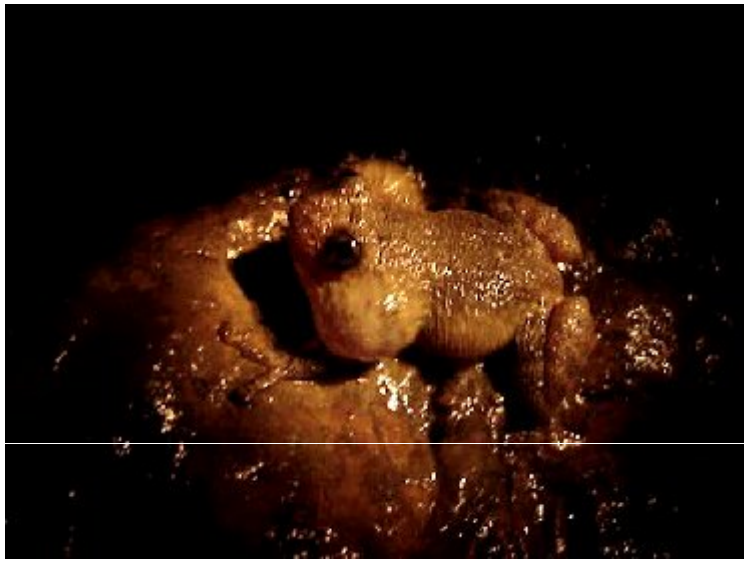
Ecological Statistics: An oxymoron to understand complex world!

Gururaja KV
IISc, Bangalore
gururajakv@gmail.com

Starting with frog movies…



Few more…

Recall (short term memory…loss)

- Systematic stratified random sampling
- Night survey, from 2003 – 2006, seasonal, search for all
- Identify and record species, numbers etc.
- Secondary data on Vegetation studies, RS and GIS
- Opportunistic observations also included for overall diversity in the region
- Shannon's index ($H' = - \Sigma p_i \ln p_i$),
- Simpson's index ($D = 1 / \Sigma p_i^2$)

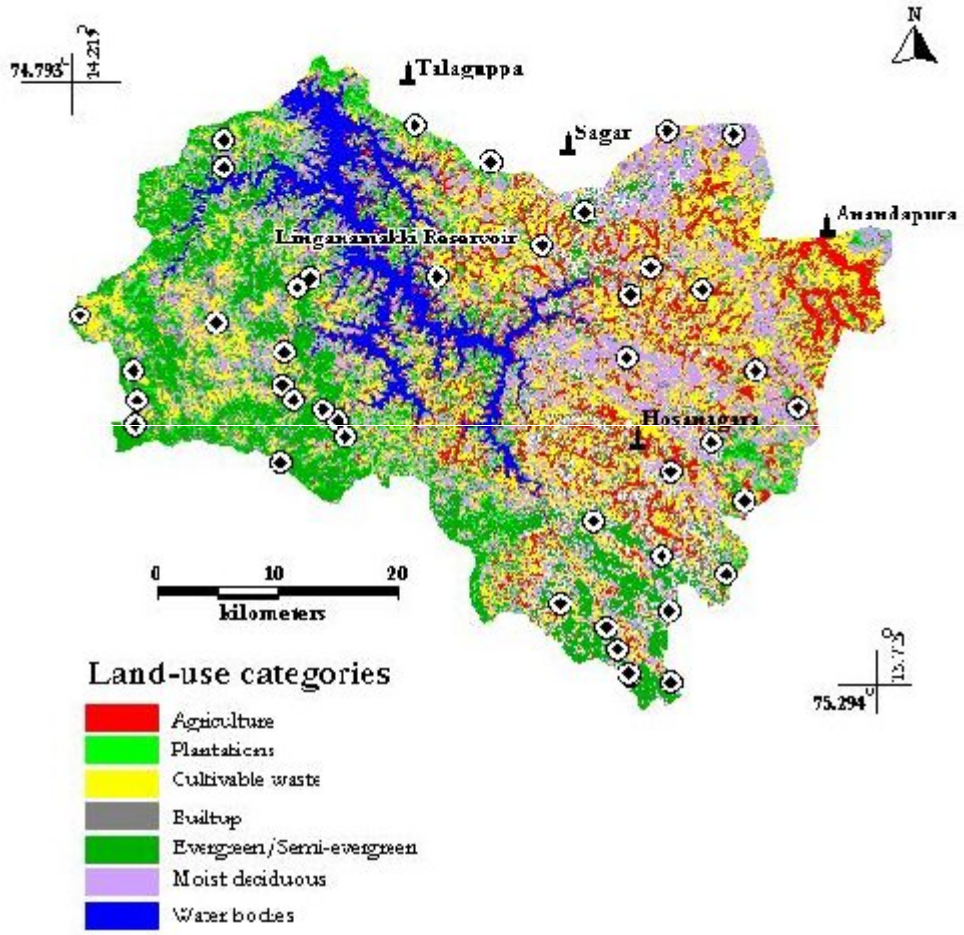
What data I got?

- Varieties of species (Species richness)
- Abundance of species (#indviduals/man hour search)
- Derived data (Endemics, Endemic abundance)
- Secondary source (IUCN red list)
- Percent Evergreens, rainfall, percent landuses, forest fragmentation data,

Sampling localities

Sampling sites	Longitude (°E)	Latitude(°N)	Altitude (m)
1	75.0896	13.773	584
2	75.0804	13.8532	586
3	74.8839	13.9269	580
4	74.7268	13.965	563
5	75.1055	13.9735	602
6	74.8428	13.9786	598
7	75.1084	14.0209	559
8	75.1245	14.0418	557
9	75.073	14.0831	680
10	75.04	14.058	596
11	75.109	13.7379	696
12…			

On the map Or image



Species list (checklist!)

Euphlyctis cyanophlyctis				11		1		1
Euphlyctis hexadactylus								1
Hoplobatrachus tigerinus	1			1		2	1	1
Indirana beddomii				1				
Indirana semipalmatus		1		5		1		2
Fejervarya granosa				1				
Fejervarya caperata	4	1	5	2		3	3	4
Minervarya sahyadris		1						
Micrixalus fuscus								6
Micrixalus saxicola				2				
Nyctibatrachus aliciae				2				3
Nyctibatrachus major								15
Philautus sp1				1	2			
Philautus amboli	2	6	37	6	4	3	7	8
Philautus sp2			2	3	1	1		1
Philautus tuberohumerus		3	2	2	1			
Polypedates maculatus			1					

List is incomplete…species wise..as well as site wise

Other data

Sub-basin	OA	ESE	MD	P	A	OF	RF	ET	E	IF	PF	EF
Nandiholé	11.14	3.31	38.12	5.52	11.33	30.56	1650.30	27.00	43.00	9.58	19.37	7.15
Haridravathiholé	13.86	2.28	28.18	5.58	18.19	31.92	1409.00	9.00	16.00	5.02	13.15	5.36
Mavinaholé	11.60	4.37	41.62	7.88	9.81	24.68	2950.40	12.10	14.60	9.96	21.26	7.83
Sharavathiholé	12.49	19.16	22.95	14.66	10.32	20.42	5041.30	48.45	84.48	8.09	18.96	6.27
Hilkunjiholé	5.57	43.27	22.46	11.56	4.18	12.96	5041.30	43.20	83.60	21.02	32.98	8.52
Hurliholé	7.56	32.78	27.91	10.74	1.97	18.17	4073.40	50.00	89.90	16.47	30.05	9.74
Nagodiholé	7.46	52.14	16.58	13.66	1.08	9.07	5689.00	86.70	43.10	24.91	28.98	10.38
Yenneholé	10.10	37.89	19.76	15.88	1.37	14.86	6531.00	39.40	80.40	14.69	25.20	10.12

And more…

So what to do now?

- Back to the questions asked
- Statistical application…find out the relationship (among variables and between amphibian diversity and these variables)

Welcome to Ecological Statistics!

- Ecology deals with "The study of spatial and temporal patterns of distribution and abundance of organisms, including causes and consequences" (Scheiner and Willig, 2007)

  - A statistic is a estimate of the value of a parameter.
  - A set of procedures and rules for reducing large masses of data into manageable proportions allowing us to draw conclusions from those data

  - Limitations of statistics
    - Snap shot, data

  - Unlimited Ecology…
    - Dynamics…feedback…complex…non linear

Ecological statistics deals with…

- Part I
  - Framing a good question?
  - Setting the objectives?
  - Sampling design, Replication, Randomnization
- Part II
  - Parameteric vs Non parametric
  - Goodness of fit
  - Measuring central tendency
  - Hypothesis testing
  - Measuring relationships
- Part III
  - Measuring biodiversity
  - Ordination, Multivariate analysis
  - Spatial analysis

Basic terms …

- Measurement – assignment of a number to something
- Data – information on species, variable, parameter
- Sample – Small portion of a population
- Population – all possible units of a given area
- Variable – Varies with the influence of other.
- Parameter – A measurable unit (eg. Temperature, pH)

Types of Measurements

- Ordinal – rank order, (1st,2nd,3rd,etc.)
- Nominal – categorized or labeled data (red, green, blue, male, female)
- Ratio (Interval) – indicates order as well as magnitude

