Statistical advices for biologists

Script to a Higher Level Course in Data Analysis and Statistics for Students of Biology and Environmental Protection

Werner Ulrich

UMK Toruń
2007
Contents

1. Introduction ................................................................................................................................. 3
2. Planning scientific studies ........................................................................................................... 4
2.1 Choosing the right statistics .................................................................................................... 7
2.2 How to build a model .............................................................................................................. 7
3. Bivariate comparisons .................................................................................................................. 9
3.1 Important distributions ........................................................................................................... 11
4. Bivariate regression techniques .................................................................................................. 16
5. Vectors and Matrices .................................................................................................................. 22
5.1 Vectors ................................................................................................................................... 22
5.2 Matrix algebra ....................................................................................................................... 26
6. Multiple regression ..................................................................................................................... 43
6.1 How to interpret beta-values ................................................................................................. 47
6.2 Advices for using multiple regression ................................................................................... 52
6.3 Path analysis and linear structure models .............................................................................. 53
6.4 Logistic and other regression models ...................................................................................... 56
6.5 Assessing the goodness of fit ............................................................................................... 58
7. Analysis of variance .................................................................................................................... 60
7.1 Advices for using ANOVA ...................................................................................................... 62
7.2 Comparisons a priori and a posteriori ................................................................................... 74
8. Cluster analysis .......................................................................................................................... 75
8.1 Advices for using cluster analysis ......................................................................................... 79
8.2 K-means cluster analysis ....................................................................................................... 80
8.3 Neighbour joining .................................................................................................................. 81
9. Factor analysis ........................................................................................................................... 83
9.1 Advices for using factor analysis .......................................................................................... 88
10. Canonical regression analysis .................................................................................................. 90
10.1 Advices for using canonical regression analysis ................................................................. 91
11. Discriminant analysis ............................................................................................................... 92
11.1 Advices for using discriminant analysis .............................................................................. 95
12. Multidimensional scaling .......................................................................................................... 96
12.1 Advices for using multidimensional scaling ......................................................................... 99
13. Bootstrapping and jackknifing ................................................................................................. 100
13.1 Reshuffling methods ............................................................................................................ 103
13.2 Mantel test and Moran’s I ................................................................................................... 107
14. Markov chains .......................................................................................................................... 109
15. Time series analysis ................................................................................................................... 116
17. Statistics and the Internet ........................................................................................................ 121
18. Links to multivariate techniques ............................................................................................. 122
19. Some mathematical sites ......................................................................................................... 123

Latest update: 04.07.2007
1. Introduction

Our world undergoes a phase of mathematization. Yet 20 years ago in the age before the third industrial, the PC revolution, it was popular among biology students - at least in our part of the world - to answer the question why they study biology with the remark that they don't like mathematics. Today such an answer seems ridiculous. Mathematical and statistical skills become more and more important.

The following text is not a textbook. There is no need to write a textbook again. The internet provides many very good texts for every statistical method and problem and students should consult these web sites (which are given at the end of this script) for detailed information on each of the methods described below. This text is intended as a script to a higher level course in statistics. Its main purpose is to give some practical advises for using multivariate statistical methods.

I rejected to providing real examples for each method although they would be readily at hand. Instead, most methods are explained using data matrices that contain random numbers. I want to show how easy it is to get seemingly significant results from nothing. Multivariate statistical techniques are on the one hand powerful tools that might provide us with a manifold of information about dependencies between variables. They are on the other hand dangerous techniques in the hand of those how are not acquainted to the many sources of errors connected with them.

The text deals with ten major groups of statistical techniques:

- The **analysis of variance** compares means of dependent variables under the influence of interacting independent variables.
- **Bivariate and multiple regression** tries to describe dependent variables from one or more sets of linear algebraic functions made up of independent variables.
- **Cluster analysis** groups a set of entities according to the expression of a set of variables defining these entities.
- **Factor analysis** is designed to condense a larger set of variables or observations that share some qualities into a new and smaller set of artificial variables that are easier to handle and to interpret.
- **Discriminant analysis** is a regression analysis that tries to classify elements of a grouping variables into predefined groups.
- **Canonical regression analysis** tries to relate two variable sets, a set of dependent and a set of independent variables, via a combined principle component and regression analysis.
- **Multidimensional scaling** arranges a set of variables in a n-dimensional space so that the distances of these variables to the axes become minimal.
- **Resampling techniques, bootstrapping** and **jackknifing** are techniques to infer measures of reliability from series of samples.
- **Markov chains** enable us estimating variable states on the basis of transition probabilities.
- **Time series and spectral analysis** tries to detect regularities (autocorrelation) in series of data.
2. Planning scientific studies

Any sound statistical analysis needs appropriate data. **Any application of statistics to inappropriate or even wrong data is as unscientific as sampling data without prior theoretical work and hypothesis building.** Nevertheless both types of making 'science' have quite a long history in biology. One the one hand there is a strong mainly European tradition of natural history (faunistic, floristic, systematic, morphological) where data acquisition and simple description dominate. There is an untranslatable German phrase 'Datenhuberei' for this. Datenhuberei resulted, for instance, in large collections of morphological, floristic, and faunistic data, manifested in mountains of insects in alcohol tubes or insect boxes. In many cases, the scientific value of these data is zero. Another, mainly American tradition is called (also untranslatable) 'Sinnhuberei', the exaggeration of theoretical speculations and sophisticated statistical and mathematical analysis. The result was a huge number of theoretical papers, full of statistics, speculation, and discussion of other papers of the same type, but without foundation in reality.

Science but has to be founded both in good data and in good theory. **The first thing however must be theory.** Because this is still not always clear to students and even scientists I shall give a short introduction to the planning of scientific work. The scheme beside shows a very general model of how to plan a scientific study. It is based on a more specific scheme of Jürgen Bortz (Statistik für Sozialwissenschaftler, Springer 1999) designed for the social sciences. We can identify six main phases of a scientific study. The first phase must always be a detailed analysis of the existing literature, where the problem to be solved is clearly stated. Based on this the second theory building phase consists in the formulation or adaptation of a sound and logical theory. A theory is a set of axioms and hypothesis that are formalized in a technical language. What we need is a test whether our theory is logically sound. We also need models for testing it and criteria for accepting or modifying it. Hence, we must formulate specific hypotheses that follow from our theory. These hypotheses have to be tested by experiments or observations. And we must formulate exactly when to reject our hypotheses.

Next come the planning and data acquisition phases. Planning has to be done with the way of data analysis in mind. Often **pilot studies** are necessary. Data sizes have to be estimated by **power analysis.** The analytical phase consists of setting appropriate significance levels for rejecting the hypothesis and of choosing appropriate test statistics. In this phase many errors occur when choosing tests for which the data set is inappropriate. The data have to be checked whether they met the prerequisites of each test. Therefore, be aware of the possibilities and limitations of the statistical tests you envision prior to data analysis.

