Structure – a FORTRAN program for ecological pattern analysis

Version 1.0

Werner Ulrich

Nicolaus Copernicus University in Toruń

Department of Animal Ecology

Gagarina 9, 87-100 Toruń; Poland

e-mail: ulrichw @ uni.torun.pl

Latest update: 12.04.2005

1. Introduction

Structure is a small program that computes basic statistics for data sets like body size ratios, abundance ratios, internal variances or kernel densities.

The program was especially designed for analyzing body size distributions. The present version contains six modules

Basic statistics

Histograms

Species / genus ratios

Variance test (log and linear scale)

Ratio test (log and linear scale)

Frequencies

Pairwise regression

Kernel density estimate

2. Data structure

ComStruc needs unformatted ASCII files (text files) as shown in the table beside. The first line must always be a comment line that starts with an asterisk (*). The data set need not to be sorted. Data have to separated by one or more spaces. Matrix size is only limited by the available computer memory. However, analysing larger matrices may become very time consuming.

The data file has to contain three input columns, two columns for the filter variables (for instance genus and species names or family and

Genus	Species	Length in mm
Halidamia	affinis	5.50
Periclista	albida	6.00
Periclista	albiventris	6.50
Periclista	albipennis	6.50
Periclista	lineolata	6.50
Periclista	pubescens	7.50
Ardis	brunniventris	5.50
Cladaris	elongatulus	7.00
Monardis	plana	6.50
Monardis	semicinetus	5.25
Pareophora	pruni	5.50
Eupareophora	exarmata	6.00
Tomostethus	nigritus	7.50
Eutomostethus	luteiventris	6.00
Eutomostethus	gagathinus	5.50
Eutomostethus	punctatus	5.50
Eutomostethus	ephippium	4.50
Phymatocera	aterrima	8.50
Rhadinoceraea	micans	7.50
Rhadinoceraea	reitteri	8.50
Rhadinoceraea	nodicornis	6.50
Rhadinoceraea	gracilicornis	5.50

Item1	Item2	Occ	Mean	StD	CV	Skewness	Range
Halidamia	affinis	1	5.500000	0.000000	0.000000	0.000000	0.000000
Periclista	albida	5	6.600000	0.489898	0.074227	1.807351	1.500000
Ardis	brunniventris	1	5.500000	0.000000	0.000000	0.000000	0.000000
Cladaris	elongatulus	1	7.000000	0.000000	0.000000	0.000000	0.000000
Monardis	plana	2	5.875000	0.625000	0.106383	0.000000	1.250000
Pareophora	pruni	1	5.500000	0.000000	0.000000	0.000000	0.000000
Eupareophora	exarmata	1	6.000000	0.000000	0.000000	0.000000	0.000000
Tomostethus	nigritus	1	7.500000	0.000000	0.000000	0.000000	0.000000
Eutomostethus	luteiventris	4	5.375000	0.544862	0.101370	-1.738730	1.500000
Phymatocera	aterrima	1	8.500000	0.000000	0.000000	0.000000	0.000000
Rhadinoceraea	micans	4	7.000000	1.118034	0.159719	0.000000	3.000000

guild membership name) and one column (the third) that contains a metrically scaled variable in the Fw.x format like body size, abundance or variability.

The Table on page 1 shows an example of the data file. It contains data on mean body length of 11 sawfly genera.

3. Program run

After the input of the data file name the programs counts the number of data sets and shows the module list

Next the program asks for the missing value. The default is zero but if you want to include zeros you should give another value for instance –999. Lastly, you have to give the precision of your data. This is important for the randomisation test. In the data set above, for instance you measured body length in all but one cases with an accuracy of 0.1mm. In one case the length is given with an accuracy of 0.05 mm. In this case you can give the accuracy level either as 0.1 or as 0.05. Different inputs will effect the results of the Monte Carlo simulations used for the ratio and the variance tests.

The program generates a single output file 'Output.txt'

4. The basic and histogram modules

ComStruc groups the data according to the first data column (the grouping variable) and gives the number of occurrences, the mean, the standard deviation, the coefficient of variation, skewness and the range of data (maximum—minimum value).

