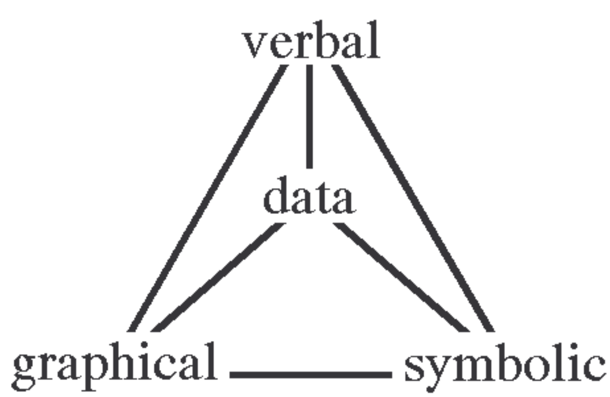
Handouts in Quantitative Biology

D. C. Schneider
Memorial University of Newfoundland
St. John's

September 2004



Part I Units and Dimensions

Table 1. Base and supplementary units in the SI system.

Quantity	Unit	Abbreviation
Length	metre	m
Mass	kilogram	kg
Time	second	s
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd
Electrical current	ampere	A
Planar angle	radian	rad
Solid angle	steradian	sr

Table 2. Standard multiples of ratio scale units.

Name	Multiple	Abbreviation	Example
pico	10^{-12}	p	pW
nano	10^{-9}	n	nW
micro	10^{-6}	μ	μW
milli	10^{-3}	m	mW
centi	10^{-2}	c	cW
deci	10^{-1}	d	dW
	10^{0}		W
deca	10^{1}	da	daW
hecto	10^{2}	h	hW
kilo	10^{3}	k	kW
mega	10^{6}	M	MW
giga	10^{9}	G	GW

Table 3. Units that commonly occur in biology.

Quantity		Unit Name	Unit Symbol	Equivalent Units
Acceleration	angular			$rad \cdot s^{-2}$
	linear			$m \cdot s^{-2}$
Area		square metre	m^2	
		hectare	ha	$10^4 \cdot m^2$
Concentration				$mol \cdot m^{-3}$
Energy (work)		joule	J	N·m
		kilocalorie	kcal	4185·J
Energy flux				$J \cdot m^{-2} \cdot s^{-1}$
Force		newton	N	$kg \cdot m \cdot s^{-2}$
Frequency		hertz	Hz	s^{-1}
Light	Luminance			$cd \cdot m^{-2}$
	Luminous flux	lumen	lm	cd·sr
	Illuminance	lux	lx	$lm \cdot m^{-2}$
		footcandle	fc	10.764·lx
	Photon flux	einstein	E	1·mole
Mass density				$kg \cdot m^{-1}$
Mass flow				$kg \cdot s^{-1}$
Mass flux				$kg \cdot m^{-2} \cdot s^{-1}$
Power		watt	W	$J \cdot s^{-1}$
Pressure (stress)		pascal	Pa	$N \cdot m^{-2}$
Surface tension				$N \cdot m^{-1}$
Velocity	angular			$rad \cdot s^{-1}$
	linear			$m \cdot s^{-1}$
Viscosity	dynamic			Pa·s
	kinematic			$m^2 \cdot s^{-1}$
Volume		cubic metre	m^3	
		litre	l	$10^{-3}m^3$
Volume flow rate				$m^3 \cdot s^{-1}$
Wavelength				m
Wavenumber				m^{-1}

Table 4. Rules for working with dimensions.
From D.S. Riggs (1963) *The Mathematical Approach to Physiological Problems*. MIT Press.

1. All terms in equation must have the same dimensions.
   Terms separated by + − or = .
2. Multiplication and division must be consistent with rule 1.
3. Dimensions are independent of magnitude.
   dx/dt is the ratio of infinitesimals,
   but still has dimensions of x/t = Length/Time.
4. Pure numbers (e, π) have no dimensions.
   Exponents and percentages have no dimensions.
5. Multiplication by a dimensionless number does not change dimensions.

Working with Dimensions--Examples.

1. According to Holligan et al 1984 (*Marine Ecology Progress Series* 17:201) the vertical flux of nutrients through the ocean's thermocline is:

$$F_N = K_V \quad \Delta N / \Delta Z$$

were F_N is the vertical flux of nutrients (milligram-atoms $m^{-2}\ s^{-1}$)
K_V is the vertical eddy diffusivity ($10^{-4}\ m^2\ s^{-1}$)
ΔN is the nitrate difference across the thermocline (mg-atoms)
ΔZ is the thickness of the thermocline (metres)

Write out dimensions beneath each symbol in the equation.
Is this equation dimensionally homogeneous? ____________

Work out the dimensions of ΔN required to make the equation homogeneous _________

Work out the units of ΔN required to make the equation homogeneous _________

M = Mass $\quad$ $M\ L^{-1}$ = mass gradient
$M\ L^{-2}$ = mass density $\quad$ $M\ L^{-3}$ = mass concentration

Based on this, ΔN must be the difference in nitrate ____________ across the thermocline.