Bob Cosatner is greatly acknowledged for sharing his slides

Types of Variables

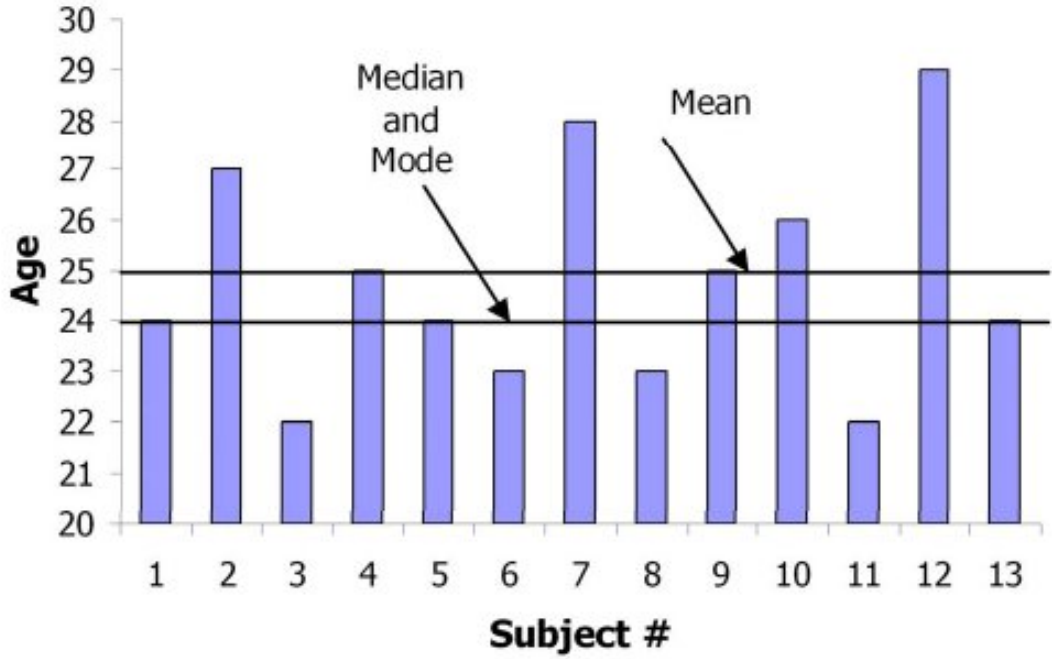
- Independent Variable – controlled or manipulated by the researcher; causes a change in the dependent variable, generally represented on x-axis
- Dependent Variable – the variable being measured, represented on y-axis
- Discreet Variable – has a fixed value
- Continuous Variable – can assume any value

Measure of Central tendency

- Mean (average)
- Median (middle)
- Mode (most frequent)

How to measure central tendency?

Subject #	Age		
1	24		**mean** - add up all ages and divide by the toal
2	27		Excel command is =average(b2:b25)
3	22		(24+27+22+25+24+23+28+23+25+26+22+29−24) / 13
4	25		25
5	25		
6	24		**median** - halfway point, equal number of variables on both sides
7	23		Excel command is =median(b2:b23)
8	28		22,22,23,23,24,24,24,25,25,26,27,28,29
9	23		24
10	25		
11	26		**mode** - most frequent
12	22		Excel command is =mode(b2:b23)
13	29		22,22,23,23,24,24,24,25,25,26,27,28,29
14	24		24



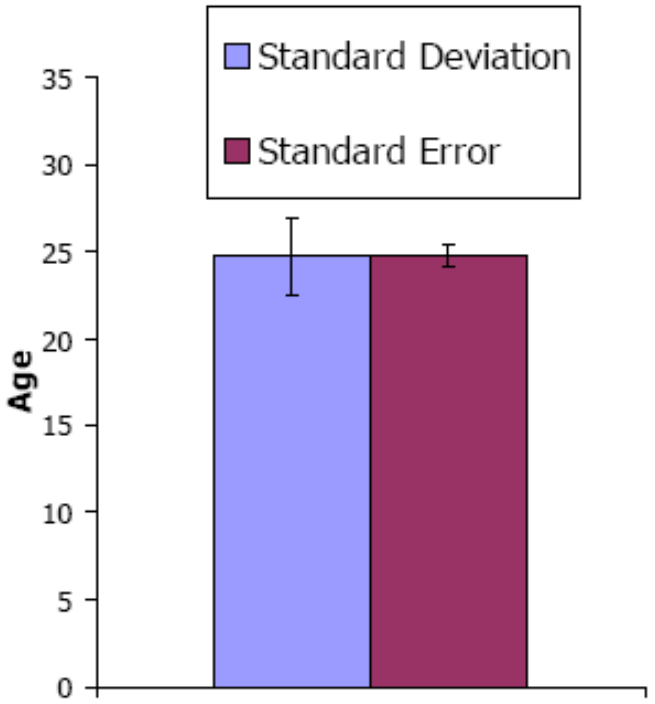
Measures of Dispersion

- Variance
- standard deviation
- standard error

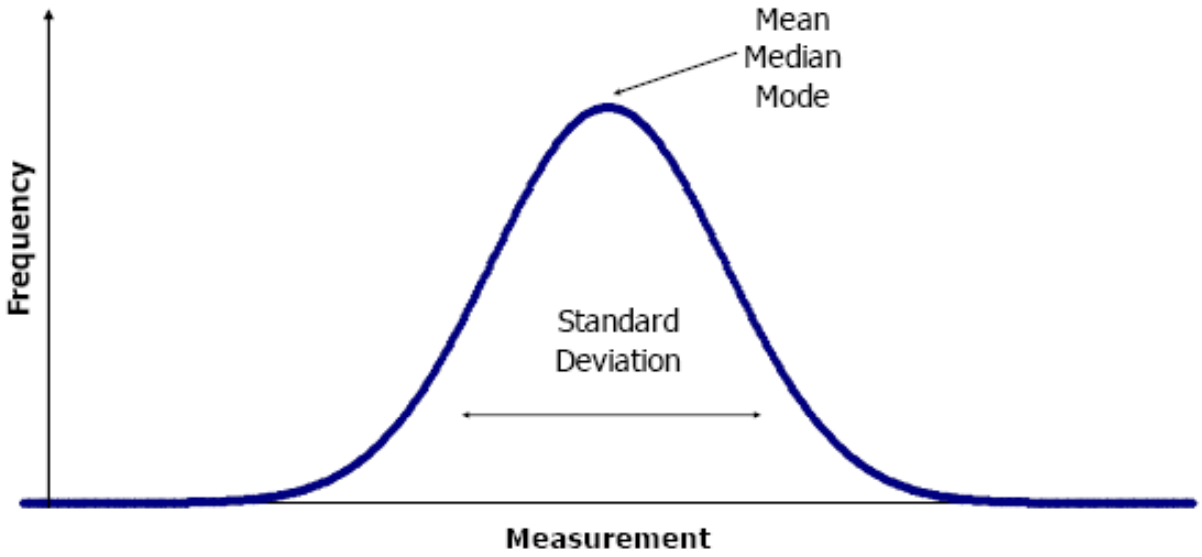
Measure of dispersion

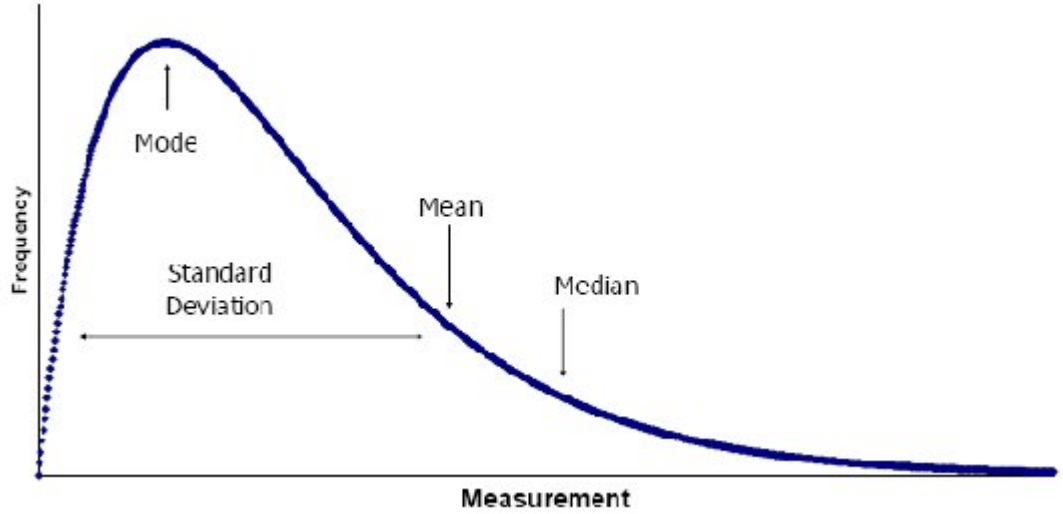
Subject #	Age	
1	24	**mean** - add up all ages and divide by the total
2	27	Excel Command is =average(b2:b14)
3	22	(24+27+22+25+24+23+28+23+25+26+22+29+24)/13 =
4	25	25
5	24	
6	23	**Standard Deviation** - square root of the sum of the squared
7	28	individual differences with the mean divided by
8	23	the total number of data points minus 1.
9	25	S.D. = $\sqrt{[\Sigma(yi - ymean)2/(N - 1)]}$
10	26	Excel command is = stdev(b2:b14)
11	22	2.2
12	29	
13	24	**Standard Error** - Represents the spread in means if many
14		samples of the same size are taken from the population.
15		S.E. = S.D. / $\sqrt{N}$

Sample variance is summation of square of deviance from each data point divided by n-1, In excel it is given by VAR(data)



Why we ask for random sample?