Skewness is computed from

$$\gamma = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{S} \left(\frac{n_i - \bar{x}}{s} \right)^3$$

The standard error of the skewness can be approximated by

$$SE(\gamma) = \sqrt{\frac{6}{n}}$$

Item1	Item2	Occ	MeanData	StDData	MeanRand	StDRand
Periclista	albida	5	1.059295	0.064321	1.058371	0.043499
Monardis	plana	2	1.238095	0.000000	1.238130	0.000000
Eutomostethus	luteiventris	4	1.104377	0.091220	1.103209	0.069760
Rhadinoceraea	micans	4	1.156332	0.019873	1.162380	0.111990

With the option *histo* the program computes a histogram of frequencies per body weight class. The number of classes is assumed to be a quarter of the number of species.

5. The ratio module

Structure computes the ratio test of Strong et al. (1979; see also Tonkyn and Cole 1986) either at a log scale or at a linear scale. Assume you have S species of a genus X and measured body weights of them. The log ratio test now asks whether the ratios of weights w_k / w_{k-1} of ranked species (from largest to smallest) are more or less regular than expected by chance. The program determines the S-1 observed ratios and determines the mean ratio and its standard deviation. Next it generates 5000 random assemblages with S species, which have body weights that are either linearly randomly distribution inside the observed range of weight or have a truncated normal distribution (mean = (max + min / 2); standard deviation = (max - min) / 4) or stem from an empirical distribution. In the case of the linear random distribution the minimum and maximum values are set by the smallest and largest species and are fixed in the Monte Carlo assemblages. Hence S - 2 species weights are generated at random. Next the program determines the mean size ratio of these 5000 assemblages and the associated standard deviation. If the body size distribution of the respective species pool is known, this empirical distribution should be used. Required is a data file that has the same structure than the input data file shown above.

The Table of the previous side shows the output file for the log ratio test with the data set on page 1. The mean ratio of the data are nearly identical with the mean of the simulation (linear random numbers).

Sometime (for linearly scaled data it seems more appropriate to use absolute differences w_k - w_{k-1}. This option is computed after the input rlin.

6. The Variance test

The variance test asks whether the data are more or less evenly spaced inside the observed range. Hence, is the observed variance of w_k - w_{k-1} (varlin) or w_k / w_{k-1} (varlog) lower (even spacing) or higher (clumped distribution) than expected from a random assemblage. Again *Structure* generates 5000 random assemblages with S species, which have body weights that are either linearly randomly distributed or have a normal distribution or are obtained from an empirical distribution. The observed standard deviations are compared with the expected ones. Structure computes the standard deviation of the expected mean standard deviation, its skewness and the upper and lower 2.5 percentiles. The Table below shows the output for varlin. A difference between the observed and expected standard deviation occurs for the genus *Periclista*. However, due to the right skew of standard deviations obtained from the Monte Carlo distributions no significant deviation at the 5% error levels occurs.

Item1	Item2	Occ	MeanData	StDData	MStDRand	StDRand	SkewRand	97.5%	2.5%
Periclista	albida	5	0.375000	0.414578	0.271598	0.100486	0.358859	0.485412	0.090139
Monardis	plana	2	1.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Eutomost.	Luteiv.	4	0.500000	0.408248	0.328632	0.139116	0.189007	0.604152	0.070711
Rhadinoc.	micans	4	1.000000	0.000000	0.650299	0.278950	0.219487	1.237605	0.147196

Item1	Item2	Spec1	Spec2
Chalcis		7.50	6.75
Chalcis		7.50	7.00
Chalcis		6.75	7.00
Brachymeria		6.00	4.75
Brachymeria		6.00	4.75
Brachymeria		6.00	3.75
Brachymeria		6.00	3.25
Brachymeria		6.00	3.10
Brachymeria		6.00	3.75
Brachymeria		6.00	4.35
…..			

7. Frequencies

The frequency module counts numbers of occurrences of the first grouping variable. The output file contains only two columns. One (Ind) for the number of occurrences (for instance the number of species per genus and the second how often this occurred in the data file.

8. Regression

Pairwise regressions of species within a genus has been introduced by Smith et al. (2004) for analyzing heredity patterns within larger taxa. Regression gives all pairwise combinations within the first filter item. For items with more than 10 occurrences it selects at random 20% of all possible combinations. A typical output file shows the next table. It is part of a larger data set of mean body length of European chalcidoid species. For all species the plot below shows the regression of species 1 against species 2. There are 4551 species combinations. Although there is a highly significant regression, the coefficient of variation is quite low and the heredity of body length in the Chalcidoidea is lower than reported from other taxa.

