

A note

I would like to draw to friends attention our new, 397 page Statistical Ecology - the quantitative exploration of Nature to reveal the unexpected - prepared with the assistance of M. Mihály DFE. The soft cover version (ISBN 9781453760529) can be ordered through the AMAZON (CreateSpace) distribution system (Book # 3476529 or type Statistical Ecology). The external appendices are available on disc from scada.london@gmail.com . The book on CD is listed by eBay under item number 140426343983. For table of contents, please click link "Stat's table of contents" on this page. The book's topics traverse many problem areas in univariate and multivariate data analysis. It assumes advanced training in community and population ecology, and familiarity with first year college algebra. The topics are developed from the very basic to the more complex in a continuum, assuming no previous training in statistics The manner of presentation emphasizes reasoned methodological choices and encourages innovations consistent with the objectives, but mindful of the need to see clearly the regularity conditions which set limits for valid application of statistics in ecology. The main text comes packaged with external appendices including a technical manual, over 40 specialized application programs, and many data files taken from the exercises in the main text. The programs are conversational, designed to challenge the user by requiring reasoned choices at different points as the analysis unfolds. The programs will run on 32 bit Windows (XT and up), but have to be placed high up in the directory. For further information please write to lorloci@uwo.ca.

L. Orlóci

Table of Contents

Preface	15	
Chapter 1 Terms and concepts		19
1.1 Unit, type, population, community		19
1.2 Attributes	20	
1.3 Data types	22	
1.4 Domain	23	
1.5 Randomness	25	
1.6 Context	27	
1.7 True or false	27	
1.8 Planning	28	
Chapter 2 Historic perspective		30
2.1 Problem lines	30	
2.2 The study scenario	33	
2.3 Data storage	33	
Chapter 3 Measurement theory		36
3.1 Description of objects	36	
3.2 Measuring scales	38	
3.3 Errors in measurements	40	
3.4 Measurement errors in functions	41	
3.5 Significant digits	43	
Chapter 4 Population description		47
4.1 System of moments and product moments		47
4.1.1 Moments	47	
4.1.2 Product moments	49	
4.1.3 Entropy and information	54	
4.2 Theoretical distributions	61	
4.2.1 Poisson distribution	61	
4.2.2 Bernoulli distribution	62	
4.2.3 The Normal distribution	63	
4.3 Descriptors of distribution shape	67	
4.4 Common transformations	69	
Chapter 5 Sampling		73
5.1 General	73	
5.2 Sampling frame	74	
5.3 Simple random sampling	75	
5.4 Stratified random sampling	76	
5.4.1 Allocation by stratum size	76	
5.4.2 Allocation by stratum variance	76	
5.5 Randomly sited systematic sampling	77	
5.6 Multistage sampling	78	
5.7 Preferential sampling	78	
5.8 Sample optimality	79	
5.8.1 Quadrat-based estimation	79	
5.8.2 Quadrat-based structure detection	80	
Chapter 6 Inference: sample to population		83
6.1 Estimation	83	
6.1.1 Consistent estimator	83	
6.1.2 Unbiased estimator	84	
6.1.3 Minimum sampling variance	84	
6.2 Estimation of entropy	84	
6.3 Estimation of information	88	
6.3.1 Estimation of mutual information	88	
6.3.2 Estimation of interaction information	89	
6.4 Moments and moment based quantities	92	
6.5 Product moments and related quantities	94	
6.6 Estimation in stratified random sampling	95	
6.7 Estimation in systematic sampling	96	
Chapter 7 Commonness and probability		98
7.1 Which kind of distribution?	98	
7.2 Normal probabilities	99	
7.3 Sampling distributions	102	
7.3.1 Distribution of the mean	103	
7.3.2 The chi-squared distribution	103	
7.3.3 Sampling distribution of t	104	
7.3.4 The F distribution	104	