The last phase is the decision phase. The researcher has to interpret the results and to decide, whether the hypotheses have passed the tests and the theory seems worth further testing or whether all or some hypotheses must be rejected. In the latter case the theory requires revision or must also be rejected.
Statistical advices

1. Definition of the problem, study of literature
2. Formulating a general theory
3. Logical verification
   - positive
   - negative
4. Formulating criteria for accepting hypotheses
5. Deducing hypotheses
6. Deducing appropriate null models
7. Planning of the study; Experimental design
8. Getting data
   - Data appropriate
      - yes
      - no
9. Setting significance values $\alpha$ for acceptance
10. Choosing the appropriate statistical test
11. $P < \alpha$
    - Study ill designed
    - no
    - yes
12. Criterion for accepting hypotheses fulfilled
    - no
    - yes
13. Theory useful
14. Formulating further tests of the theory

Phases:
- Searching phase
- Theoretical phase
- Planning phase
- Observational phase
- Analytical phase
- Decision phase
2.1 Choosing the right statistics

Choosing the right statistic, the statistical test for which our data set is appropriate, is not easy. The first step in planning a study should involve an idea of what we later want to do with which the data we get. We should already envision certain statistical methods. We have to undertake a power analysis to be sure that we get enough data for our envisioned test.

The scheme beside shows a short, and surely incomplete, map of how to choose a statistical test on the basis of our data structure. A first decision is always whether we want to deal with one, with two or with several variables simultaneously. We have also to decide whether we want to test predefined hypotheses or whether we want to apply methods that just generate new hypotheses. We must never mix hypothesis generating and hypothesis testing with the same data set and the same method.

If we have larger data sets we might divide these data into two (or even more groups) and apply our statistics to one set for hypothesis generating and to the second set for testing. But we must be sure that both subsets are really independent and might form different and independent samples from the population.

Beside the question about the number of variables the scheme leads us to two other important questions. First, what is the metrics of our data, metric, non-metric or mixed? Second, are our data in accordance with the general linear model (GLM) or not? This leads us to the decision whether to apply a parametric or a non-parametric test.

2.2 How to build a model

Why is it necessary to build mathematical models using experimental or observational data? There are several reasons for this. Biology has transformed from natural history (history!) to an explanatory science. It not only tries to describe phenomena in nature it tries to understand causes and relations. For this task we have to structure our observations and to look for relations between them. This is exactly the modelling process: we use the science of structures, mathematics, to uncover hidden patterns and relations. Modelling is therefore more than finding out whether sample means differ or whether we have simple correlations between data. We have to parametrize these relations. But models have many other tasks. First of all, they generate new predictions about nature, predictions that then have to be verified or falsified. This prediction generating feature is of course also a method to verify our model. Secondly, good models allow predictions to be make about the future. This is a main aim for all environmental models. They are designed to predict the future of populations, ecosystems and biodiversity. At last models reduce the chaos in our data and allow by this the development of new theories and concepts.

Models can be classified into certain classes. On one end of a continuum we have verbal models stating more or less precisely relations between a set of variables. These verbal statements may be incorporated into diagrams where the variables are connected by arrows. Then, we have a qualitative model. On the other end, there are explicit mathematical models that formalize relations. These relations may be fully parametrized and then we have a quantitative model that gives quantitative predictions about variable states. At least, models may contain exact parameter values at all stages of computation. We speak of deterministic models because
all future states of the models can in principle be computed by the initial set of values. On the other hand, the model might contain more or less stochastic variables, variables that are driven by random events. In this case, future parameter values are less sure or even chaotic. In this case we speak of stochastic models. This short discussion indicates already what we need to build a model.

- The first step is that we have a theory. Shortly speaking, a theory is a set of hypotheses stated in a formal language. We need hypotheses about nature and the relations between certain variables. Modelling without a priori theoretical reasoning will lead to nothing. Our a priori experience must lead to a selection of variables, so-called drivers, of the model. These drivers might have explicit or stochastic values. They might be parametrized (characterized by explicit values or functions) or not. In the latter case the model itself should assign values or value ranges to these parameters.

- Then, we have to collect the necessary data. These data have to match the requirements of our theory. Making experiments or observations without an explicit theory in mind will very often result in large sets of data without any value because afterwards we (suddenly) notice that one or another important variable had been ignored and not measured or that our method was inappropriate to incorporate the variable values of the model. This latter case occurs very often if we took to less replicates and the variances are too high. Problems also arise if we used different methods for observations and we later notice that these differences make it impossible to compare the data (for instance because they differ in the degree of quantitativeness).

- In a next step we have to confirm assumed relations between these drivers. We might assign qualitative or quantitative relations. If we quantify the relations (for instance from a regression analysis) we parametrize these relations.

- Then, we have to formalize the relations. This is best done by a flow diagram or flow chart. The flow diagram forces us to write each relation and each step of the model explicitly. This step often uncovers smaller or larger errors in our initial model that would have remained undiscovered in a purely verbal model formulation. Making flow diagrams learns us thinking hard!

- The following step is then a technical one. Rewriting our flow diagram into a computer algorithm. For more complex models this should be a done using a common computer language like C++, R, Matlab, Visual Basic, or Fortran, simple models can be written via a spreadsheet program like Excel.

- Our model will generate a set of output variables or whole classes of relations. We have to check these parameters, whether their values are realistic, whether they correctly predict real values and whether they are able to predict the future.

- At the end we have to modify our model in the light of its predictions and variable states.
3. Bivariate comparisons

For simple comparisons of means we use the t-test based on the t-distribution. The non-parametric alternative is the U-test. This test should be preferred at small sample sizes, skewed or multimodal distributions of errors and if we have outliers. However, many tied ranks also distort error levels of the U-test (Figs. 3.1, 3.2). In this case permutation tests should be preferred. This is shown in the next *Past* example. We have two samples and want to compare the means. The t-test rejects the hypothesis of a significant difference at $p > 0.05$. The U-test gives a similar result. However, we have a small sample size and tied ranks. A permutation test that reshuffles all values across both samples points to a significant difference.
Two distributions are compared with a $\chi^2$ test. A special important usage of the $\chi^2$ test is its application for comparing the outcomes in a twofold grouped variable. Assume we undertake a genetic study. We test 1000 fruit flies. We find 110 times that an allele A that occurred 200 times was combined with the occurrence of curled hairs. In turn, curled hairs were 600 times associated with allele B. Does allele A influence the frequency of curled hairs? We have only data from one population and single data points. We have no measure of variance. A t-test can’t be applied. Nevertheless we are able to test our hypothesis that the frequencies changed. We make the following table (Fig. 3.3), a 2x2 contingency table. It is the basis for a special case of the $\chi^2$ test, the 2x2 contingency $\chi^2$ test. We compute squares of differences, that means variances. These are $\chi^2$ distributed.

With the table we can compute expected probabilities. The total frequency of curled hairs in the population is 585/1000. Allele A has a total frequency of 0.2. Hence we expect 0.2*585 = 117 flies with allele A and curled hairs. We the same logic we get 0.8*585 = 468 flies with curled hairs and allele B. The two other fly numbers come from A and normal hairs = 200-117 = 83 and B and normal hairs = 800-468 = 332.