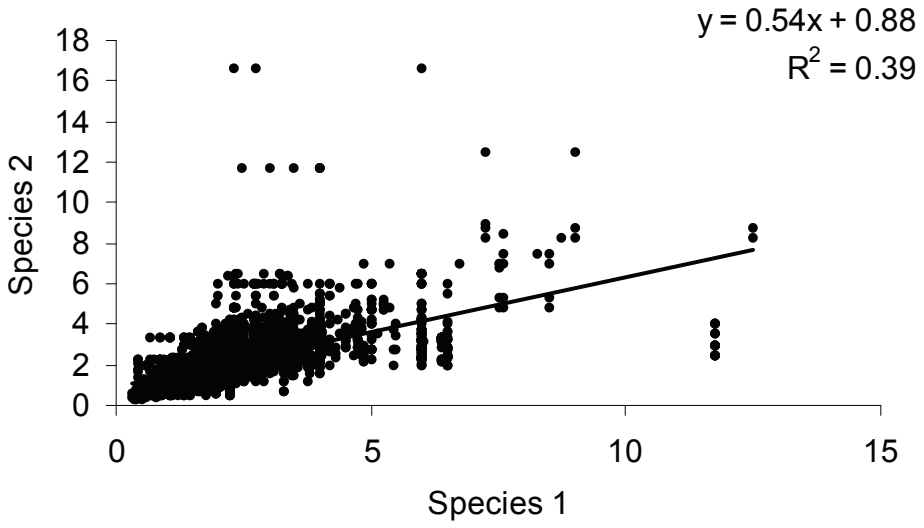
9. Kernel density analysis

Kernel density analysis is a tool for studying patterns in histograms when the way of classification influences the results. The goal of kernel density estimation is to approximate the true frequency distribution of a variate X by the sum of n predefined random distributions shifted along the range of X by steps of μ. Structure uses a normal density estimator of the form

$$f_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} Exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{h}\right)^2\right)$$

The bandwidth h can take three values

Smooth: $h = range/10$

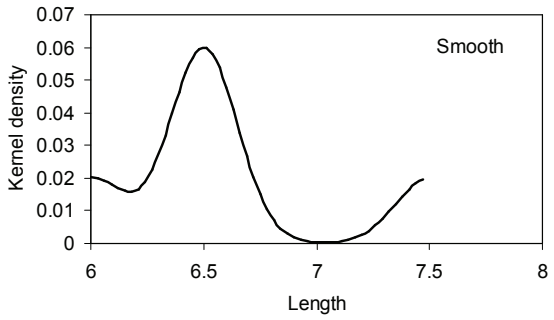


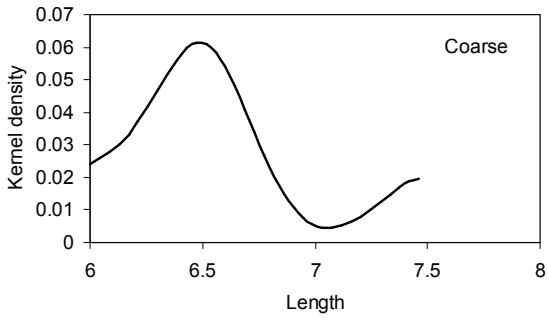
Item1	Item2	Occurr	MinSize	MaxSize	Mean	Variance	Range	Bandwidth	KernelW
Periclista	albida	5	6.000000	7.500000	6.375000	0.240000	1.500000	0.150000	0.030000

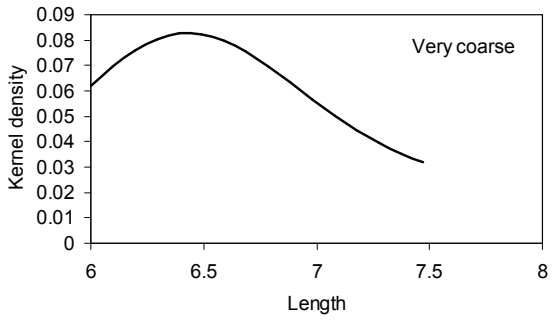
Interval	StdKernelDensity
6.000000	0.020183
6.030000	0.019998
6.060000	0.019228
6.090001	0.018093
6.120001	0.016906
6.150001	0.016035
6.180001	0.015861
6.210001	0.016723
6.240002	0.018873