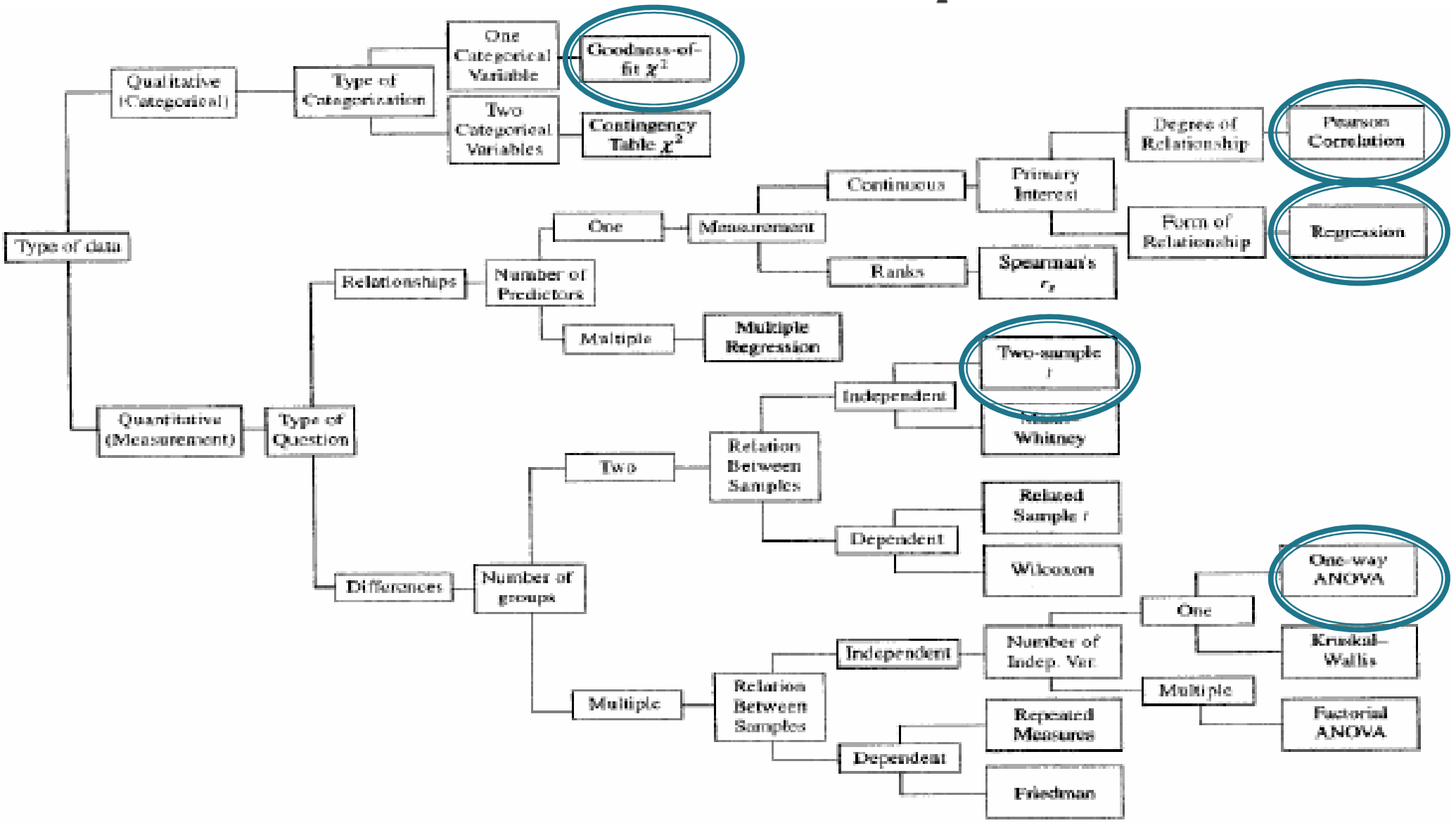


Hypothesis testing (Inferential stats)

- Earlier part was descriptive statistics!
- Null (H_0) vs alternate hypothesis(H_1)
- Null Hypothesis – Statistical hypotheses usually assume no relationship between variables. Eg. there is no association between eye color and eyesight.

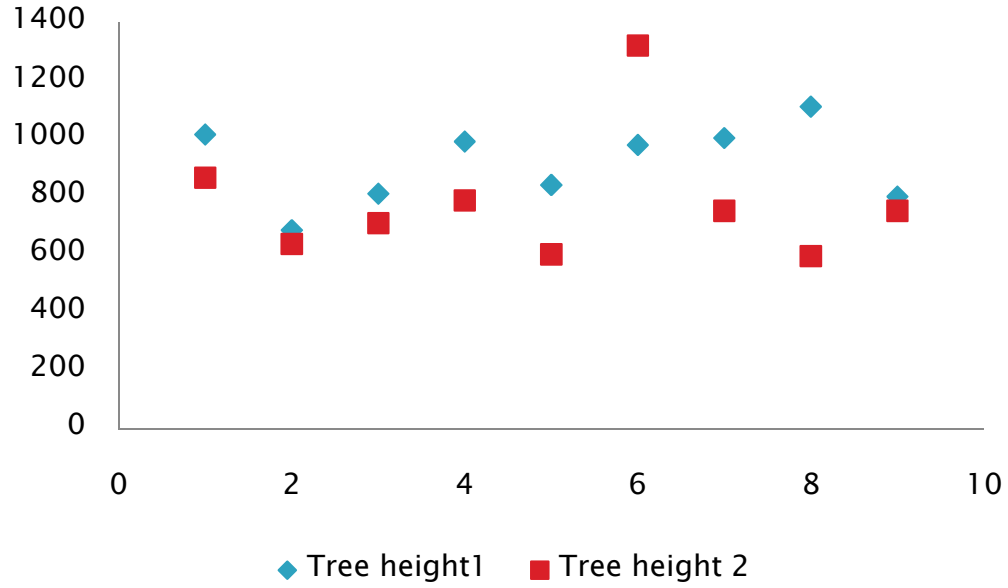
	No change	Change
Change detected	Reject (1)	Accept
No change detected	Accept	Reject (II)

What we will do today?



The data

	Tree height1 (cm)	Tree height 2 (cm)
1	1014	864
2	684	636
3	810	708
4	990	786
5	840	600
6	978	1320
7	1002	750
8	1110	594
9	800	750



Chi(kai)^2 Analysis

- Distribution free analysis
- Comparison between observed data vs expected data
- Generally for qualitative data and categorical data
- Chi square = sum $(O-E)^2/E$

Chi square computed

- At df 8, 818.1, $p<0.001$
- Null hypothesis rejected
- There is a difference between observed vs expected value in tree height

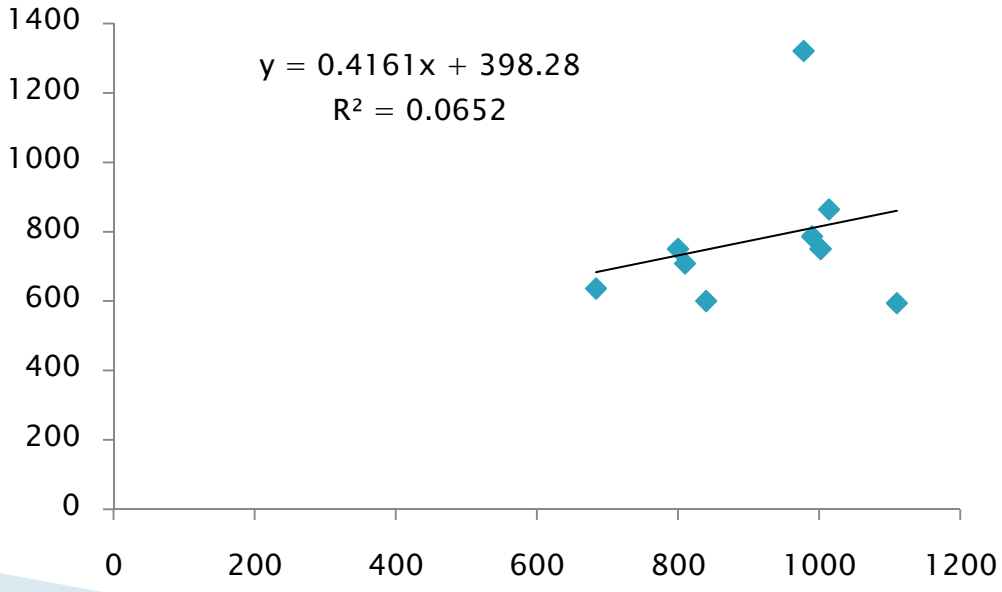
	TreeheightO	TreeheightE
1	1014	864
2	684	636
3	810	708
4	990	786
5	840	600
6	978	1320
7	1002	750
8	1110	594
9	800	750

Correlation and Regression

- Relationship between dependent and independent variables
- Quantification of relationships
- Pearsons correlation coefficient (r)

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

- Simple scatter plot will explain
- r=0.255, R=0.0652



Student's 't' test

- Are the means of two groups different?
- Groups assumed to be normally distributed and of similar size.

