Indirect Correlations





THERM → DINOS → FISH

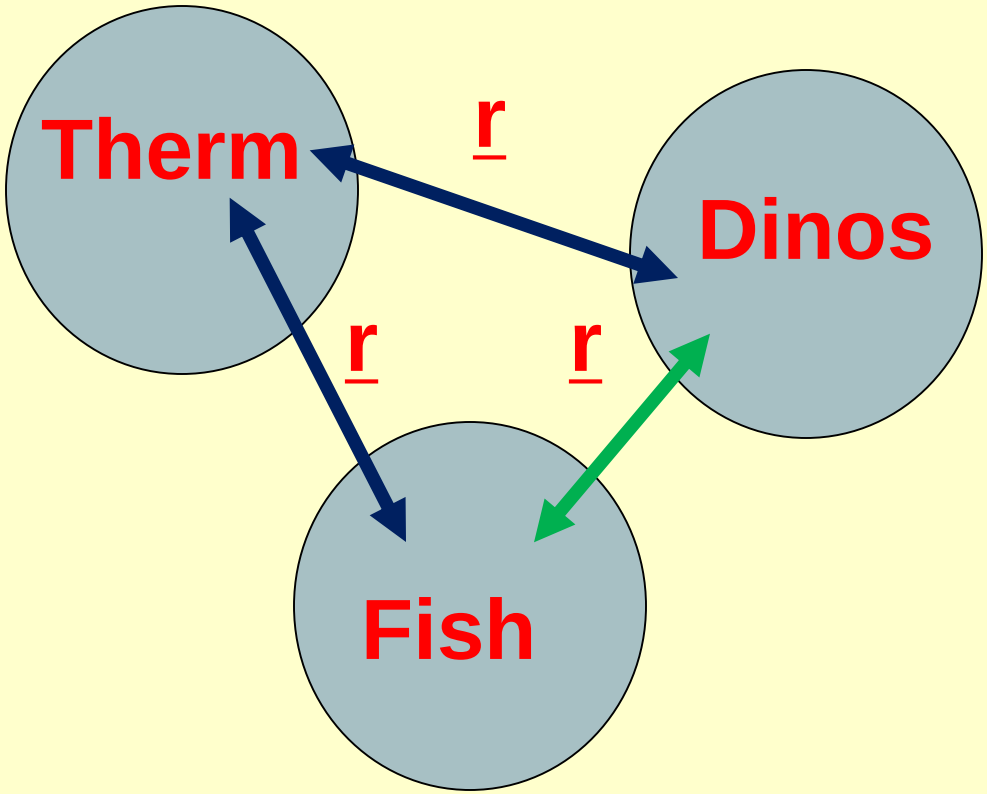
THERM → FISH

The '***stable ocean hypothesis***' (Lasker 1975) established that anchovy larvae survive and grow when periods of calm weather maintain aggregations of motile phytoplankton for long enough for larvae to develop through week-long '*first feeding*' life stage.

But mixing can also nutrient input (ocean productivity)

Approach 1: Partial Correlation

Want to quantify effect of Dinoflagellate concentration on Fish larvae abundance, while considering Thermocline influences



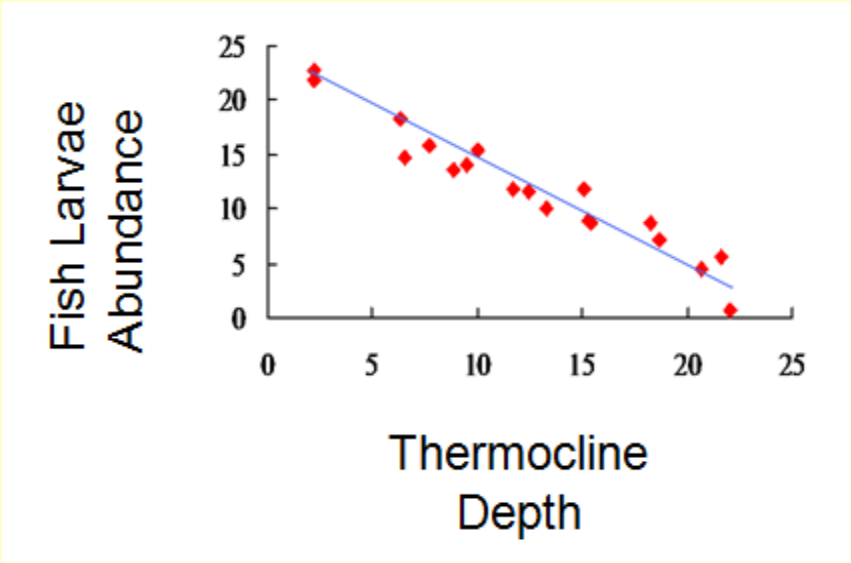
NOTE:

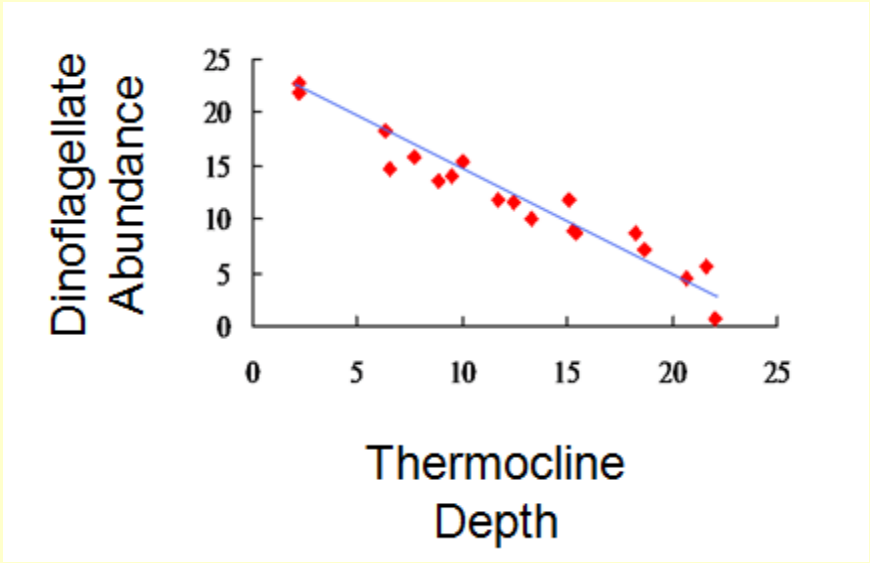
Anticipate direct Influences of thermocline depth on dinoflagellate and larval fish abundance

Approach 1:

r (dino,fish.therm)

Approach 2: Partial Regression





Method 2:

Regress Fish on Therm

Regress Dinos on Therm

Regress Fish Residuals on Dino Residuals