7.4 Empirical sampling distributions	106	13.2 Response type a plane	225
7.5 Setting confidence limits	107	13.3 Response type a polynomial	230
7.5.1 Point vs. interval estimation	107	13.4 Response type a product or exponential	232
Chapter 8 Measuring resemblance	111	13.5 Working with residuals	233
8.1 Comparison space	111	Chapter 14 Character analysis: importance a posteriori	235
8.2 Minkowski metrics	113	14.1 Multiple correlation	235
8.3 Product moment	116	14.2 Specific variance	238
8.4 Mean square contingency	118	14.3 Sum of squares	239
8.5 Indices of similarity and dissimilarity	119	14.4 Information	240
8.6 Goodall's probability index	121	14.5 Weighting variables: a discussion	246
8.7 Calhoun's distance	123	Chapter 15 Exploration of multidimensional data: the preliminaries	248
8.8 Plexus diagrams	124	15.1 Multidimensional or multivariate	248
8.9 Invariance	125	15.2 Views of the medium	249
Chapter 9 Stating and testing hypotheses	126	15.3 Broad objectives	251
9.1 Basic types	126	Chapter 16 Exploring continuous multidimensional variation	254
9.2 General	126	16.1 Two transformations	254
9.3 Procedure	127	16.2 Component analysis	255
9.4 Simple, composite and mixed hypotheses	128	16.2.1 First transformation $\square 1$	255
9.5 One and two-sided alternatives	129	16.2.2 Second transformation $\square 2$	256
9.6 Parametrised and non-parametrised H_0	130	16.2.3 Algorithm	257
9.7 Errors in probabilistic decisions	130	16.2.4 Dimensions of the significant trended variation	259
9.8 Bartlett's paradox	131	16.2.5 A complete example	260
Chapter 10 Probabilistic comparisons I	133	16.2.6 Presentation of the PCA results	262
10.1 Sample mean and a standard	133	16.3 MDSCAL: a flexible method	264
10.2 Sample variance and a standard	134	Chapter 17 Exploring group structure in the data	267
10.3 Sample distribution and a standard	135	17.1 Single link clustering	267
10.3.1 Chi-squared divergence	136	17.2 Centroid clustering	268
10.3.2 I-divergence information	137	17.3 Sum of squares clustering	272
10.3.3 The Kolmogorov-Smirnov divergence	137	17.4 Association analysis	275
10.4 Sample mean vector and standard	139	17.5 Analysis of structured tables	278
10.5 Sample covariance matrix and standard	142	17.5.1 The data table	278
Chapter 11 Probabilistic comparisons II	145	17.5.2 Compositional sharpness of blocks	280
11.1 Sample variances	145	17.5.3 Compositional gradients	282
11.2 Sample means	146	17.5.4 Dimensionality	284
11.3 Several variances and means	147	17.5.5 Identification of underlying factors	285
11.3.1 Complete randomized design	148	17.5.6 Partitioning the deviations	287
11.3.2 Randomized block design	158	Chapter 18 Exploration of affinities: identification	291
11.3.3 Latin square design	161	18.1 Approaches	291
11.3.4 TWO-FACTOR DESIGN	164	18.2 Generalized distance	292
11.4 Testing homogeneity in discrete data	166	18.3 Discriminant function	294
11.4.1 Null hypotheses for homogeneity	167	18.4 Information divergence	297
11.4.2 Test criteria	168	18.5 A case of probabilistic classification	298
11.4.3 Homogeneity of replicates	168	18.6 Group identification	300
11.4.4 Homogeneity of treatment means	169	Chapter 19 Trajectory analysis of time series data	302
11.5 Comparison of k covariance matrices	170	19.1 Historic setting	302
11.6 Comparison of k group mean vectors	172	19.1.1 The Kernerian line	302
Chapter 12 Probabilistic comparisons III	177	19.1.2 Surrogate mathematical models	303
12.1 Two continuous variables	177	19.1.3 Ecological models	303
12.2 Two binary variables	179	19.1.4 Focus on process governance	303
12.3 Two multistate discrete variables	181	19.2 Units of organisation	304
12.4 Sets of variables	184	19.3 Components of change, multi scaling	304
12.5 Nested character hierarchies	190	19.4 Indicators of change	305
12.5.1 Basic concepts	191	19.5 Phase space: the reference system	305
12.5.2 Additive partitions	193	19.6 The model and the data	307
12.5.3 Comparison of entire communities	195	19.7 First-order objectives	309
12.5.4 Testing the significance of r	196	19.8 Compositional transition scalars	310
12.5.5 Interpretations of correlation profiles	198	19.8.1 Euclidean distance	310
12.6 Unbalanced nested hierarchy	200	19.8.2 Acute angle	311
12.6.1 Overview	200	19.8.3 Compositional transition velocity	311
12.6.2 The isolation problem in general	201	19.8.4 Acceleration, deceleration	312
12.6.3 The hierarchical relev�	203	19.8.5 Angular velocity	312
12.6.4 Technical details	208	19.9 Synchronicity scalars	313
12.6.5 Decomposition of sum of squares	209	19.9.1 Product moment correlation	313
12.6.6 Decomposition of product moment	210	19.9.2 The topological similarity coefficient	316
12.6.7 Further on the partial variance	211	19.10 The Hausdorff (fractal) dimension	322
12.6.8 Results and interpretation	212	19.11 Divergence scalars	323
12.6.9 Remarks	215		
12.6.10 Conclusions	216		
Chapter 13 Trend seeking: univariate response	219		
13.1 Response type a straight line	219		

19.11.1 Rényi's entropy of order α	323	20.8 Tests on the markovity of an observed series	335
19.11.2 Rényi's information of order α	324	20.8.1 H_0 : the series is zero order Markov	336
19.11.3 Pooled squared deviations	324	20.8.2 H_0 : series mth order Markov	337
19.12 Complex trajectory properties	324	20.9 Comparison of test options	338
19.12.1 Phase structure	324	Chapter 21 Diversity partitions	340
19.12.2 Determinism	325	21.1 General remarks and data coding	340
19.12.3 Periodicity	325	21.2 Entropy and information partitions	341
Chapter 20 The Markov chain	327	21.3 Example	343
20.1 General introduction	327	21.4 A discussion	346
20.2 A simple example	328	21.5 Appendix	347
20.3 Transition probabilities	328	Bibliography	364
20.4 What defines a Markov chain?	329	Glossary	376
20.5 Populations and the transition matrix	330	Index	386
20.6 The calculus of transition probabilities	332	Greek alphabet	395
20.7 Fitting the model	334		

Preface

I find it appropriate to begin the preface with recollections regarding the general state of the Earth process. Thomas Berry (1990) used this term and gave it well-definable meaning. Our text has in its focus manifestations of this process when presenting the concepts and methods of statistical data analysis to those in ecology and related fields.