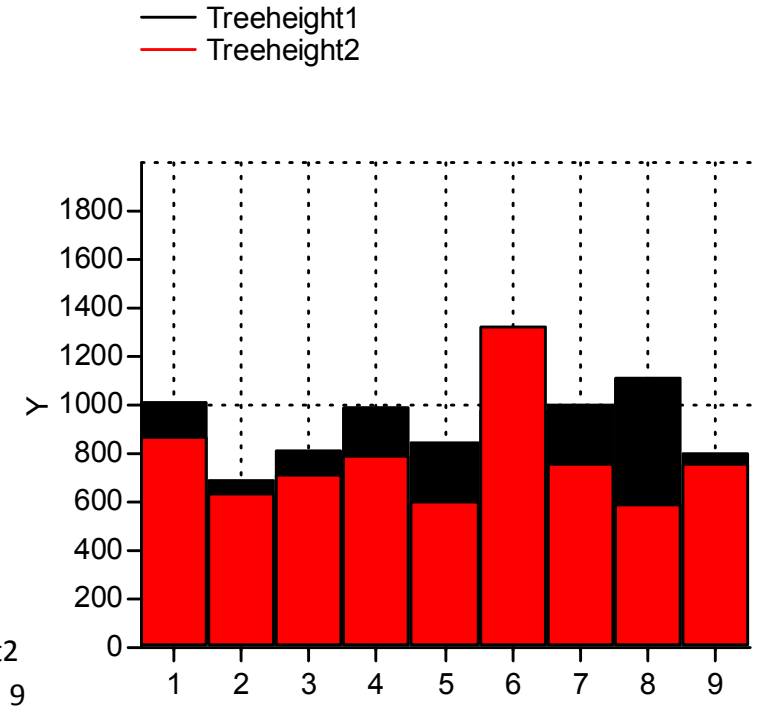
$$t_{\alpha,\nu} = (Y_1 - Y_2) / \sqrt{[(\sigma_1^2 - \sigma_2^2) / n]}$$ (equal sample sizes)

Y_1 and Y_2 are the means of each group
σ_1 and σ_2 are the standard deviations
n is the number of data points in each group
α is the significance level (usually 0.05)
ν is the degrees of freedom

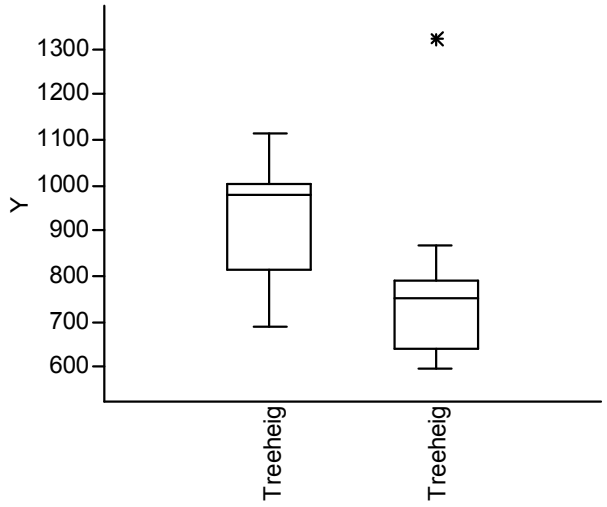
t test calculated!

- t=1.562, p>0.05

	Treeheight1	Treeheight2
1	1014	864
2	684	636
3	810	708
4	990	786
5	840	600
6	978	1320
7	1002	750
8	1110	594
9	800	750
10		



	Treeheight1	Treeheight2
N	9	9
Min	684	594
Max	1110	1320
Sum	8228	7008
Mean	914.222	778.667
Std. error	45.3719	73.9174
Variance	18527.4	49174
Stand. dev	136.116	221.752
Median	978	750
25 prcntil	805	618
75 prcntil	1008	825

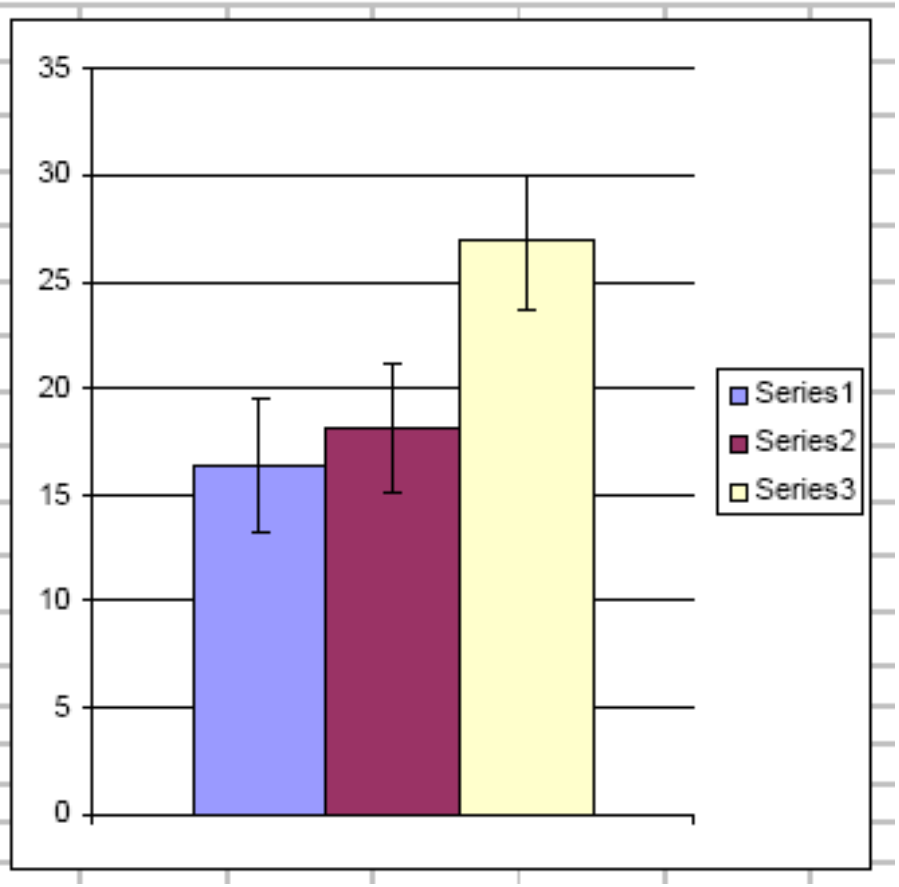


Analysis of Variance (ANOVA)

- F= var between groups/var within group
- Compares the variance of 2 or more groups
- With following Assumptions:
  - Groups relatively equal.
  - Standard deviations similar. (Homogeneity of variance)
  - Data normally distributed.
  - Sampling should be randomized.
  - Independence of errors.

ANOVA…

	12	15	23
	13	15	23
	15	16	25
	15	17	26
	17	17	27
	19	21	30
	20	21	30
	20	23	31
Mean	16.38	18.13	26.88
StDev	3.114	3.091	3.182
StErr	0.794	0.747	0.626



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	506.333	2	253.167	25.855	.000
Within Groups	205.625	21	9.792		
Total	711.958	23			

Coming back to my results…

- 42 species, 7 families
- 27 endemic to Western Ghats (64%)
- 12 vulnerable
- 17 near threatened
- Nandiholé least richness, abundance
- Yenneholé highest

Sub-basin	Richness	Abund.	Endemic	En.Abu	Non-end.	Non.Abu	Simpson	Shannon
Nandi	10	36	4	12	6	24	5.456	1.963
Haridravathi	14	49	6	28	8	21	8.747	2.356
Mavinhole	14	48	8	28	6	20	7.629	2.298
Sharavathi	14	33	9	27	5	6	7.316	2.298
Hilkunji	20	48	11	31	9	17	10.128	2.653
Nagodi	18	59	11	45	7	14	7.548	2.436
Hurli	15	38	9	26	6	12	11.100	2.544
Yenne	22	66	13	35	9	31	13.299	2.807

Results …

Correlation coefficient (r) at significance level ($P<0.05$) between endemic species richness and abundance with habitat, land-use, fragmentation and landscape metrics.