Now we have all to apply an ordinary $\chi^2$ test. We have 4 data points with associated expectancies. Because we computed the expected frequencies from four data points the number of degrees of freedom is only one. $df = 1$. We compute by hand

$$\chi^2 = \frac{(110-0.2*585)^2}{0.2*585} + \frac{(475-0.8*585)^2}{0.8*585} + \frac{(90-83)^2}{83} + \frac{(325-332)^2}{332} = 1.26$$

Excel gives =ROZKLAD.CHI(1.26,1) = 0.2614. This value (and the $\chi^2$ test value) are identical to the one of Statistica ($p = 0.2614$). Our interpretation is that A and B lead to the same frequencies of curled hairs. Below is shown how to compute the test using Excel.
3.1 Important distributions

Elementary statistical tests are based on certain statistical distributions. These can often be approximated by the normal distribution. In the first statistic course we dealt with certain discrete distributions (Bernoulli, hypergeometric, Pascal, Poisson) and with the continuous normal (and lognormal) distribution.

Now we deal with other important continuous distributions. In Fig. 3.1 the factorials are marked with black quadrates. We see that for \( x = 1 \) the function value becomes 1 for \( x = 2 \) also 1, for \( x = 3 \) \( y = 2! \), for \( x = 4 \) \( y = 3! \), for \( x = 5 \) \( y = 4! \) and so on. The factorial is a discrete function. What is if we would try to define \( 3.5! \)? We would have to generalize the factorial. This generalization is done by the **Gamma function** introduced by the great Swiss mathematician Leonhard Euler (1707-1783, his *Introductio in analysin infinitorum* founded the analysis as an independent part of mathematics).

For all positive real numbers \( x \) is \( \Gamma(x) \) defined as

\[
\Gamma(x+1) = x \Gamma(x)
\]  

(3.1.1)

Hence \( \Gamma(3.5) = 2.5 \Gamma(2.5) = 2.5 \times 1.5 \Gamma(1.5) = 2.5 \times 1.5 \times 0.5 \Gamma(0.5) \).

For natural \( x \) holds therefore

\( \Gamma(x+1) = x! \)

\( \Gamma(1) = \Gamma(2) = 1 \)

But we have no possibility to compute \( \Gamma(1.5) \). Euler found a solution and defined \( \Gamma(x) \) by the **Euler integral**

\[
\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt; \ (x>0)
\]

(3.1.2)

From this we get

\( \Gamma(1) = \int_0^\infty e^{-t} dt = 1 \)

and

\( \Gamma(2) = \int_0^\infty te^{-t} dt = 1 \)

(3.1.3)

The latter function is our already known Euler distribution.

Additionally holds

\( \Gamma(x) \Gamma(1-x) = \pi / \sin(\pi x) \)
and therefore for $z = 0.5$

$$\Gamma(1/2) = \pi^{1/2}$$

The gamma function defines an important distribution, the gamma distribution defined by

$$f(x, \alpha) = \frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)}$$  \hspace{1cm} (3.1.4)

With this function it is easy to compute the gamma function. Excel computes $f(x, \alpha)$. Hence to compute $\Gamma(\alpha)$ you transform

$$\Gamma(\alpha) = \exp(-x)\frac{x^{\alpha-1}}{f(x, \alpha)}.$$

Fig. 3.1.2 shows an example of the gamma distribution $f(x, 3)$. The Fig also shows one reason for the importance of this distribution. It is frequently applied to model skewed natural distributions because the model parameter $\alpha$ can be appropriately adjusted.

Another reason for the importance is the fact that via the gamma function discrete probability distributions that contain factorials can be transformed into continuous ones.

For instance the negative binomial is often written in the form

$$f(x, k) = \frac{\Gamma(x + k - 1)}{\Gamma(x)\Gamma(k - 1)} p^k (1 - p)^x$$

One special case of the gamma distribution is the $\chi^2$-distribution. Assume we have $n$ independent random variates. All are $Z$-transformed and have the mean zero and the variance one. According to the central limit theorem the sum of these variates should be asymptotically normally distributed with a mean of zero (the sum of all means) and a variance of $n$. In effect we add up means. What is if we add up variances? Simplified we compute

$$\chi^2 = \sum_{i=1}^n X_i^2$$  \hspace{1cm} (3.1.5)

This is the $\chi^2$-distribution introduced by the German mathematician R. F. Helmert (1841-1917). Its density function is

$$f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} e^{-x/2} x^{(n-2)/2}$$  \hspace{1cm} (3.1.6)

As the gamma distribution the $\chi^2$-distribution is not a single distribution. It denotes a whole class of distributions depending on the number of random variables $n$. $n$ is called the degree of freedom of the distribution.

$\chi^2$ distributions are additive that means

$$\chi^2_n = \chi^2_1 + \chi^2_{n-1}$$
For large \( n \) \( \chi^2 \) approaches again a normal distribution. If we add up the variances (all having the value one) the mean, the variance, and the skewness of the new \( \chi^2 \) random variate has

\[
\mu = n \\
\sigma^2 = 2n \\
\gamma = \frac{8}{n} 
\]

(3.1.7)

For large \( n \) we can transform a \( \chi^2 \)-distribution into a standard normal distribution by applying the \( Z \)-transformation

\[
z = \frac{\chi^2 - n}{\sqrt{2n}}
\]

(3.1.8)

The next important distribution we have to deal with is the student or t-distribution introduced 1908 by the British mathematician William Gosset (1876-1937) under the pseudonym 'student'. Assume a normal distributed random variate \( Z \) with \( \mu = 0 \) and \( \sigma^2 = 1 \). Additionally we have a \( \chi^2 \)-distribution with \( n \) degrees of freedom. Gosset now defined a t value of

\[
t_n = \frac{z}{\sqrt{\chi^2/n}}
\]

(3.1.9)

The density function of the t-distribution is very complicated and not interesting for us. Interesting instead is that the t-distribution has a mean and a variance of

\[
\mu = 0 \\
\sigma^2 = \frac{n}{n-2}
\]

with \( n \) being again the number of degrees of freedom. Unfortunately Excel only computes cumulative \( \chi^2 \)- and t-distributions. However every statistic package has a probability calculator and gives the appropriate density function values.

If we have two \( \chi^2 \)-distributions with \( n_1 \) and \( n_2 \) degrees of freedom we define

\[
F_{n_1,n_2} = \frac{n_2 \chi^2_1}{n_1 \chi^2_2}
\]

(3.1.10)

as the F-distribution (introduced by the British biostatistician Sir Ronald Fisher (1890-1962). For large \( n \) this distribution again approximates a normal distribution. t-, \( \chi^2 \)-, and F- distributions are closely connected.

\[
t_n^2 = n \frac{z^2}{\chi^2_n} = \frac{n \chi^2_1}{\chi^2_n} = F_{1,n}
\]

(3.1.11)

The last important distribution we deal with is the Weibull distribution (after the Swedish mathematician Waloddi Weibull, 1887-1979)
We get the cumulative density distribution from the integral

\[ F(\alpha, \beta) = 1 - e^{-\alpha x^\beta} \]  

(3.1.13)

For \( \beta = 1 \) the Weibull distributions equals a simple exponential function. For \( \beta = 3 \) the distributions approximates (but not equals) a normal, for larger \( \beta \) the distribution becomes more and more left skewed.