Item1	Item2	Interval	KernelDensity	Peaks/Dole
Periclista	albida	6.000000	0.020183	1.000000
Periclista	albida	6.180001	0.015861	0.000000
Periclista	albida	6.510004	0.059783	1.000000
Periclista	albida	7.020007	0.000266	0.000000
Periclista	albida	7.470010	0.019557	1.000000

*Item1	Item2	Occ	Peaks	Doles	Mode	Skewness	ModeRatio
>Periclista	albida	5	3	2	6.5100	1.1649	0.3376







Coarse: $$h = 1.06n^{-0.2}\ \min(\sigma_x; range/5.36)$$

Very coarse: $$h = \sigma_x$$

(cf. Silvermann 1986) with n being the number of observations, range the range size of x (max_x—min_x), σ_x the standard deviation of x. The step width is h/5 starting from min_x. The sum of all $f_h(x,\mu)$ gives the ultimate frequency distribution. Kernel density analysis

*
Pterostichus oblongopunctatus 6.60
Oxypselaphus obscurus 23.00
Patrobus atrorufus 18.70
Synuchus vivalis 13.00
Pterostichus strennus 10.00
Pterostichus melanarius 6.80
Pterostichus diligens 28.00
Carabus granulatus 11.60
Harpalus 4-punctatus 6.20
Pterostichus nigrita 7.60
Carabus nemoralis 8.50
Amara brunea 13.00
Badister bullatus 6.20
Stomis pumicatus 6.00
Platynus assimilis 5.70
Leistus terminatus 13.00
Pseudoophonus rufipes 16.10
Pterostichus anthracinus 12.10
Pterostichus minor 8.60
Notiophilus palustris 8.00

transforms therefore a discrete into a continuous frequency distribution.

The Table on the next side shows the output for the genus Periclista of the data file on page 1. A smooth estimate was computed. The program first gives some basic statistics and shows then the kernel density estimates for each interval μ. Plots of f(h,μ) against μ for smooth, coarse and very coarse estimates show the following Figures. The number of peaks depends therefore on the bandwidth. In most cases a coarse bandwidth will give the most stable results.

Structure gives the number of peaks and doles and their locations, the mode of the estimate and the skewness of the density function.

Lastly it computes the mode ratio. This is the ratio of the peak values of the major and the second mode of the distribution. In the Table above this is the quotient of 0.020183 / 0.059783 = 0.3376.

S/G	StDev	Sim S/G	SimStDev	2.5%	97.5%
1.538	1.599	1.411	0.162	1.176	1.818

10. Species / genus ratios

The analysis of S/G ration needs two input files, a first file containing the data and a second file containing the distribution in the whole sample universe (the metacommunity). *Structure* compares the observed S/G ratios (file 1, example of ground beetle body length beside) with 1000 randomly drawn samples of the same size as the data from the universe file. *Structure* samples therefore species at random from the universe. The universe file must be larger than the data file. The output is simple and shown above. The program gives S/G, its standard deviation and the mean simulated S/G and its standard deviation. Additionally the upper and lower 2.5 percentiles of the simulation are given.

11. Citing *Structure*

Structure is freeware but nevertheless if you use *Structure* in scientific work you should cite *Structure* as follows:

Ulrich W. 2005 - Structure – a FORTRAN program for ecological pattern analysis - www.uni.torun.pl/~ulrichw

12. Acknowledgements

The development of this program was supported by a grant of the Polish Science Committee (KBN, 3 F04F 03422).

13. References

Silvermann B. W. 1986. Density estimation for statistics and data analysis. Chapman & Hall, New York.

Smith F. A. et al. 2004. Similarity of mammalian body sizes across the taxonomic hierarchy and across space and time. Am. Nat. 163: 672-691

Strong D. R., Szyska L. A., Simberloff D. 1979. Tests of community wide character displacement against null hypotheses. Evolution 33: 897-913.

Tonkyn D.W., Cole B. J. 1986. The statistical analysis of size ratios. Am. Nat. 128: 66-81.