	Richness	Abundance
Habitat variables		
Tree endemism (%)	0.513	
Evergreenness (%)	0.544	
Stream flow (%)	0.817	0.607
Canopy (%)	0.643	0.58
Rainfall (mm)	0.892	0.7
Land-use		
Evergreen–semievergreen (%)	0.853	0.617
Moist deciduous (%)	-0.737	-0.735
Agriculture (%)	-0.734	-0.585
Open land (%)	-0.783	-0.659
Forest fragmentation index		
Interior (%)	0.635	
Patch (%)	-0.709	-0.577
Landscape pattern metrics		
Shape index	0.791	
Contiguous patch (m^2)	-0.809	
Shannon's index	0.842	0.618
Total edge (m)	0.832	0.551
Edge density (#/area)	0.715	

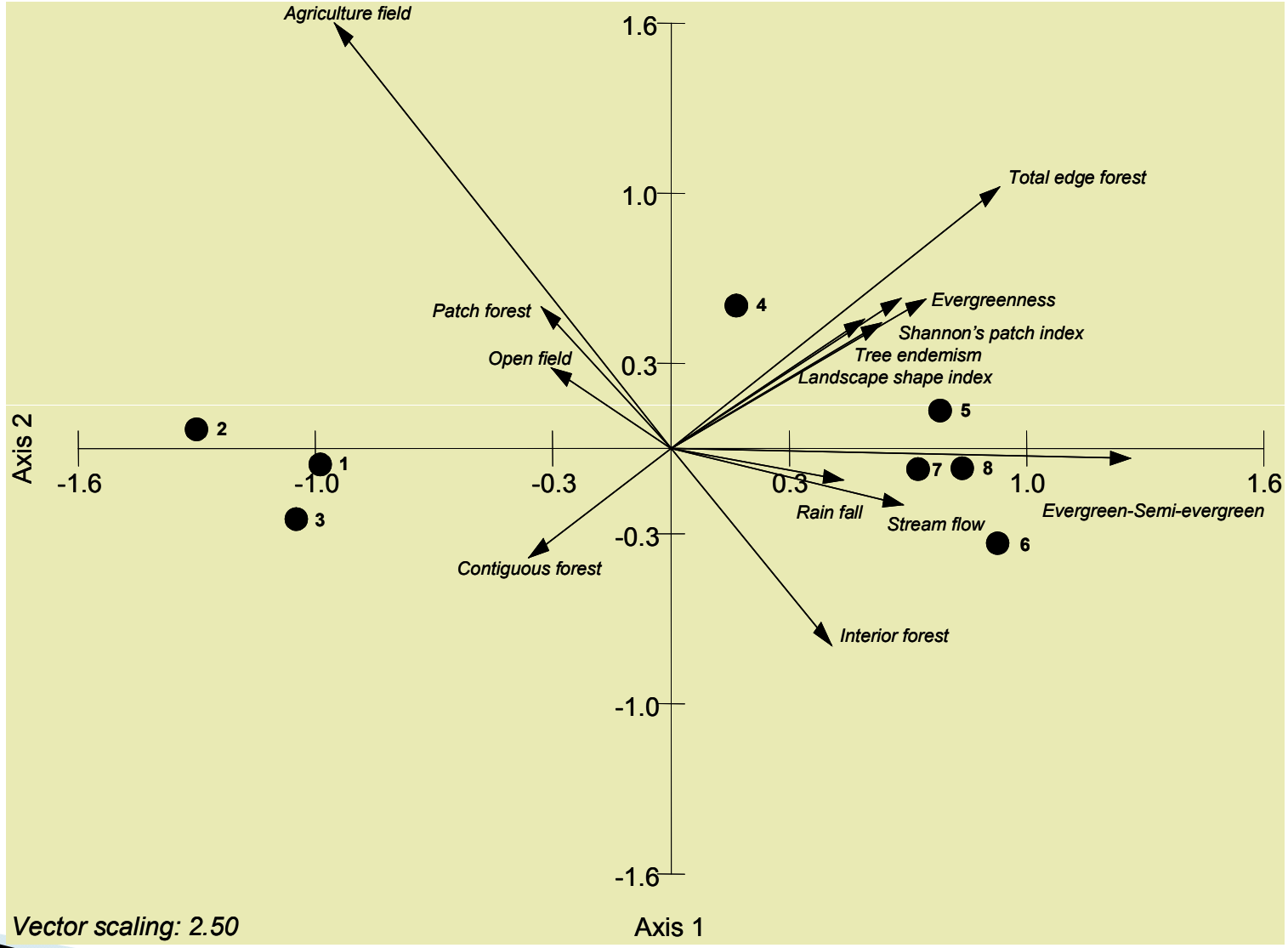
Results …

Correlation coefficient (r) at significance level ($P<0.05$) among the environmental descriptors.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	0.985															
3	0.812	0.855														
4	0.675	0.704	0.628													
5	0.791	0.824	0.973	0.749												
6	0.858	0.878	0.965	0.771	0.992											
7	-0.6	-0.7	-0.8	-0.81	-0.83	-0.8										
8	-0.67	-0.7	-0.92		-0.89	-0.86	0.689									
9	-0.7	-0.73	-0.87	-0.78	-0.93	-0.92	0.755	0.845								
10	0.701	0.695	0.806	0.596	0.841	0.847	-0.54	-0.87	-0.94							
11	0.736	0.706	0.799	0.526	0.828	0.849		-0.8	-0.89	0.962						
12	-0.67	-0.69	-0.86	-0.64	-0.9	-0.88	0.652	0.891	0.976	-0.98	-0.93					
13	0.873	0.9	0.907	0.762	0.925	0.952	-0.76	-0.7	-0.82	0.711	0.765	-0.75				
14	-0.87	-0.89	-0.91	-0.76	-0.92	-0.94	0.782	0.69	0.761	-0.63	-0.68	0.687	-0.98			
15	0.831	0.864	0.923	0.754	0.934	0.944	-0.82	-0.72	-0.77	0.63	0.665	-0.7	0.965	-0.99		
16	0.812	0.858	0.896	0.722	0.907	0.921	-0.8	-0.68	-0.75	0.614	0.661	-0.67	0.977	-0.97	0.969	
17	0.935	0.943	0.906	0.77	0.917	0.956	-0.71	-0.72	-0.84	0.77	0.814	-0.79	0.978	-0.96	0.941	0.921

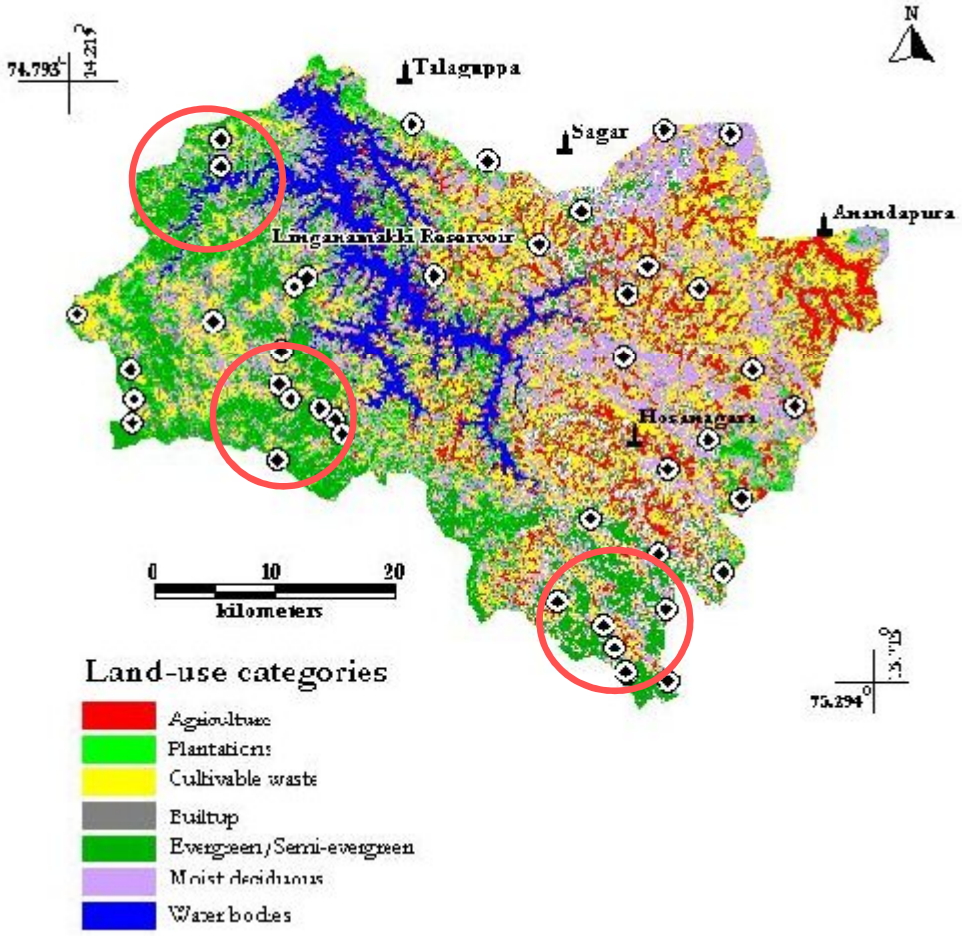
1. Tree endemism; 2. Evergreenness; 3. Stream flow; 4. Canopy; 5. Rainfall; 6. Evergreen-semievergreen; 7. Moistdeciduous; 8. Agriculture; 9. Open land; 10. Interior; 11. Perforated; 12. Patch; 13. Shape index; 14. Contiguous patch; 15. Shannon's index; 16. Total edge; 17. Edge density

Principal Component Analysis...