The Weibull distribution is particularly used in the analysis of life expectancies and mortality rates.

We simply model the mortality rate \( m \) using a general power function model

\[ m(t) = m_0 t^{\beta - 1} \]

Using \( S(t) \) as the distribution of survival and modelling this via an exponential model under the assumption that the mortality rate is constant we get

\[ S(t) = e^{-m_0 t^\beta} \]

For this usage the Weibull distribution is rewritten in a two parametric from

\[ f(\beta) = \frac{\beta t}{T} \left( \frac{T}{t} \right)^{\beta - 1} e^{-\left( \frac{T}{t} \right)^{\beta}} \]  

(3.1.14)

where \( T \) denotes the characteristic life expectancy and \( t \) the age. \( f(\beta) \) gives then the probability that a given person will die at age \( t \). \( T \) is the age at which 63.2% of the population already died. We get \( T \) from eq. 3.1.13 with \( \alpha = 1 \) by setting \( t = T \).

\[ F(1, \beta) = 1 - e^{-1} = 1 - e^{-1} = 1 - \frac{1}{e} = 0.632 \]

Having now data on age specific mortality rates \( f(\beta) \) we can estimate the characteristic life expectancy \( T \) and the shape parameter \( \beta \). The parameterized model then allows for calculation survival and mortality rates and associated demographic variables at any given time \( t \) (Fig. 3.1.4). The Fig. shows the cumulative mortality rates in dependence of time for \( T = 100 \) and different \( \beta \) using eq. 3.1.13.

Having data on mortality rates we can estimate the characteristic life time \( T \) from eq. 3.1.13. We use a double log transformation
Using the cumulative mortality rates of the first tables we obtain \( b \) from the slope of a plot of \( \ln[\ln(1-F)] \) against \( \ln(t) \) (Fig. 5.3) We get a slope of 1.20, typical for many insects that have an exponential mortality-time distribution. The intercept \( b \) is

\[
b = -\beta \ln(T) \rightarrow T = e^{-\frac{-0.89}{1.2}} = 2.09
\]

This is the characteristic life expectancy. Interpolating the first second column of the initial table give for 630 individuals to have died a very similar result around two years.
4. Bivariate regression

If we have two variables we often want to infer whether both variables are related. If these variables are metrically scaled we can apply a regression. In the case of ordinary or nominal scales some association index might be applied. Association indices are dealt with below. A regression is most simply done using Excel or other spreadsheet or statistics programs. You have a x and a y variable and plot x versus y. Excel automatically gives the associated linear or non-linear (power, logarithmic, exponential, or algebraic) function and the coefficient of determination R². To calculate the regression equation Excel and other programs use ordinary least squares regression (Fig. 4.1). The sum of all squared distances $\Delta y$ is minimized:

$$D = \sum_{i=1}^{n} (\Delta y)^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

(4.1)

Hence we have to solve

$$\frac{\delta D}{\delta a} = -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b) = 0$$

$$\frac{\delta D}{\delta b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

(4.2)

and get

$$a = \frac{\sigma_{xy}}{\sigma_x^2}$$

(4.3)

with $\sigma_{xy}$ being the covariance of x an y.

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}$$

(4.4)

The coefficient of correlation is then defined as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

(4.5)

This is well known and needs no further explanation. However, at this stage most errors start. First we have to make clear what we really did. Did we have any well founded hypothesis about the relationships between X and Y. There are several possibilities:

- A change in x causes respective changes in y. Under this hypothesis x is the independent and y the dependent variable.
A change in $y$ causes respective changes in $x$ and we have to invert our relationship.

- $x$ and $y$ are mutually related.
- $x$ and $y$ are not directly related and the observed pattern is caused by a third or more hidden factors that influence both, $x$ and $y$. In this case there is no causal relationship between $x$ and $y$. We speak of a pseudo-correlation.
- $x$ and $y$ are neither directly nor indirectly related. The observed pattern is produced accidentally.

Hence, we have to decide whether $x$ is really independent and whether $y$ is really dependent on $x$. Further we first have to establish whether the prerequisites of linear least squares regression are met. Often this relationship is not so clear as it seems. Look at Figures 4.2 to 4.4. In the first case I plotted numbers of predator species against numbers of prey species. With some justification we can assume that predators dependent on prey although it is known from ecological modelling that there are mutualistic relationships and prey can also be seen as being dependent on predators. Anyway the relationship appears to be linear and we can apply ordinary least squares.

This is not the case in Fig. 4.3. Of course brain weight depends on body weight but this relationship is best described by a power function. Hence we have a non-linear relationship. Now we have two possibilities. We might linearize the power function

\[ W = aB^z \rightarrow \ln W = \ln a + z \ln B \]

\[ \downarrow \]

\[ \ln W = \ln a + z \ln B + \varepsilon \]

\[ \downarrow \]

\[ W = aB^z \exp(\varepsilon) \]

(4.6)

with $\varepsilon$ being the error to be minimized by the regression.

Hence our model to fit contains the error term $\varepsilon$ in a multiplicative form. In other words, we assume that errors are lognormally distributed around the means and stem from multiplicative processes. This has to be established before applying this type of regression.

As an alternative we might apply non-linear regression. Most statistics packages contain non-linear re-
gression modules. They also use least squares but compute them directly from our non-linear function we intend to fit. In this case we apply the following model

\[
W = aB^2 \\
\downarrow \\
W = aB^2 + c
\]

Now we treat the errors as being normally distributed around the mean and result from additive error generating processes. Both models make different assumptions about the underlying processes and predict different regression functions. In Figs. 4.5 and 4.6 I computed exponential and power function distributed values using one time an additive and one time a multiplicative error term. In Fig. 4.5 I used \( Y = e^{0.1X} + \text{norm}(0, Y) \), where norm was a normally distributed random variate. A fit by the nonlinear estimation module of Statistica (the red trend line) gave a result quite close to the model parameters. The Excel build in fitting function, in turn, predicted a too high intercept and a too low slope. This function uses log-transformation (\( \ln Y = aX + \varepsilon \)) and assumes therefore a different distribution of errors. The log transformation reduces the absolute differences between the Y values. Hence the largest Y values have less influence on the regression. The log transformation gives more weight to smaller values. In Fig. 4.6 I used a multiplicative error (\( Y = X^{0.5}e^{\text{norm}(0, Y)} \)). The non-linear estimation predicts a too low slope and a too high intercept. However, the Excel fit, although using log-transformed values and therefore the correct error distribution, predict a too high slope and a too low intercept. In general the nonlinear estimation appears to be closer to the generating parameter values. Further we see that in this latter case the larger X values have a higher influence on the non-linear fit and force the regression line to a lower slope.