2. A series of experimental measurements by Holligan *et al* suggest that the vertical flux of nutrients through the thermocline follows an exponential relation:

$$F_N = \alpha(K_V \Delta N / \Delta Z)^{3/4}$$

What units does α have? ___________________________

What dimensions does α have? ________________________

3. Another series of experiments by Holligan *et al* suggest that nutrient flux depends upon the temperature gradient across the thermocline.

$$F_N = \beta\,(\Delta T/\Delta Z)^{-1/3}$$

$$\Delta T/\Delta Z = {}^{o}\mathrm{C/metre}$$

What units does β have? _________________________

What dimensions does β have? _________________________

Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests:
1 without critical attention to the underlying model involved;
2 without due regard to the precise distribution of sampling errors;
3 with little concern for the scale of measurement;
4 careless of dimensional homogeneity;
5 without considering the ideal transformation;
6 without any attempt at model simplification;
7 with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.

M.J. Crawley. 1993. *GLIM for Ecologists*. (London, Blackwell)

Euclidean and Fractal Dimensions in Biology -- References

Gunther, B. 1975. Dimensional analysis and the theory of biological similarity. *Physiological Reviews* 55: 659-698.

Hastings, H. M. and G. Sugihara. 1993. *Fractals: a User's Guide for the Natural Sciences.* Cambridge University Press.

Mandelbrot, B.B. 1977. *Fractals: Form, Chance, and Dimension.* San Francisco: Freeman.

Pennycuick, C.J. *Newton Rules Biology: A Physical Approach to Biological Problems.* Oxford University Press.

Platt, T.R. and W. Silvert. 1981. Ecology, physiology, allometry, and dimensionality. *Journal of Theoretical Biology* 93: 855-860.

Schneider, D.C. 1994. *Quantitative Ecology: Spatial and Temporal Scaling.* San Diego: Academic Press.

Stahl, W.R. 1961, 1962. Dimensional analysis in mathematical biology. *Bulletin of Mathematical Biophysics* 23: 355-376, 24: 81-108.

Sugihara, G., B. Grenfell, and R.M. May. 1990. Applications of fractals in ecology. *Trends in Resereach in Ecology and Evolution*. 5: 79-87.

<short, highly readable account, including how to estimate km^{d}>

West, B.J. and A.L. Goldberger. 1987. Physiology in fractal dimensions. *American Scientist* 75: 351-365.

Part II. The General Linear Model.

Notation for Frequency Distributions and Probability Functions.

There is no standard notation for frequency distributions and probability functions: the notation will vary from text to text. Here are some notational conventions that tend to be widely used. Equivalent notation is also shown.

An empirical distribution constructed from a sample of size n can be expressed in any of four different ways:

$F(Q = k)$	histogram of values	frequencies
$F(Q = k)/n$	histogram of proportions	relative frequencies
$F(Q \leq k)$	histogram of cumulative values	cumulative frequencies
$F(Q \leq k)/n$	histogram of proportions	cumulative relative frequencies

Theoretical distributions can be either discrete (binomial, Poisson) or continuous (normal, chisquare, F, t). These are functional expressions. The probability density function pdf is a function for the probability, or relative frequency. The cumulative density function cdf is for the cumulative probability, or cumulative frequency. These function can thus be considered models for the frequency distribution obtained from data.

	Observed	Expected	k is discrete	Q is measured
	n = sample	N = population	x is continuous	X is continuous

Frequency	$F(Q = k)$	Frequency of Q in the sample of size n	(the histogram)
	$n{\cdot}Pr(Q \leq k)$	Expected frequency that Q in sample, limited to k values	
	$n{\cdot}Pr(X \leq x)$	Expected frequency X in sample, X continuous	
	$N{\cdot}Pr(Q \leq k)$	Expected frequency that Q in population, k values only	
	$N{\cdot}Pr(X \leq x)$	Expected frequency X in population, X continuous	
Relative Frequency	$F(Q = k)/n$	Proportion of Q in the sample of size n	
	$Pr(Q = k)$	Probability that Q = k	probability mass function, pmf
	$Pr(X=x)$	Probability that X = x	probability density function, pdf
Cumulative Frequency	$F(Q \leq k)$	Cumulative frequency of Q	
	$n{\cdot}Pr(Q \leq k)$	Expected frequency that $Q \leq k$ in sample, limited to k values	
	$n{\cdot}Pr(X \leq x)$	Expected frequency $X \leq x$ in sample, X continuous	
	$N{\cdot}Pr(Q \leq k)$	Expected frequency that $Q \leq k$ in population, k values only	
	$N{\cdot}Pr(X \leq x)$	Expected frequency $X \leq x$ in population, X continuous	
Cum. Relative Frequency	$F(Q \leq k)/n$	Proportion of $Q \leq k$ in the sample of size n	
	$Pr(Q \leq k)$	Probability that $Q \leq k$	cumulative mass function, cmf
	$Pr(X \leq x)$	Probability that $X \leq x$	cumulative density function, cdf

Notation for Frequency Distributions and Probability Functions.