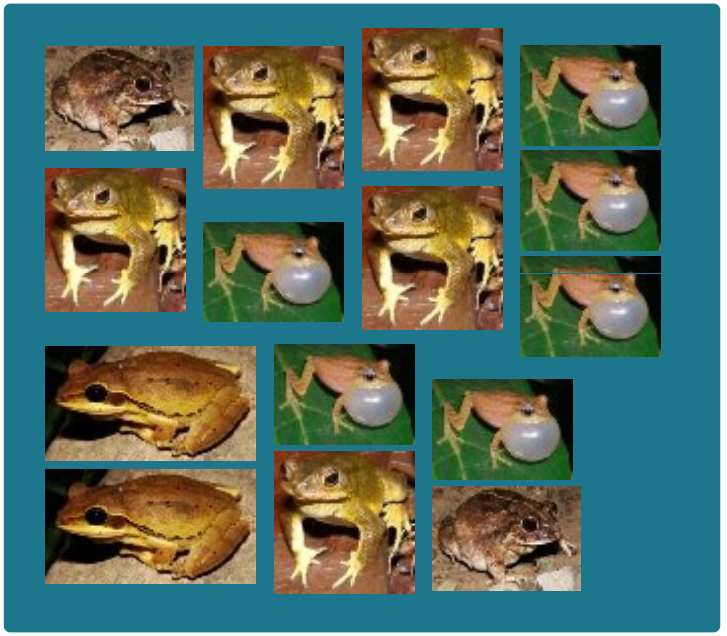
1. Nandihole, 2. Haridravathi, 3. Mavinhole, 4. Sharavathi, 5. Hilkunji, 6. Hurli, 7. Nagodi and 8. Yenneh

Finally…


N
Talaguppa
Sagar
Anandapura
Linganamakki Reservoir
Honnavara
0
10
20
kilometers
Land-use categories
Agriculture
Plantations
Cultivable waste
Builtup
Evergreen/Semi-evergreen
Moist deciduous
Water bodies

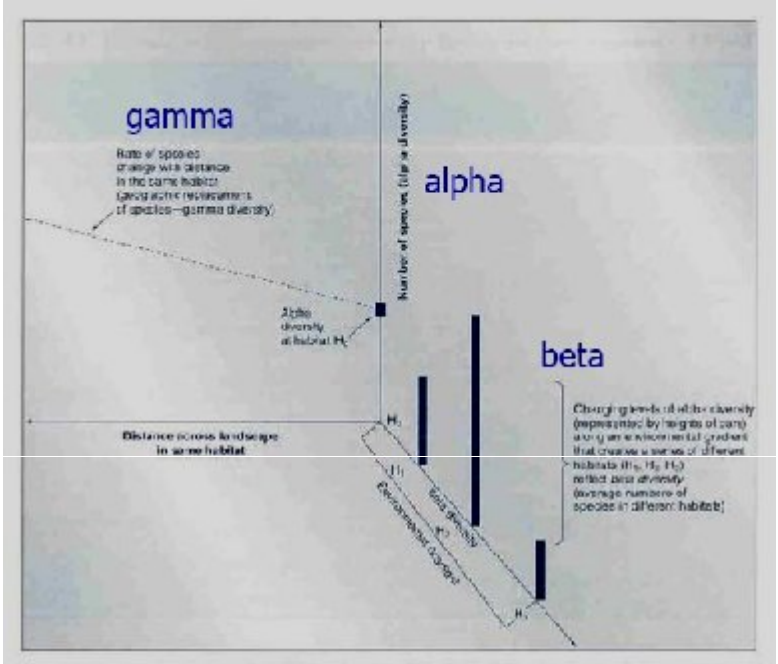
How to Measure Biodiversity?

- More importantly why to measure? ……
- Two unique features
  - Species richness: Varieties in an area
  - Species abundance/evenness: Relative number of individuals of species per unit area, how evenly they are
  - Species diversity: Combination of richness and evenness



Scale based

- Alpha (α) – Number of species in a given habitat or natural community
- Beta (β) – Degree of variation in diversity from patch to patch
- Gamma and Epsilon (γ and ε) – Species richness of a range of habitats in a geographical area (Island)/region. Consequence of Alpha and Beta diversity
- α and γ scalar, only magnitude, β vector both scalar and vector