Next look at Fig. 4.4. In this case we compared a plantation and a field. There are no clear dependent and independent variables. Further, both variables have error terms, whereas ordinary least squares regression (OLR) assumes only the dependent variable of have errors. Hence important prerequisites of OLR are not met. If the errors in measuring x would be large our least square regression might give a quite misleading result. This problem has long been known and one method to overcome this problem is surely to use Spearman’s rank order correlation. However we may also use a different method for minimizing the deviations of data points from the assumed regression line. Look
Statistical advices

at Fig. 4.7. Ordinary least square regression depends on the variability of $y$ but assumes $x$ to be without error ($\text{OLR}_y$). But why not using OLR based on the variability of $x$ only ($\text{OLR}_x$)? The slope $a$ of $\text{OLR}_x$ is (why?)

$$a_{\text{OLR}_x} = \frac{s_x^2}{s_{xy}}$$  \hspace{2cm} (4.8)

Now suppose we compute $\text{OLE}_y$ and $\text{OLE}_x$ separately. Both slopes differ. One depends solely on the errors of $y$, the other solely on the errors of $x$.

We are tempted to use the mean value of both slopes as a better approximation of the ‘real’ slope. We use the geometric mean of both slopes and get.

$$a_{\text{RMA}} = \sqrt{a_{\text{OLR}_y} a_{\text{OLR}_x}} = \sqrt{\frac{s_x^2}{s_{xy}} \frac{s_y^2}{s_{xy}} } = \frac{s_y}{s_x}$$ \hspace{2cm} (4.9)

In effect this regression reduces the diagonal distances of the rectangle made up by $\Delta y$ and $\Delta x$ as shown in Fig. 4.7. Due to the use of the geometric mean it is called the geometric mean regression or the reduced major axis regression (also standard major axis regression). Due to the use of standard deviations the method uses standardized data instead of raw data. Eq. 4.9 can be written in a slightly different form. We get (why?)

$$a_{\text{RMA}} = \frac{s_y}{s_x} = \frac{a_{\text{OLR}_y}}{r}$$ \hspace{2cm} (4.10)

The RMA regression slope is therefore always larger than the ordinary least square regression slope $a_{\text{OLR}_y}$. Hence, if an ordinary least square regression is significant the RMA regression is significant too.

Lastly, the most natural way of distance and minimizing distances to reach in a regression slope is to use the Euclidean distance (Fig. 4.9). If we use this concept we again consider the errors of $x$ and those of $y$. This technique is called major axis regression (MAR). However, in this case the mathematical background becomes already quite complicated. The MAR slope becomes

$$a_{\text{MAR}} = \frac{s_y}{\lambda - s_y^2}$$ \hspace{2cm} (4.11)

with $\lambda$ being
OLRx and OLRy are termed the **model I regression**. RMA and MAR are also termed **model II regression**.

Of course, the intercept of all these regression techniques is further computed from the means of X and Y

\[ b = \bar{y} - ax \]  

(4.13)

Figure 4.10 shows a comparison of all four regression techniques. We see that at a high correlation coefficient of 0.77 the differences in slope values between all four techniques are quite small. Larger differences
but would occur at lower correlation coefficients. The Excel model shows also how to compute a model II regression. They are not implemented into Excel and major statistic packages although they are recently very popular among biologists. The program PAST is recently the only common package that computes default RMA slopes (Fig. 4.9). Figs. 4.10 and 4.11 show how the four regression models behave if we introduce outliers. We see that OLRx and MAR react strong on a single outlier whereas RAM and OLRy behave more moderate. RMA appears to be least affected by the outlier while in all cases giving reasonable regression lines.

This leads us to the question of model choice (Fig. 4.12). When to use which type of regression? In general model II regression should be used when we have no clear hypothesis what is the dependent and what the independent variable. The reason is clear. If we don’t know what is x and y we also can’t decide which errors to leave out. Hence we have to consider both errors, those of x and those of y. If we clearly suspect one variable to be independent we also should use model II regression if this variable has large measurement errors. As rule of thumb for model I regression the errors of x should be smaller than about 20% of the errors in y. Hence $s_y < 0.2s_x$. Lastly, if the we intend to use a model II regression, we should use MAR if our data are of the same dimension (of the same type, for instance weight and weight, mol and mol, length and length and so on). Otherwise a RMA regression is indicated.

However, all these advices have only importance if we deal with loosely correlated data. For variables having a correlation coefficient above 0.9 the differences in slope become quite small and we can safely use an ordinary least square regression OLRy. Additionally, there is an ongoing discussion about the interpretation of RMA. RMA deals only with variances. The slope term does not contain any interaction of the variables (it lacks the covariance). Hence, how can RMA describe a regression of $y$ on $x$? Nevertheless especially RMA and MAR have become very popular when dealing with variables for which common trends have to be estimated.
5. Vectors and Matrices

5.1 Vectors

Given a point in space we can shift this point to another place. The arrow that goes from the original point to its new place is called a vector. In Fig. 5.1.1 we have three vectors \( A, \) and \( B, \) and a vector \( C \) that point back to itself. This is called a null vector. In a Cartesian system vectors are given by the coordinates of the endpoint minus the coordinates of the starting point. Hence in Fig. 5.1.1

\[
A = \begin{pmatrix} 20 - 5 \\ 20 - 10 \end{pmatrix} = \begin{pmatrix} 15 \\ 10 \end{pmatrix}
\]

Hence all vectors with identical \( x \) and \( y \) values are identical. Further

\[
B = \begin{pmatrix} 5 - 20 \\ 2 - 12 \end{pmatrix} = \begin{pmatrix} -15 \\ -10 \end{pmatrix} = -1 \begin{pmatrix} 15 \\ 10 \end{pmatrix} = -A
\]

The vector \( B \) is parallel to \( A \) but points in the opposite direction. That means \( B = -A. \) In Fig. 5.1.2 the vectors \( i \) and \( j \) are given by

\[
i = \begin{pmatrix} 1 \\ 0 \end{pmatrix} ; \quad j = \begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]

The vector \( A \) can be seen as the multiplication of a number \( a_i \) with \( i \) and \( a_j \) with \( j. \) In vector algebra numbers are called scalars. Hence

\[
A = \begin{pmatrix} a_i \cdot i \\ a_j \cdot j \end{pmatrix}
\]

We call the scalars \( a_i \) and \( a_j \) the coordinates of the vectors \( A. \) The null vector is therefore defined by \( o = \{0,0\} \) and the unity vectors are \( I = \{1,0\} \) and \( J = \{0,1\}. \)

We can define vectors \( I \) more than two dimensions. The general form of a vector in an \( n \)-dimensional space is

\[
V = \begin{pmatrix} a_i \\ \vdots \\ a_n \end{pmatrix}
\]

The examples above provide a natural introduction to basic vector operations. The addition and subtraction of vectors are defined as
In two-dimensional space this can be interpreted as generating a parallelogram from the vectors \( A \) and \( B \) that has the longer diagonal of \( C \) (Fig. 5.1.3). Consequently a subtraction \( A - B \) is defined as an addition of the antivector of \(-B\) and \( A \) (Fig. 5.1.4).

Both definitions of course hold for additional dimensions too.

Basic theorems for addition and subtractions hold for vectors too. Both operations are commutative, and associative. Hence

\[
A + B = B + A
\]
\[
A + (B + C) = (A + B) + C
\]
\[
A + o = A
\]
\[
A - A = o
\]

An addition of \( A \) with the null vectors gives \( A \) and \( A - A \) gives the null vector \( o \).