Equivalent notation	$Pr(Q = k)$	$f(x)$	pmf	$P(Q = k)$	for discrete variables
	$Pr(X = x)$	$f(x)$	pdf	$P(X = x)$	for continuous
	$Pr(Q \leq k)$	$F(x)$	cmf	$P(Q \leq k)$	for discrete variables
	$Pr(X \leq x)$	$F(x)$	cdf	$P(X \leq x)$	for continuous

Table 5. Key for choosing the frequency distribution of a statistic.

Statistic is the population mean	
If data are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30	t distribution
If data are Poisson	Poisson distribution
If data are Binomial	Binomial distribution
If data do not cluster around central value, examine residuals (deviations from the mean)	
If residuals are normal or cluster around a central value	
If sample is large (n > 30)	Normal distribution
If sample is small (n < 30)	t distribution
If residuals are not normal	Empirical (bootstrap)
Statistic is the population variance	
If data are normal or cluster around a central value	Chi-square
If data do not cluster around central value	
If sample is large (n > 30)	Chi-square
If sample is small (n < 30	Empirical (bootstrap)
Statistic is the ratio of two variances (ANOVA tables)	
If data are normal or cluster around a central value	F-distribution
If data do not cluster around a central value, calculate residuals	
If residuals are normal or cluster around a central value	F-distribution
If residuals do not cluster around central values	
If sample is large (n > 30)	F-distribution
If sample is small (n < 30)	Empirical
Statistic is none of the above	
Search statistical literature for appropriate distribution or confer with statistician	
If not in literature or cannot be found	Empirical

Empirical distributions are generated by taking all permutations, by sampling permutations, or by subsampling (bootstrap methods).

Table 6. Generic recipe for calculating a confidence limit.

1. State population; state the statistic of interest.
2. Calculate an estimate of the statistic from data
3. Determine the distribution of the estimate.
4. State tolerance for Type I error.
5. Write a probability statement about the estimate or statistic.
6. Plug values into the statement to obtain confidence limits.
7. Make a statement about the probability that the line
   (or limits) include the true value.
   This statement is not about the statistic or estimate.

Strangely, the motto chosen by the founders of the Statistical Society in 1834 was *Aliis exterendum*, which means "Let others thrash it out." William Cochran confessed that "it is a little embarrassing that statisticians started out by proclaiming what they will not do."
E. A. Gehan and N. A. Lemak. 1995. *Statistics in Medical Research: Developments in Clinical Trials* (Plenum Press).

Fisher's famous paper of 1922, which quantified information almost half a century ago, may be taken as the fountainhead from which developed a flow of statistical papers, soon to become a flood. This flood, as most floods, contains flotsam much of which, unfortunately, has come to rest in many text books. Everyone will have his own pet assortment of flotsam; mine include most of the theory of significance testing, including multiple comparison tests, and non parametric statistics.

John Nelder, Rothamsted Experimental Station. (Fisher's successor as Director of the Statistics Department, and pioneer of generalised linear models). From: *Mathematical Models in Ecology*, British Ecological Society Symposium 1971.

Table 7. Generic recipe for decision making with statistics.

1. State population, conditions for taking sample.
2. State the model or measure of pattern . . . . ST
3. State Null Hypothesis about the population . . . . H_o
4. State Alternative Hypothesis . . . . H_a
5. State criterion (tolerance) for Type I error . . . . α
6. State frequency distribution that gives probability of outcomes when the Null Hypothesis is true. Choices are:
   Permutations, i.e. distribution of all possible outcomes when H_o is true;
   Empirical distribution obtained by random sampling of all possible outcomes when H_o is true;
   Cumulative distribution function (cdf) that applies when H_o is true;
   State assumptions when using a cdf such as normal, F, t, or chisquare.
7. Calculate the statistic. This is the observed outcome.
8. Calculate the p-value for the observed outcome relative to distribution of outcomes when H_o is true.
9. If p less than α then reject H_o and accept H_a
   If p greater than α then accept H_o.
10. Report statistic, p-value, sample size.
    Declare decision.

Equivalent method (less informative) based on just a statistical table, no computer

8. Calculate outcome corresponding to α
9. If observed outcome > outcome @ α then reject H_o, accept H_a.
   If observed outcome $\leq$ outcome @ α then accept H_o.
10. Report statistic, p-value, and sample size. Declare decision.

This latter method is less informative, because the observed p-value does not get reported. This method was made necessary by the cumbersome tables for frequency distribution. With modern computers it is possible to calculate an exact p-value for any statistic. The method of reporting an exact p-value is preferred to the method based on tables.