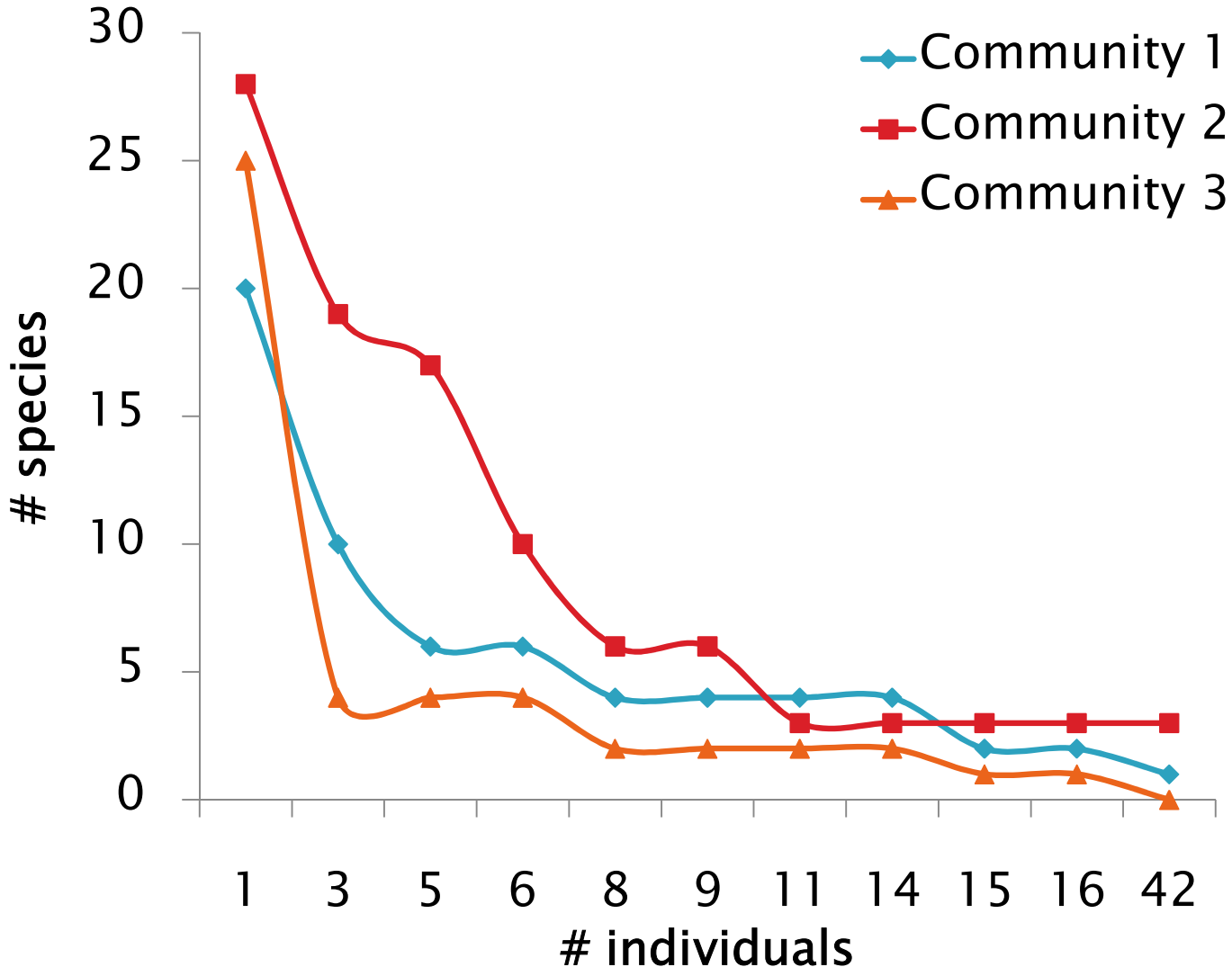


©Van dyke, 2003

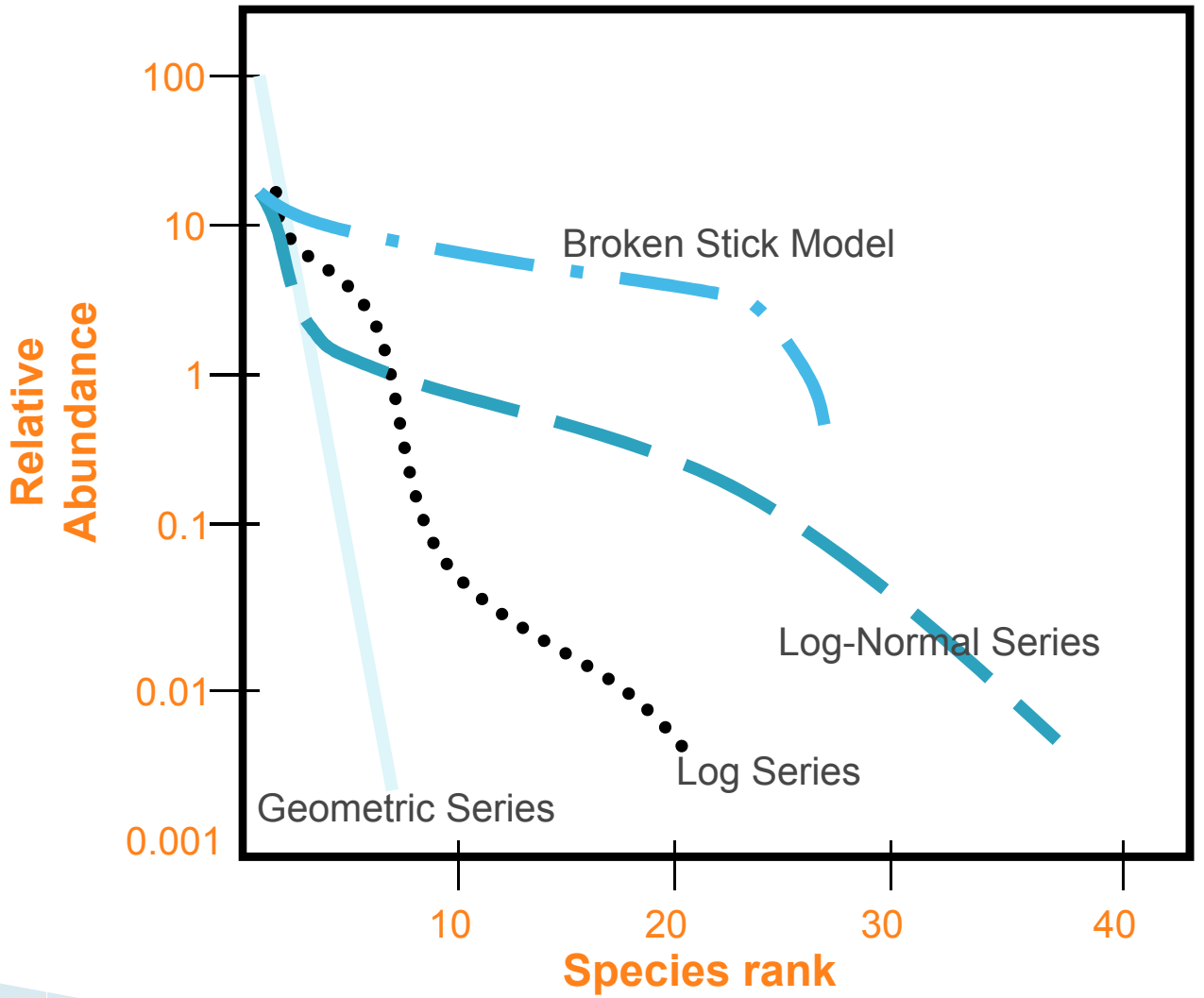
Distribution based

- Parametric – Assumes normal distribution, eg. log-normal, etc.
- Non-parametric – Distribution free, Shannon's, Simpson's, etc.

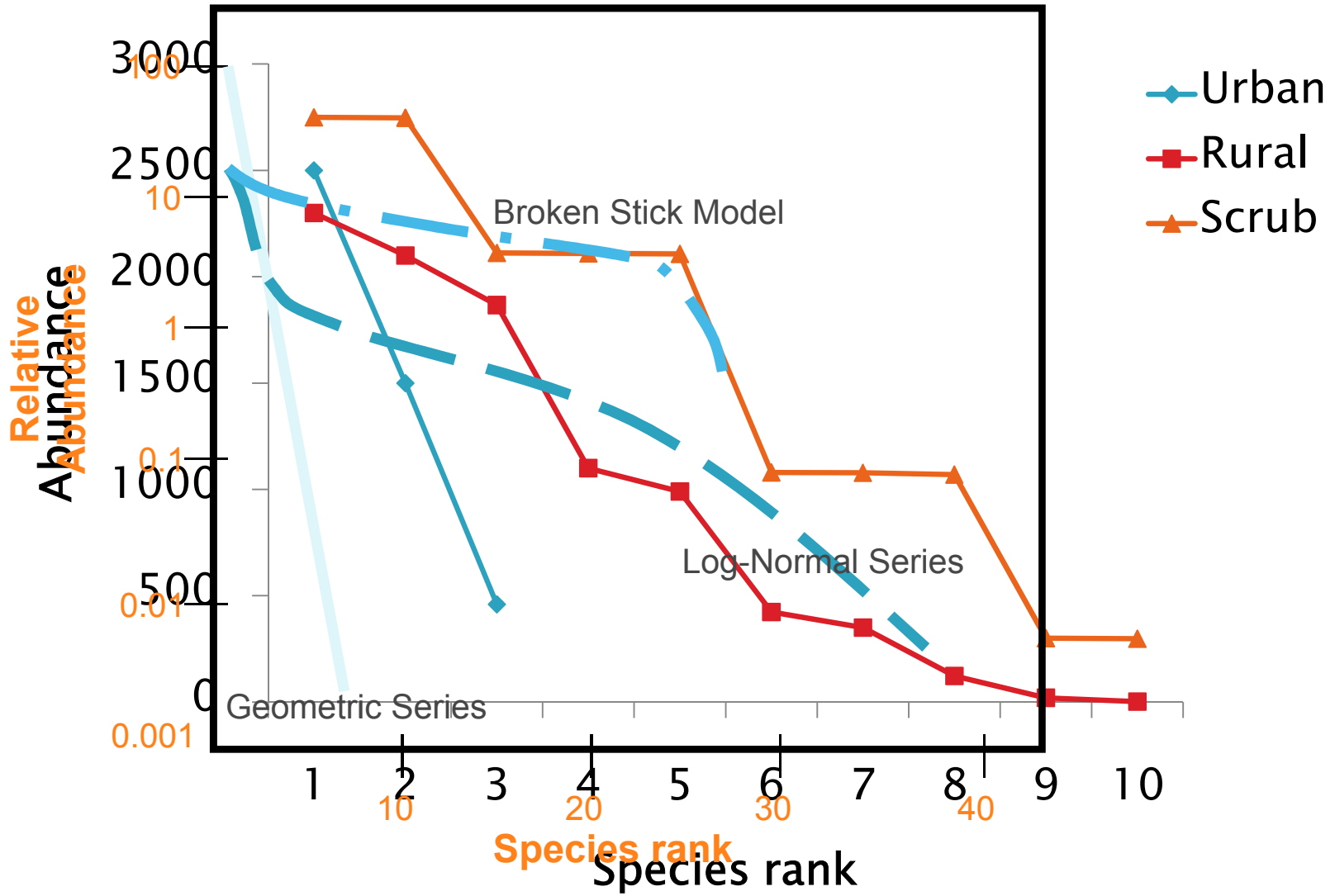
Simplest – Species abundance curve



Curves …Which one to follow?



Rank-abundance curves



Measuring species richness

- Margalef's diversity index
  - $D_{Mg} = (S-1)/\ln N$
- Menhinick's index
  - $D_{Mn} = S/\sqrt{N}$
    - Where S is # of species recorded, N is total # of individuals of all S species
  - $D_{Mg}=1.702$
  - $D_{Mn}=1.201$

Species	# individuals
House crow	4
Jungle crow	12
Common myna	8
Jungle myna	2
House sparrow	5
Pied bush chat	2
Little egret	1
S = 7	**N = 34**
ln N = 3.526	**$\sqrt{N}$ = 5.831**

Alpha Diversity Indices

- Heterogeneity Indices
  - Consider both evenness and richness
  - Species abundance models only consider evenness
- No assumptions made about species abundance distributions
  - Cause of distribution
  - Shape of curve
- Non-parametric
  - Free of assumptions of normality

Two General Categories

- Information Theory
  - Diversity (or information) of a natural system is similar to info in a code or message
  - Example: Shannon-Wiener
- Species Dominance Measures
  - Weighted towards abundance of the commonest species
  - Total species richness is down weighted relative to evenness
  - Example: Simpson, McIntosh

Simpson Index (D)

- Probability of 2 individuals being conspecifics, if drawn randomly from an infinitely large community
- Summarized by letter D or1/D
- D decreases with increasing diversity
  - Can go from 1 – 30+
  - Probability that two species are conspecifics ⇑ with ⇓ diversity
- 1/D increases with increasing diversity
  - $0.0 < 1/D < 10+$

Simpson Index

- Heavily weighted towards most abundant species
  - Less sensitive to changes in species richness
  - Once richness > 10, underlying species abundance is important in determining the index value
  - Inappropriate for some models
    - Log Series & Geometric

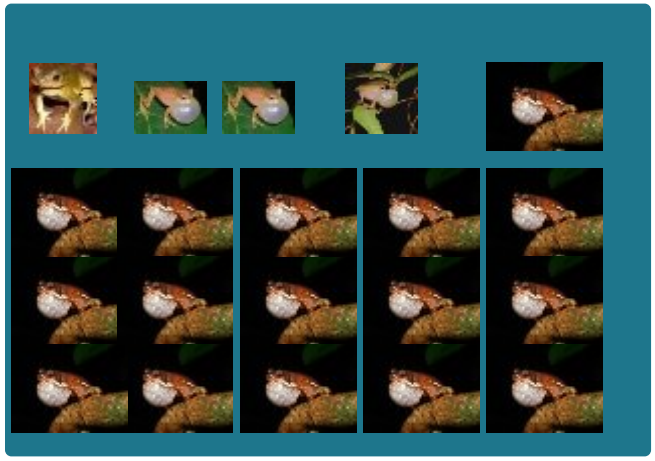
Simpson Calculation …

- Calculate n (individual abundance) and N (total abundance)
- Calculate D
  - $D = \Sigma p_i^2$
  - p_i = proportion of individuals found in the ith species
- Calculate 1/D
  - Increases with increasing diversity

Simpson index

Species	Site 1	Site 2
Pedostibes tuberculosus	5	16
Philautus neelanethrus	5	2
Philautus amboli	5	1
Duttaphrynus melanostictus	5	1

Species	Site 1	Site 2
Pedostibes tuberculosus	.25	.8
Philautus neelanethrus	.25	.1
Philautus amboli	.25	.05
Duttaphrynus melanostictus	.25	.05



Site 1, D= .25, 1/D=4
Site 2, D=.655, 1/D=1.53

Shannon–Wiener Index

- Claude Shannon and Warren Weaver in 1949
  - Developed a general model of communication and information theory
  - Initially developed to separate noise from information carrying signals
  - Subsequently, mathematician Norbert Wiener contributed to the model as part of his work in developing cybernetic technology
- Assumptions
  - All individuals are randomly sampled
  - Population is indefinitely large, or effectively infinite
  - All species in the community are represented