Next we define the multiplication. There are three types of vector multiplication. The first is the multiplication with a scalar, the **scalar multiplication (or S-product)**. Fig. 5.1.5 shows a natural definition of the S-product. We get

\[
\lambda A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \lambda a_1 \\ \vdots \\ \lambda a_n \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}
\]

The S-multiplication if commutative and distributive (Fig. 5.1.5).

\[
\gamma (\lambda A) = (\gamma \lambda) A
\]
\[
\lambda (A + B) = \lambda B + \lambda A
\]
\[
(\lambda + \gamma) A = \lambda A + \gamma A
\]
\[
oA = o
\]
The length of a vector is of course given by the law of Pythagoras

\[ |A| = \sqrt{a_1^2 + \ldots + a_n^2} \]  

(5.1.6)

This definition implies that the length of \(|A + B| \leq |A| + |B|\).

We can multiply two vectors to get a scalar. This is called the **scalar product** and is defined as

\[
\begin{pmatrix}
a_1 \\ 
\vdots \\ 
a_n 
\end{pmatrix}
\cdot
\begin{pmatrix}
b_1 \\ 
\vdots \\ 
b_n 
\end{pmatrix} = a_1b_1 + \ldots + a_nb_n = \sum_{i=1}^{n} a_ib_i
\]  

(5.1.7)

The commutative, distributive and the mixed associative laws hold

\[
A \cdot B = B \cdot A \\
A \cdot (B + C) = (A \cdot B) + (A \cdot C) \\
(\lambda A) \cdot B = \lambda (A \cdot B) \\
A \cdot 0 = 0
\]  

(5.1.7)

The simple associative law does not hold \(A \cdot (B \cdot C) \neq (A \cdot B) \cdot C\) because ne time we get a vector in direction \(A\) and the other a vector in direction \(C\).

Improtant is the case when we multiply two vectors that are perpendicular. Two dimensional perpendicular vectors have the following structure (Fig. 5.1.6)

\[
A = \begin{pmatrix} x \\ y \end{pmatrix}; B = \begin{pmatrix} \lambda y \\ -\lambda x \end{pmatrix} \rightarrow A \cdot B = \lambda xy - \lambda xy = 0
\]  

(5.1.8)

Hence if the scalar product of two non-sero vectors is zero they are perpendicular. Important is also that the equation \(A \cdot X = b\) has an indefinite number of solutions. **Further the division through a vector**

is not defined.

The scalar product has a simple geometrical interpretation. It is the product of the length of \(A\) with the perpendicular projection \(c_1\) of \(B\) on \(A\) (Fig. 5.1.7)(why?). Because of \(c_1 / |B| = \cos(\alpha)\) we get

\[
A \cdot B = |A| \cdot |B| \cos(\alpha)
\]  

(5.1.9)

If \(\alpha = \pi/2\) \(A \cdot B = 0\). Further we get \(A \cdot A = |A|^2\).

The use of vectors allow for some easy proofs of geometrical theorems.

For instance the cosine theorem can be derived
using a triangle made of three vectors \( C = A - B \) and \( A \cdot B = |B| |B| \cos(\gamma) \). Hence

\[
C^2 = c^2 = (A - B)^2 = A^2 - 2AB + B^2 = c^2 = a^2 + b^2 - 2ab \cos(\lambda)
\]

For \( \gamma = \pi/2 \) we get the theorem of Pythagoras.

Important fields to use vectors are trigonometry and analytical geometry. Using vectors of unit length we can define an angle \( \alpha \) from

\[
E_1 \cdot E_2 = |E_1||E_2| \cos(\alpha) = \cos(\alpha)
\]

Further we can define projections of geometrical objects. For instance a parallel shift of a length \( A \) by a vector \( V \) (Fig. 5.1.9) is given by

\[
A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \rightarrow A' = a_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + v_1 \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} a_1 + v_1 \\ a_2 + v_2 \end{pmatrix}
\]

Using vectors we can define straight lines in space (Fig. 5.1.10). The line \( A \) to point \( P \) is given by

\[
A = r_1 + \lambda (r_2 - r_1)
\]  
(5.1.10)

This equation makes is easy to calculate geometrical relationships. For instance do the straight lines defined by the points \( A_1 = \{1,2,3\} \) and \( A_2 = \{4,5,6\} \) and \( B_1 = \{2,3,4\} \) and \( B_2 = \{5,4,3\} \) cross? We compute

\[
A = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \lambda \begin{pmatrix} 4 \\ -1 \\ 5 \end{pmatrix} = B = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} + \gamma \begin{pmatrix} 5 \\ -2 \\ 4 \end{pmatrix}
\]

\[
\begin{pmatrix} \lambda \\ 3 \\ -\gamma \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}
\]

and see that \( \gamma = 0 \) and \( \lambda = 1/3 \) fulfil this equation. Both straight lines cross in the point \( P = \{2,3,4\} \).

If we have two point \( A \) and \( B \) on a straight line and the direction vector \( U \) the straight line is defined by the vector \( P \) from \( A \) to \( B \) by

\[
P = \lambda u
\]
5.2 Matrix algebra

Biological data bases are most often structured in form of a matrix. Typical examples are our spreadsheet matrices, for instance using Excel, Access, or Matlab (short for Matrix laboratory). The Excel examples below show typical biological data sets. Species are in rows and these are described by a set of nominally, ordinally or metrically scaled variables (descriptors). In the first matrix below we have species at four sites and the values are total catches. In ecology we often have only data about the absence or presence of a certain species. In this case we deal with presence absence matrices and presences are coded with a 1 and absence with a 0.

In general we write matrices in form of rows and columns (Fig. 5.1). An important special case is a matrix that has the same number of rows and columns. This is a square matrix. Matrices with only one row or one column (row or column matrices) are vectors. Hence matrices can be seen as being composed of several vectors.

There are several types of matrices that have

<table>
<thead>
<tr>
<th>Objects</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Descriptors

\[
A = \begin{pmatrix}
  a_{11} & \ldots & a_{1n} \\
  \vdots & \ddots & \vdots \\
  a_{m1} & \ldots & a_{mn}
\end{pmatrix}
\]

Fig. 5.1

\[
V = \begin{pmatrix}
  a_{11} & a_{12} & a_{13} \\
  a_{21} & a_{22} & a_{23} \\
  a_{31} & a_{32} & a_{33}
\end{pmatrix}
\]

\[
V = \begin{pmatrix}
  a_1 \\
  a_2 \\
  a_3 \\
  a_4
\end{pmatrix}
\]
Statistical advices

special properties. Let’s first look again to our species x sites matrix. We want to infer whether the site abundances and species occurrences per site are related. In order to do so we use different measures of association (distance).

Distance measures can operate on presence absence matrices and count site overlap or operate on metrically scaled variables. The simplest count measure is the well known **Soerensen index** of species overlap (= Czekanowski index ≈ Jaccard measure):

\[
D = \frac{2S_{jk}}{S_j + S_k}
\]

(5.2.1)

where \(S_j\) and \(S_k\) are the species number of site \(j\) and \(k\) and \(S_{jk}\) is the number of shared species.