Table 8 Generic Recipe for data analysis with the General Linear Model.

1. Construct model. Begin with verbal and graphical model.
   Distinguish response from explanatory variables
   Assign symbols, state units and type of measurement scale for each.
   Write out statistical model.
2. Execute model — Place data in model format, code model statement.
   Compute fitted values from parameter estimates.
   Compute residuals and plot against fitted values.
3. Evaluate the model, using residuals.
   If straight line inappropriate, revise the model (back to step 1).
   If errors not homogeneous, consider using generalized linear model (step 1)
   If n small, evaluate assumptions for using chisquare, t, or F distribution.
      residuals homogeneous ? (residual versus fit plot)
      residuals independent ? (plot residuals versus residuals at lag 1)
      residuals normal ? (histogram of residuals, quantile or normal score plot)
   If not met, empirical distribution (by randomization) may be necessary
4. State population and whether the sample is representative
5. Decide on mode of inference. Is hypothesis testing appropriate?
   If yes step 6, otherwise, skip to step 10.
6. State H_o/H_A pair (some analyses may require several pairs).
   State test statistic, its distribution (t or F), and tolerance of Type I error.
7. ANOVA: Partition df and SS according to model.
   Table Source, SS, df, MS, F-ratio.
   Type I error (p-value) from distribution(F or t).
8. Recompute p-value if necessary.
   If assumptions not met compute better p-value by randomization if:
   sample small ($n < 30$) and if p near α.
9. Declare decision about model terms: If $p < \alpha$ then reject H_o and accept H_A
   If $p \geq \alpha$ then accept H_o and reject H_A
   Report conclusion with evidence: Either the ANOVA table or
   F-ratio (df1,df2) or t-statistics (df) and p-value (not α) for terms of interest.
10. Report and interpret parameters of biological interest (means, slopes)
    along with one measure of uncertainty (st. error, st. dev., or conf. intervals).
    Use appropriate distribution (step 8) to compute confidence limits.

This is a modification of the Generic Recipe for Hypothesis testing.
The pattern is stated as an equation; the summary statistic is the F-ratio.
The equation links one or more response variables to one or more explanatory variables,
via parameters (means and slopes).
This equation is used to set up the ANOVA table, to partition the degrees of freedom, and to partition the total sum of squares: $SS_{total} = (n-1) * Var(Y) = (n-1) * s^2$

For reports, use the methods section to:
state the critical value α;
state that the residuals were examined for normality, homogeneity, and independence;
state that randomization methods were used to compute Type I error, if assumptions were not met.

Table 9. Commonly used tests, based on the General Linear Model.

Analysis	Response Variable	Explanatory Variable	Interaction?	Comments
t-test	1 ratio	1 nominal	Absent	compares two means
1-way ANOVA	1 ratio	1 nominal	Absent	compares 3 or more means in 1 category
2-way ANOVA	1 ratio	2 nominal	Present	tests for interactive effects compares means in 2 categories, if no interaction
Paired Comparison	1 ratio	2 nominal	Assumed Absent	compares 2 means in 1 category, controlled for 2nd category (blocks or units)
Randomized Blocks	1 ratio	2 nominal	Assumed Absent	compares 3 or more means in 1 category, controlled for 2nd category (blocks or sampling units)
Hierarchical ANOVA	1 ratio	≥2 nominal	Absent	nested comparisons of means
ANCOVA	1 ratio	≥ 1 ratio ≥ 1 nominal	Present	compares two or more slopes
			Absent	compares means, controlled for slopes
Regression	1 ratio	1 ratio	Absent	tests linear relation of response to explanatory
Multiple Regression	1 ratio	≥ ratio	Assumed Absent	tests linear relation to 2 explanatory variables relation expressed as a plane

Correlation (srbx15_7.out)

Thorax length data from Box 15.7 in Sokal and Rohlf (1995), p 594.

```
MTB > read 'a:srbx15_7.dat' c1 c2;
SUBC> nobs = 15.
     15 ROWS READ

MTB > name c1 'ltot' c2 'thor'
MTB > plot c2 c1
         -
    6.40+                                                        *
         -                                                 *
thor     -        *                *     *                     *
         -                         *                *
         -            *                     * *
    5.60+                 *      *
         -
         -
         -
         -
    4.80+
         -     *
         -
         -
         -       *
    4.00+
          --+---------+---------+---------+---------+---------+----ltot
          6.0       7.2       8.4       9.6      10.8      12.0

MTB > plot c1 c2
    12.0+                                                            *
         -                                                 *
ltot     -                                                       *
         -                                              *
         -
    10.0+                                        2
         -
         -                                                *
         -                                               *   *
         -                                          *
     8.0+                                        *
         -
         -                                             *
         -
         - *            *                                     *
     6.0+
         -
          ------+---------+---------+---------+---------+---------+thor
             4.40      4.80      5.20      5.60      6.00      6.40
```

Total length of 15 aphid stem mothers and the mean thorax length of their parthenogenetic offspring.