The critical state of the Earth process and the dangerous course it is running on is no longer a theory whose validity is disputed. Few do not agree that the high volume of greenhouse gases in the atmosphere, mainly from the burning of fossil fuels, is causing global warming; dumping of limitless quantities of chemical pollutants are poisoning the water supply; and vegetation destruction by everyday land use is triggering erosion and desertification over vast tracks of land. The biota responds with extinctions in the extreme, but not in trivial numbers. In fact conservative estimates put the extinction rate at the staggering 10,000 species annually with all forms of life counted. These suggest that one half of all the species now existing will be eradicated before the closing of the 21st century (Willson 1992, 2001).

If one were to rank the deleterious effects by potential, most would probably consider global warming as society's public enemy number one. The Manabe ($2\times\text{CO}_2$) – Mason scenario (Manabe 1990, Mason 1990) is a benchmark prediction reinforced by almost two decades of research advances (IPCC 2001, 2007, Gore 2006, Orlóci 2008). According to this scenario, the Earth's climate will have undergone surface warming by about 2.5 °C on average in about seven decades (about 3.6 °C in the Century) counting from 1990. Mason's expectation of the oceans' thermal inertia to be overcome and atmospheric warming to manifest itself measurably became reality. The Tundra permafrost is melting and the polar Ice is doing the same.

More recent estimates (IPCC 2001, 2007, Gore 2006, Orlóci 2008) of the atmospheric warming rate are much worse. But climate warming at even the early predictions is quite sufficient to force dramatic changes in the World's biota. How dramatic? Consider a potential case from a typical site in the Boreal region near Timmins, Ontario (Orlóci 1994):

<i>Annual mean precipitation mm</i>	<i>Annual mean temperature °C</i>	<i>Thermal flux rate °C</i>	<i>Temperature increase by 2060° C</i>	<i>Expected temperature by 2060° C</i>
711	1.3	3.7	9.2	10.5

Note: thermal flux is defined as the rise in local temperature per one degree rise in the global average temperature.

The necessary outcome of the process if left unchecked is a global disaster of unseen proportions. Recognition of this has led to the enactment of progressive environmental laws by many nations which mandate a coupling of technical planning and environmental protection.

Clearly the statistical approach in sampling and data analysis must come up to par with the new sweeping standards of the mandated, large scale environmental studies. Assessment and prediction are the main tasks. These are concerned with the present state of the environment, its evolutionary past, and anticipated

future. The complexities of implementation place premium on choices that emphasise empiricism, power, and very much, a clear local relevance.

These points were uppermost in our mind when we selected the topics for presentation. We had to go beyond the deceptively simple Fisherian sampling environment (Orlóci 1993,2001b) into the real world whose complexities we know from Poore (1962), Mandelbrot (1967,1977) and Lorenz (1963). In world they describe *process* is in centre in its full natural colours: complex (non-linear), fractal (irregular, fragmented), and chaotic (disorderly, confused).

We begin these notes with the sampling environment, recognising the sharp dichotomy in conceptualisations with Fisherian statistics (FS) in one direction and Poorean successive approximation (PSA) in the other. The defining differences are substantive and should be easily grasped by any with only a minimal exposure to ecological ideas:

- 1) FS assumes an ideal sampling environment of global regularity as if it were ruled by strict experimental controls. PSA does not idealise the sampling environment, but takes it as it comes. Consequently, the constraints on sampling and inference in FS are different from those in PSA.
- 2) PSA allows the statistical conclusion to grow and to come closer and closer in approximation to the truth through recursive sampling and analysis. FS lacks inherent facilities to allow such an evolution of the conclusion by virtue of its idealisation of the sampling environment.
- 3) FS focuses on the "average". PSA defines a role for the "type" and for the "typical" event as well. "Average" and "typical" need not be the same.

In the organization of the book's contents problem-oriented lines are followed, giving due weight for *concepts* and *modus operandi* of both FS and PSA. The text begins with definitions of terms and a discussion of general ideas. Independent chapters treat data management, biological variables and their measurement, population description, sampling, estimation, and sampling distributions. Subsequent chapters cover the methods of comparison (variables, individuals, groups), character weighting (ranking), trend seeking (regression, ordination), and classification (cluster analysis, identification). Reference list, problems, glossary and subject index conclude the book. Numerous step-by-step examples are included. Three external appendices are closely integrated with the main text and serve as a basis of hands-on practice sessions: the APICE exercise book, sample data set, and application programs.

Márta Mihály, Forest Engineer, gave invaluable technical and lectoral assistance throughout the preparation of the book. For these and for her patience I express my sincerest thanks.

László Orlóci

Winter 2010, Kailua, Hawaii