The general metrically based distance measure that includes the **Manhattan** (= taxi driver) distance \((z = 1)\) and the **Euclidean distance** \((z = 2)\) is the **Minkowski metric**:

\[
D = \sqrt{\sum_{i=1}^{n} (x_{i,j} + x_{i,k})^z}
\]

(5.2.2)

Other important distance measures are the **Bray - Curtis** (Czekanowski) metric

\[
D = \frac{\sum_{i=1}^{n} |x_{i,j} - x_{i,k}|}{\sum_{i=1}^{n} x_{i,j} + \sum_{i=1}^{n} x_{i,k}}
\]

(5.2.3)

This metric can be used on raw and ranked data. In ecology the **index of proportional similarity of Colwell and Futuyma** is often used

\[
I_{jk} = 1 - 0.5 \sum_{i=1}^{n} |p_{i,j} - p_{i,k}|
\]

(5.2.4)

where \(p_{i,j}\) and \(p_{i,k}\) denote the frequencies \(p_i = N_i / N_{total}\) of species \(i\) at sites \(j\) and \(k\).

Finally, the Pearson and Spearman correlation coefficients provide measures of distance. This is shown in Fig. 5.2.1. Data from four sites are cross correlated and give a symmetric 4x4 correlation matrix with diagonal elements of one. Such a matrix is the basis for many multivariate statistical techniques.

Association matrices are always square. Square matrices are those with equal numbers of rows and columns. In statistics square matrices are of great importance. A special type of square matrices are **diagonal matrices** where all elements apart from the diagonal are zero

\[
V = \begin{pmatrix}
a_{11} & 0 & 0 \\
0 & a_{22} & 0 \\
0 & 0 & a_{33}
\end{pmatrix}
\]

If all values of a in a diagonal matrix are 1 we speak of a **unit or identity matrix**

\[
V = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

(5.5)
Identity matrices are equivalent to the 1 of ordinary numbers.

The **transpose matrix** $A'$ is a matrix $(m*n)$ obtained from an original matrix $A$ $(n*m)$ where rows and columns are changed.

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix} \rightarrow A' = \begin{bmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{bmatrix}$$

A transpose matrix that is identical to the original is square and symmetric

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 10 \\ 11 \end{bmatrix} \rightarrow A' = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & 10 \\ 3 & 10 & 11 \end{bmatrix}$$

The Figure below shows again important matrix types.

We now start to define matrix operations. First look at a vector, that means a matrix with only one column (or row). Note that a scalar (a simple number) can be viewed as a vector with only one row or a matrix with only one row and one column. Further we can view a matrix as a grouping of single vectors. A vector denotes a point in space. It length is given by the law of Pythagoras.

$$V = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \rightarrow L = \sqrt{1^2 + 2^2 + 3^2} = 14^{1/2}$$

We see immediately that any vector can be normalized by dividing its elements through the length. The new vector will have a length of 1.

The first operation to introduce is matrix addition. Assume you have insect counts of 4 species (rows) at 3 sites (columns) during 3 months. This can be formulated in matrix...
The total catch per site and species is the sum of the respective matrix elements. We see that matrix addition is only defined for matrices with identical numbers of rows and columns.

Matrix addition immediately leads to the first type of multiplication, the S-product. We have

\[
A = \begin{pmatrix}
1 & 2 & 3 \\
2 & 2 & 4 \\
3 & 1 & 0 \\
\end{pmatrix} + \begin{pmatrix}
1 & 2 & 3 \\
2 & 2 & 4 \\
3 & 1 & 0 \\
\end{pmatrix} + \begin{pmatrix}
2 & 4 & 0 \\
1 & 2 & 0 \\
6 & 9 & 1 \\
\end{pmatrix} + \begin{pmatrix}
2 & 8 & 1 \\
7 & 5 & 5 \\
0 & 0 & 1 \\
\end{pmatrix} = \begin{pmatrix}
5 & 14 & 4 \\
10 & 9 & 9 \\
9 & 14 & 9 \\
\end{pmatrix}
\]

The next example introduces the multiplication of matrices. Assume you have production data (in tons) of winter wheat (15 t), summer wheat (20 t), and barley (30 t). In the next year weather condition reduced the winter wheat production by 20%, the summer wheat production by 10% and the barley production by 30%. How many tons do you get the next year? Of course \((15 \times 0.8 + 20 \times 0.9 + 30 \times 0.7) t = 51 t\). In matrix notion

\[
P = \begin{pmatrix}
15 & 20 & 30 \\
\end{pmatrix} \begin{pmatrix}
0.8 \\
0.9 \\
0.7 \\
\end{pmatrix} = 15 \times 0.8 + 20 \times 0.9 + 30 \times 0.7 = 51
\]

this type of multiplication is called a scalar (or dot) product because it results in a number (a scalar in matrix terminology). In general

\[
A \cdot B = (a_1 \ldots a_n) \cdot (b_1 \ldots b_n) = \sum_{i=1}^{n} a_i b_i = \text{scalar}
\]

We can easily extend this example to deal with matrices. We add another year and ask how many cereals we get if the second year is good and gives 10% more of winter wheat, 20% more of summer wheat and 25% more of barley. For both yes we start counting with the original data and get a vector with one row that is the result of a two step process. First we compute the first year value and then the second year value and combine both scalars in a new row vector with two columns denoting both years.
Now we consider three sites with different harvest. Recall that species are in columns, sites in rows. We get an intuitional definition of the **scalar or dot or inner product** of matrices. The final values give total production at three sites and two years. The result is not a scalar but a matrix.

\[
\begin{bmatrix}
15 & 20 & 30
\end{bmatrix} \cdot \begin{bmatrix}
0.8 & 1.1
0.9 & 1.2
0.7 & 1.25
\end{bmatrix} = \begin{bmatrix}
15*0.8 + 20*0.9 + 30*0.7 & 15*1.1 + 20*1.2 + 30*1.25
\end{bmatrix} = \begin{bmatrix}
51 & 78
\end{bmatrix}
\]

In general we get

\[
A \cdot B = \begin{bmatrix}
\begin{array}{ccc}
\ldots & \ldots & \ldots
\end{array}
\end{bmatrix} \cdot \begin{bmatrix}
\begin{array}{ccc}
\ldots & \ldots & \ldots
\end{array}
\end{bmatrix} = \begin{bmatrix}
\sum_{i=1}^{m} a_i b_i & \ldots & \sum_{i=1}^{m} a_i b_k
\end{bmatrix} = \begin{bmatrix}
A_i B_i & \ldots & A_i B_k
\end{bmatrix}
\]

Hence the dot product can be viewed as a step by step procedure where each row and each column are subjects to single dot products of two vectors. Further, a dot product is only defined if the number of columns of \( A \) is equal to the number of rows in \( B \). The new matrix has the same number of columns than in \( B \) and the same number of rows than in \( A \). Further, in most cases \( A \cdot B \neq B \cdot A \). Next, \( B \cdot B \) only exists if \( B \) is a square matrix.