Judging from these graphs, a linear model of association did not look acceptable. The following models were then investigated by transforming one or both variables, plotting, and examining the plot to see if it was linear (no bowls or arches).

ltot	log(lthor)
log(lot)	lthor
log(ltot)	log(lthor)
ltot	1/lthor
ltot	lthor^3

The last two were a slight improvement over the first three, but none of the plots could be viewed as linear.

Next, try a model based on monotonic relation: thorax length increases monotonically with total length. That is, variables are associated on a rank scale.

```
MTB > rank c1 c3
MTB > rank c2 c4
MTB > name c3 'Rltot'
MTB > name c4 'Rthor'

MTB > plot c3 c4
        -
    15.0+                                                         *
        -                                         *
Rltot   -                                                     *
        -                             *
        -                       *
    10.0+                       *
        -                                     *
        -                                 *
        -                                             *
        -             *
     5.0+         *
        -                 *
        -                                                 *
        - *
        -     *
     0.0+
          ------+---------+---------+---------+---------+---------+Rthor
              2.5       5.0       7.5      10.0      12.5      15.0

 MTB > corr c3 c4

Correlation of Rltot and Rthor = 0.649
```

This is called the Spearman Rank correlation coefficient. It is a measure of monotonic relation. It measures the linear relation between the **ranks** of the variables.

How does this measure of monotonic association compare with a measure of linear association?

```
MTB > corr c1 c2 m1
Correlation of ltot and lthor = 0.650
```

This is the Pearson correlation, a measure of the linear association between the variables. In this example, the measure of linear association turns out to be the same as the measure of monotonic association.

So far 6 different models have been tried, none could be considered acceptable, based on lack of bowls or arches in the residuals (deviations from line), as judged by eye. Perhaps the problem is that the data are heterogeneous. There appears to be a positive relation, but some of the data points do not conform to this relation. In particular, it seems that any thorax length is possible at low total lengths (ltot < 7 micrometer units). Let's assume that something different is happening at low total lengths, and just examine the relation between variables when ltot > 7 micrometer units.

```
MTB > let c1(5) = 0/0
MTB > let c1(5) = 0/0
                   J
*** VALUES OUT OF BOUNDS DURING OPERATION AT J

MTB > let c1(8) = 0/0
MTB > let c1(9) = 0/0
MTB > plot c1 c2
ltot    -
        -
        -                                                         *
        -                                  *
    11.2+                                                   *
        -
        -                            *
        -             *
        -             *
     9.6+
        -                                *
        -
        -          *                  *          *
        -
     8.0+
        -   *
        -            *
        -
          ------+---------+---------+---------+---------+---------+lthor
             5.60      5.76      5.92      6.08      6.24      6.40

         N* = 3
```

This looks acceptably linear.

Now compute Pearson correlation, placing the coefficient into k1 for later use.

```
MTB > corr c1 c2 m1

Correlation of ltot and lthor = 0.664

MTB > copy m1 c3 c4
MTB > let k1 = c3(2)
MTB > print k1
K1       0.663741
```

Next compute t-statistic, with H_o that the true correlation is zero.

```
MTB > let k2 = k1*sqrt((12-2)/(1-k1**2))
MTB > print k2
K2       2.80620
```

Compute p-value from cumulative distribution function, for t distribution.

```
MTB > cdf k2;
SUBC> t 10.
    2.8062    0.9907
MTB > let k3 = (1-.9907)*2
MTB > print k3
K3       0.0186000
```

Note multiplication by 2, the cumulative distribution function yields proportion of outcomes smaller than t = 2.8062, which comes to 99.07% of the outcomes.
The right tail is thus approximately 1 - 0.9907 = 0.93% and both tails together comes to approximately 1.8% (p = 0.0186 exactly).

Summary.
For non-linear (monotonic) model, use ranks. Compute rank correlation.
For linear model (relation described by straight line) use Pearson correlation.

Multivariate Analysis -- References

Cooley, W. W. and P. R. Lohnes (1971). *Multivariate Data Analysis*. Wiley & Sons, New York.

Gittens, R. Canonical Analysis. *Biomathematics* **12**. Springer-Verlag, Berlin.

Ludwig, J. A. and J. F. Reynolds (1988). *Statistical Ecology*. Wiley & Sons, New York.

Kim, J. and C. W. Mueller (1978). *Introduction to Factor Analysis. What it is and How to do it*. Sage Publications, London.

Morrison, D. F. (1976). *Multivariate Statistical Methods*. McGraw-Hill, New York.

Pielou, E. C. (1984). *The Interpretation of Ecological Data*. Wiley & Sons, New York.

Seal, H. L. (1964). *Multivariate Statistical Analysis for Biologists*. Methuen, London.