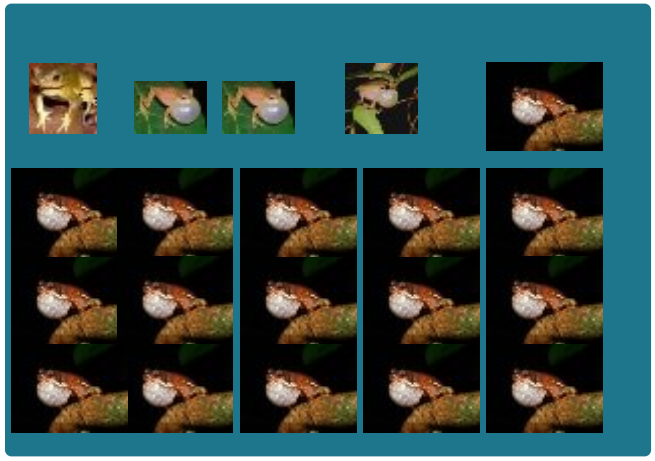
Shannon-Wiener …

- Equation
  - $H' = -\Sigma\, p_i \ln p_i$
    - p_i = proportion of individuals found in the ith species
    - Unknowable, estimated using n_i / N
      - Flawed estimation, need more sophisticated equation (2.18 in Magurran)
  - Error
    - Mostly from inadequate sampling
    - Flawed estimate of p_i is negligible in most instances from this simple estimate

Shannon Wiener (H') index

Species	Site 1	Site 2
Pedostibes tuberculosus	5	16
Philautus neelanethrus	5	2
Philautus amboli	5	1
Duttaphrynus melanostictus	5	1

Species	Site 1	Site 2
Pedostibes tuberculosus	.25	.8
Philautus neelanethrus	.25	.1
Philautus amboli	.25	.05
Duttaphrynus melanostictus	.25	.05



Site 1, H'= 1.39
Site 2, H'=.71

Beta Diversity measures

- Jaccard Index
  - $C_j = j/(a+b-j)$, j is common to both a and b community, a species in community a, b species in community in b,
- Bray-Curtis dissimilarity

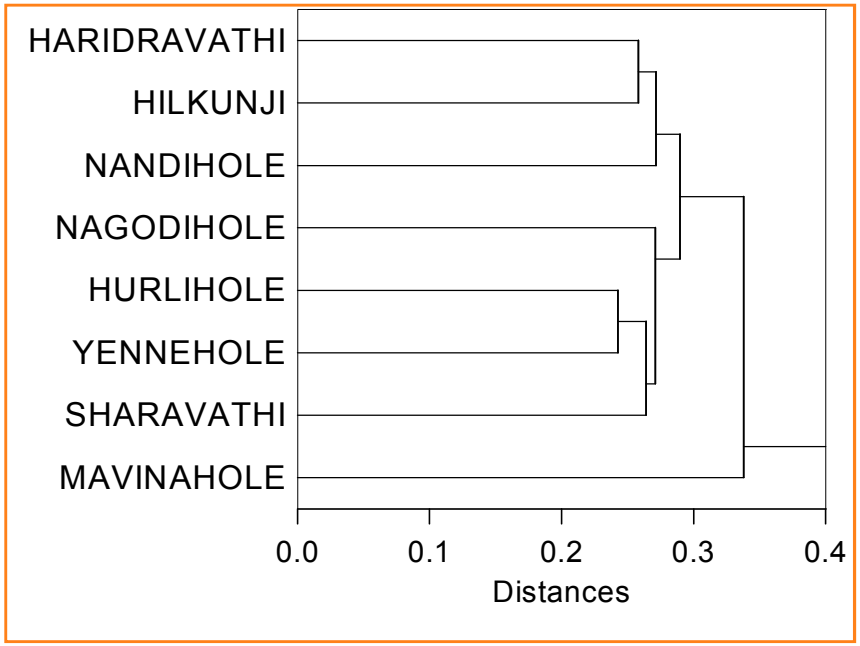
  S = (b+c)/(2a+b+c),

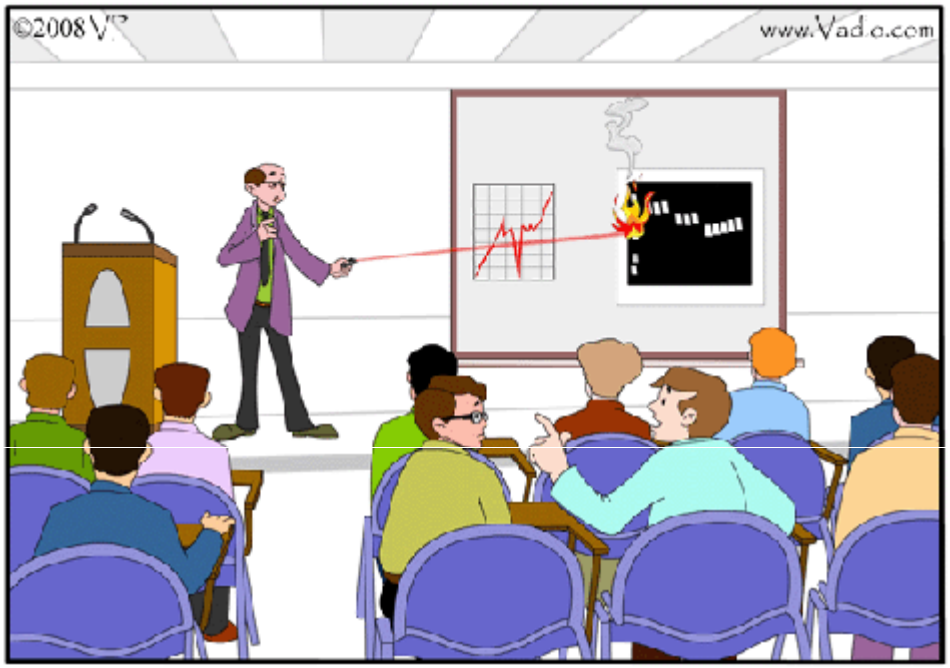
  'b', *'c'* are unique species *'a'* is common to both

Bray–Curtis dissimilarity …

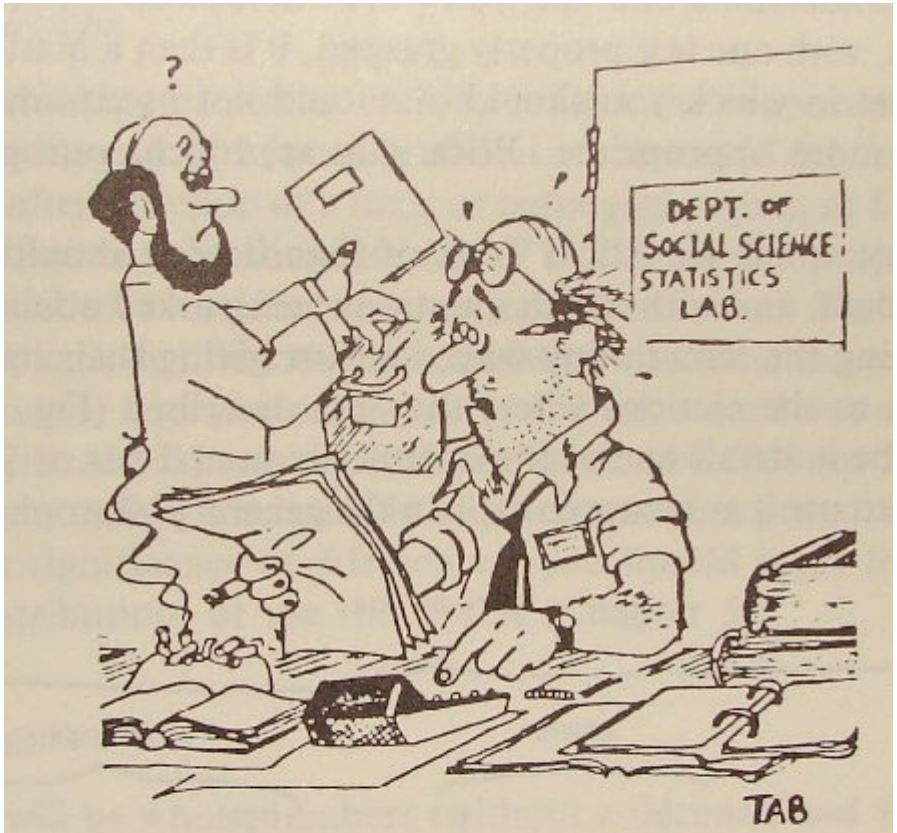
D=b+c/2a+b+c

	HIL	*HAR*	*NAN*	*NAG*	*MAV*	*HUR*	*SHA*
HAR	0.5						
NAN	0.5	0.57					
NAG	0.83	0.71	0.57				
MAV	0.75	0.66	0.67	0.67			
HUR	0.53	0.58	0.58	0.58	0.65		
SHA	0.68	0.74	0.62	0.62	0.68	0.68	
YEN	0.62	0.64	0.57	0.5	0.63	0.46	0.49





No wonder. He was stuck on this slide for last 15 minutes!



Dash it all! 'Good' and 'Evil' MUST be significantly different!!
(Courtesy of *Trends in Biochemical Sciences* and *Elsevier Science Publishers*)

Thank you