Van de Geer, J. P. (1971). *Introduction to Multivariate Analysis for the Social Sciences*. W. H. Freeman, San Francisco.

Most statistical packages (such as SAS, BMDP, SYSTAT, SPSS) include references.

> There are aspects of statistics other than its being intellectually difficult that are barriers to learning. For one thing, statistics does not benefit from a glamorous image that motivates students to persist through tedious and frustrating lessons....there are no TV dramas with a good-looking statistician playing the lead, and few mother's chests swell with pride as they introduce their son or daughter as "the statistician."
> C.T. Le and J.R. Boen. 1995. *Health and Numbers: Basic Statistical Methods*. Wiley.

Autocorrelated Data -- References

Box, G. E. P. and G. H. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

<the basic text in time series analysis>

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. John Wiley, New York

<extensive treatment of topic, fairly mathematical>

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

<somewhat mathematical, emphasizes use of randomization tests>

Griffith, D. A. (1987). *Spatial Autocorrelation*. Resource Publications in Geography, American Society of Geographers.

<accessible treatment with examples>

Platt, T. and K. L. Denman (1975). Spectral analysis in ecology. *Annual Review of Ecology and Systematics* **6**: 189-210.

<reviews one technique: analysis in the frequency domain>

Ripley, B. D. (1981). *Spatial Statistics*. Academic Press, London.

<comprehensive coverage of topics, fairly mathematical>

Upton, G. J. and B. Fingleton (1985). *Spatial Data Analysis by Example*. Vol. I. Point Pattern and Quantitative Data. John Wiley & Sons, Chichester.

<highly accessible because of examples; short on conceptual linkages>

Most statistical packages (such as SAS, BMDP, SYSTAT, SPSS) include references.

GLM: Autocorrelated Data (codacf.out)

Cod (*Gadus morhua*) catch data.

Catches from the northwest Atlantic, NAFO division 2J3KL are divided into Canadian offshore, other offshore, and inshore.

$Total_{offshore} = Other + Can_{offshore}$. Catches in tonnes = 10^3 kg.

```
MTB > read 'a:cod.dat' c1-c4;
SUBC> nobs = 30.
MTB > let c5 = c3 - c2
MTB > name c1 'yr' c2 'other' c3 'totoff' c4 'inshore' c5 'canoff'
MTB > plot c4 c1
  160000+      * *
        -
inshore -           *
        -           *
        -              *
  120000+         *
        -                * *                         *
        -                   * *                        *       *
        -                       *                 *      *
        -                                       *
   80000+                        *             *    *     *   *
        -                                    *              *
        -                          * *
        -                                  *
        -
   40000+                             *   *
        -                               *
          +---------+---------+---------+---------+---------+------yr
      1956.0    1962.0    1968.0    1974.0    1980.0    1986.0
```

Are the inshore catches serially correlated?

```
MTB > acf c4
 ACF of inshore
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.816                          XXXXXXXXXXXXXXXXXXXXX
  2   0.636                          XXXXXXXXXXXXXXXXX
  3   0.537                          XXXXXXXXXXXXXX
  4   0.401                          XXXXXXXXXXXX
  5   0.222                          XXXXXXX
  6   0.074                          XXX
  7  -0.069                        XXX
  8  -0.170                      XXXXX
  9  -0.245                    XXXXXXX
 10  -0.299                   XXXXXXXX
 11  -0.360                 XXXXXXXXXX
 12  -0.360                 XXXXXXXXXX
 13  -0.343                 XXXXXXXXXX
 14  -0.335                  XXXXXXXXX
 15  -0.293                   XXXXXXXX
```

Yes. Inshore catches are strongly correlated. r = +0.816 at lag of 1 year. This means that if catches are high in one year, they will be high the year before or the year after. Catches negatively correlated at lag of 11 years (r = −0.36).

What is best model to describe the relation? The two choices are moving average and autoregressive. Moving average means that catch in any one year depends on combined effects of several previous years. Autoregressive means that catch in any one year is related primarily to effects during a fixed time previously.

The shape of the autocorrelation function suggests that this catch is best described as moving average. Check this by computing the partial autocorrelation with PACF command

```
MTB > pacf c4
PACF of inshore
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.816                         XXXXXXXXXXXXXXXXXXXXX
  2  -0.089                       XXX
  3   0.134                         XXXX
  4  -0.183                    XXXXXX
  5  -0.183                    XXXXXX
  6  -0.082                       XXX
  7  -0.160                     XXXXX
  8   0.028                         XX
  9  -0.052                        XX
 10  -0.010                         X
 11  -0.131                      XXXX
 12   0.057                         XX
 13  -0.063                       XXX
 14  -0.054                        XX
 15   0.047                         XX
```

The shape of the partial autocorrelation function also indicates that catch is related to several prior years (moving average) rather than to year at fixed time in past.

Conclusions:

Inshore catches strongly autocorrelated.
A moving average model is best guess for a statistical model.

Next Analysis: Can inshore catches be predicted from offshore catches?

```
MTB > regress c4 1 c5;
SUBC> residuals c8.

The regression equation is
inshore = 95000 - 0.028 canoff

Predictor        Coef        Stdev    t-ratio        p
Constant        95000         7851      12.10    0.000
canoff        -0.0285       0.1338      -0.21    0.833

s = 32914       R-sq = 0.2%      R-sq(adj) = 0.0%

Analysis of Variance

SOURCE       DF           SS           MS         F        p
Regression    1     49014084     49014084      0.05    0.833
Error        28  30333534208   1083340544
Total        29  30382548992

Unusual Observations
Obs.  canoff   inshore       Fit  Stdev.Fit  Residual   St.Resid
  1     4515    159492     94871       7477     64621      2.02R
R denotes an obs. with a large st. resid.
```

Is this model acceptable? Check assumption A, linear relation.

```
MTB > plot c8 c5

C8       -
         -
         -   **
   50000+    *
         -   * *
         -
         -   *2                                    *
         -      *   *                                   2           *
       0+        *            *
         -        *         *        *                    *  *
         -    *  *                                          *
         -  **   *
         -
  -50000+   2
         - *
         -
         -
           +---------+---------+---------+---------+---------+------canoff
           0      25000     50000     75000    100000    125000
```

No bowls or arches, so linear model acceptable.

Next, investigate the assumptions concerning errors.

B1 sum(errors) = 0 ? Yes, because least squares used in regression.

B2 errors independent ?
The catches are strongly autocorrelated, so residuals are also likely to be autocorrelated. If the residuals are autocorrelated, then p-values based on this model will be in error because the residuals won't be independent.

```
MTB > acf c8                                                are residuals autocorrelated?
    ACF of C8
           -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
             +----+----+----+----+----+----+----+----+----+----+
   1   0.815                          XXXXXXXXXXXXXXXXXXXXX
   2   0.636                          XXXXXXXXXXXXXXXXX
   3   0.536                          XXXXXXXXXXXXXX
   4   0.400                          XXXXXXXXXXX
   5   0.218                          XXXXXX
   6   0.067                          XXX
   7  -0.082                        XXX
   8  -0.185                     XXXXXX
   9  -0.262                   XXXXXXXX
  10  -0.318                  XXXXXXXXX
  11  -0.381                XXXXXXXXXXX
  12  -0.381                XXXXXXXXXXX
  13  -0.362                 XXXXXXXXXX
  14  -0.351                 XXXXXXXXXX
  15  -0.303                  XXXXXXXXX
```

The residuals are not independent. p-value cannot be trusted.

```
MTB > differences 1 c4 c6
MTB > name c6 'inshd1'
MTB > print c4 c6

 ROW  inshore  inshd1

   1   159492       *          16    35181   -6467
   2   157286   -2206          17    41213    6032
   3   119363  -37923          18    59939   18726
   4   138511   19148          19    72623   12684
   5   144548    6037          20    81455    8832
   6   131328  -13220          21    85822    4367
   7   110527  -20801          22    96523   10701
   8   110843     316          23    80038  -16485
   9   101859   -8984          24   113049   33011
  10   101037    -822          25   106423   -6626
  11    97224   -3813          26    97721   -8702
  12    76588  -20636          27    79883  -17838
  13    62539  -14049          28    72369   -7514
  14    62052    -487          29    78747    6378
  15    41648  -20404          30   101925   23178
```

To solve the problem take the differences from one year to the next, in the response variable (inshore catch). Taking the difference usually reduces the autocorrelation.

To check this, examine autocorrelation of the differenced variable.

```
MTB > acf c6
ACF of inshd1
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.006                          X
  2  -0.003                          X
  3  -0.048                         XX
  4   0.099                          XXX
  5  -0.034                         XX
  6   0.171                          XXXXX
  7  -0.164                      XXXXX
  8  -0.061                        XXX
  9  -0.081                        XXX
 10   0.064                          XXX
 11  -0.072                        XXX
 12   0.066                          XXX
 13   0.058                          XX
 14   0.037                          XX
 15  -0.152                      XXXXX
```

```
MTB > pacf c6
PACF of inshd1
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1   0.006                          X
  2  -0.003                          X
  3  -0.048                         XX
  4   0.100                          XXX
  5  -0.036                         XX
  6   0.172                          XXXXX
  7  -0.168                      XXXXX
  8  -0.063                        XXX
  9  -0.065                        XXX
 10   0.021                          XX
 11  -0.039                         XX
 12   0.042                          XX
 13   0.129                          XXXX
 14   0.014                          X
 15  -0.144                      XXXXX
```

Autocorrelation in response variable is usually reduced by taking differences.

Now examine whether **change** in the inshore catch (inshore catch after differencing) is related to offshore catch.

```
MTB > regress c6 1 c5;
SUBC> residuals c9.

The regression equation is    inshd1 = - 4333 + 0.0603 canoff

29 cases used 1 cases contain missing values    (1956 lost from analysis)

Predictor        Coef        Stdev    t-ratio        p
Constant        -4333         3798      -1.14    0.264
canoff        0.06033      0.06364       0.95    0.352

s = 15509       R-sq = 3.2%      R-sq(adj) = 0.0%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    1   216159680   216159680      0.90    0.352
Error        27  6493937152   240516192
Total        28  6710096896

Unusual Observations
Obs.  canoff    inshd1       Fit Stdev.Fit  Residual   St.Resid
  3     4676    -37923     -4051      3611    -33872     -2.25R
 24    94457     33011      1366      4559     31645      2.13R
```

Check the residuals for autocorrelation.

```
MTB > acf c9
ACF of C9
          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1  -0.002                          X
  2   0.001                          X
  3  -0.070                        XXX
  4   0.051                          XX
  5  -0.103                       XXXX
  6   0.095                          XXX
  7  -0.224                    XXXXXXX
  8  -0.130                       XXXX
  9  -0.132                       XXXX
 10   0.031                          XX
 11  -0.090                        XXX
 12   0.077                          XXX
 13   0.095                          XXX
 14   0.094                          XXX
 15  -0.094                        XXX
```

Residuals no longer autocorrelated for new model (based on differencing)

Conclusion: When we remove the autocorrelation present in the inshore catch series, we find that the inshore catches are not related to offshore catches.

Numerical Methods. Finding the sample size (srex9_6.out)

Exercise 9.6 from Sokal and Rohlf (1995), page 268

What sample size should be used to be 80% certain of observing a true difference between two means as small as a tenth of a millimeter, at the 5% level of significance?

First compute the error Mean square = 0.2496

This is better estimate than total variance = 25.6819/99 = 0.2594

```
MTB > read 'srex9_5.dat' c1-c5;
SUBC> nobs=20.
MTB > stack c1-c5 c6;
SUBC> subscripts c7.
MTB > name c6 'b_lngth' c7 'gr'
MTB > anova c6 = c7

Analysis of Variance for b_lngth

Source      DF        SS        MS       F      P
gr           4    1.9734    0.4933    1.98  0.104
Error       95   23.7085    0.2496
Total       99   25.6819
```

n = unknown

σ^2 estimated as $s^2 = 0.2496$ (see above)

$\delta = 0.10$ and $\delta^2 = 0.01$

$\nu = a\,(n - 1)$

$\alpha = 5\%$

P = 80%

match cdf computations in Minitab to t-values for example in Box 9.14 page 263

$t_{0.05[\nu]} = 2.642$ in text, for $\nu = 4(20-1) = 76$

$t_{2(1-0.80)[\nu]} = 0.847$ in text, for $\nu = 4(20-1) = 76$

```
MTB > invcdf .01;
SUBC> t 76.
    0.0100   -2.3764
MTB > invcdf .005;
SUBC> t 76.
    0.0050   -2.6421
MTB > invcdf .4;
SUBC> t 76.
    0.4000   -0.2542
MTB > invcdf .2;
SUBC> t 76.
    0.2000   -0.8464
```

use 0.005 and 0.20 for box 9.14

Use 0.005 and 0.20 for box 9.14 therefore use 0.025 and 0.20 for exercise 9.6

Compute k1 = $2(\sigma/\delta)^2$

```
MTB > let k1 = 2*(0.2496)/(0.01)
```

Guess n = 20, hence ν = 2*(20−1) = 38

```
MTB > invcdf 0.025 k2;
SUBC> t 38.
MTB > invcdf 0.2 k3;
SUBC> t 38.
MTB > let k4 = k1*(k2 + k3)**2      ≤ n
MTB > print k1 k2 k3 k4
K1        49.9200
K2        -2.02439
K3        -0.851178
K4        412.782    ≤ n
```

t value stored into k2

t value stored into k3

≤ n in Box 9.14
Both t-values are negative, the sum becomes positive when squared.

```
MTB > invcdf 0.025 k2;
SUBC> t 822.
MTB > invcdf 0.2 k3;
SUBC> t 822.
MTB > let k4 = k1*(k2 + k3)**2
MTB > print k2 k3 k4
K2        -1.96285
K3        -0.842055
K4        392.745      ≤ n
```

Guess n = 412
hence ν = 822

```
MTB > invcdf .025 k2;
SUBC> t 782.
MTB > invcdf .2 k3;
SUBC> t 782.
MTB > let k4 = k1*(k2 + k3)**2
MTB > print k4 k3 k2
K4        392.804       = n
K3        -0.842103
K2        -1.96301
MTB > stop
```

Guess n = 392
hence ν = 782

No change from last iteration

Sample size is n = 392 for stated power and Type I error (= 5%).