The above definition implies that if the matrix \( B \) is a simple scalar the dot product simplifies to

\[
A \cdot B = \begin{bmatrix}
\begin{array}{ccc}
\ldots & \ldots & \ldots
\end{array}
\end{bmatrix} \cdot \begin{bmatrix}
\begin{array}{ccc}
\ldots & \ldots & \ldots
\end{array}
\end{bmatrix} = \begin{bmatrix}
A_i c & \ldots & A_i c
\end{bmatrix}
\]

As for scalars we have to look whether the dot product is distributive, associative and commutative. The dot product is generally **not commutative but associative and distributive**.

\[
A \cdot B \neq B \cdot A
\]

\[
(A + B) + C = A + (B + C)
\]

\[
(A \cdot B) \cdot C = A \cdot (B \cdot C) = A \cdot B \cdot C
\]

\[
(A + B) \cdot C = A \cdot C + B \cdot C
\]

In the case of a symmetric matrix an important relation exists

\[
(A \cdot B)' = B' \cdot A'
\]

The **trace of a symmetric matrix** is defined as the sum of all diagonal elements of this matrix.

\[
\text{Tr}(A_{n,n}) = \sum_{i=1}^{n} a_{ii}
\]
It should be noted that another product of two vectors exist, the **outer or cross product**. It gives a new vector that is perpendicular to both original vectors. This product is not used in matrix algebra.

An important transformation of square matrices $A_{nn}$ is the **determinant** $\det A$ or $|A|$. The determinant is a scalar that enables to transform matrices $A$ into new ones $B$. The associated function $B = f(A)$ has to confirm to three basic rules:

1. $B$ is a linear transformation of $A$ that means any change in $A$ results in a linear change in $B$.
2. Any change in the ordering of rows or columns in $A$ should cause a change of sign in $f(A)$.
3. $f(A)$ is determined by a scalar, called the norm or value of $A$ in such a way that the norm of the identity matrix is 1. Hence $f(I) = 1$

The value of a determinant is calculated as the sum of all possible products containing one, and only one, element from each row and each column. These products receive a sign (+ or -) according to a predefined rule. The simplest determinant is

\[
|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}
\] (5.2.12)

Determinants have a simple graphical interpretation (Fig. 5.2.2). The area under the parallelogram spanned by the vectors $\{a_1,a_2\}$ and $\{b_1,b_2\}$ is identical to the determinant of the matrix $A$.

Determinants of matrices of higher order become increasingly time-consuming to be calculated because we get $n!$ permutations of single products. However with days Math programs this is an easy task. Above is the respective norm of an example matrix calculated by *Mathematica.*

Why are determinants important? Determinants are used to solve systems of linear equations. They allow to infer some properties of a given matrix. Particularly holds

1. If a row or a column of a matrix $A$ is zero $\det (A) = 0$.
2. If a row or a column of a matrix $A$ is linearly dependent on another row or column (is proportional to another row or column) then $\det(A) = 0$.
3. If a row or a column of $A$ is multiplied by a scalar $k$ to result in another matrix $B$ then $\det(B) = k \det(A)$. 
If the determinant of a matrix is zero the matrix is called **singular**. If det $A = 0$ not all row or columns are linearly independent. Look for instance at the next matrices

\[
\begin{pmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
2 & 3 & 1 \\
\end{pmatrix}
\quad
\begin{pmatrix}
1 & 2 \\
4 & 5 \\
12 & 15 \\
\end{pmatrix}
\]

In $P$ row 2 can be obtained from row one by the transformation $r_2 = r_1 + 3$. However, this is not a linear transformation. In $Q$, in turn, row 3 is three times row 2 (row 3 is proportional to row 2). This is a **linear transformation** and the respective determinant is zero.

Determinants of upper or lower triangular matrices are easy to compute. The determinant of

\[
\begin{vmatrix}
1 & 2 & 3 \\
0 & -2 & -4 \\
0 & 0 & 2 \\
\end{vmatrix}
= \prod_{i=1}^{3} a_{ii} = 1 \cdot -2 \cdot 2 = -4
\]

Triangular distance matrices are important in multivariate statistics so is the method for computing the determinant.

For ordinary numbers (scalars) the product of a number with its inverse gives always 1. Hence $a \cdot a^{-1} = 1$. The extension of this principle to matrices looks at follows

\[
A \cdot A^{-1} = I
\]

(5.2.13)

To solve this equation we need the inverse of a matrix $A^{-1}$. We see immediately that this operation is only defined for square matrices (why?).

It can be shown that the inverse of a matrix is closely related to its determinant

\[
A \cdot A^{-1} = \frac{1}{|A|} 
\begin{pmatrix}
|A| & 0 & \ldots & 0 & 0 \\
0 & |A| & \ldots & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & 0 & |A| \\
0 & 0 & \ldots & 0 & 0 \\
\end{pmatrix}
= \begin{pmatrix}
1 & 0 & \ldots & 0 & 0 \\
0 & 1 & \ldots & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 1 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 1 \\
\end{pmatrix}
\]

(5.2.14)

The equation tell that an inverse only exists if a matrix is not singular, that is if it has a non-zero determinant. Computing the inverse of a matrix is quite tricky. A formal method provides the **Gauß algorithm** that is implemented in standard matrix software. *Mathematica* calculates inverse matrices with the inverse com-
mand. We can then check whether the calculation conforms to our definition of the inverse. Indeed the dot product of both matrices equals the identity matrix.

Inverse matrices have several important properties.
1. \((B^{-1})^{-1} = B\)
2. \(B^{-1} \cdot B = B \cdot B^{-1} = I\)
3. \(|B^{-1}| = 1 / |B|\)
4. \((A \cdot B)^{-1} = B^{-1} \cdot A^{-1} \neq A^{-1} \cdot B^{-1}\)

One important point is that the inverse of a matrix only exists if \(|B| \neq 0\).

How to apply matrices? A first well known examples deals with systems of linear equations. Take

\[
\begin{align*}
a_1x + a_{12}y &= b_1 \\
a_{21}x + a_{22}y &= b_2
\end{align*}
\]

Solving the system conventionally gives

\[
X = \frac{b_1a_{22} - b_2a_{12}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{\det A_1}{\det A} \quad Y = \frac{b_2a_{11} - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}} = \frac{\det A_2}{\det A}
\]

\[
A_1 = \begin{pmatrix} a_1 & a_{12} \\ a_{21} & a_{22} \end{pmatrix} ; A_2 = \begin{pmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{pmatrix}
\]

We see why the determinant was defined in such a curious way. It has to match the requirements for solving linear algebraic equations. The general solutions for systems with \(n\) equations and \(n\) unknown variables \(X_i\) is:

\[
X_i = \frac{A_i}{A}
\]

However, we can make things easier and solve linear systems without using determinants. Assume you have a system of four linear equations

\[
\begin{align*}
a_1 + 2a_2 + a_3 + 2a_4 &= 5 \\
2a_1 + 3a_2 + 2a_3 + 3a_4 &= 6 \\
3a_1 + 4a_2 + 4a_3 + 3a_4 &= 7 \\
5a_1 + 6a_2 + 7a_3 + 8a_4 &= 8
\end{align*}
\]

This system can be written in matrix notation
This equation has the formal structure of $A \times X = B$. Because $A^{-1} \times A = A \times A^{-1} = I$ we can multiply both sides with the inverse of $A$. We get $X = A^{-1} \times B$ and the solution for the coefficients $a_i$.

$$
\begin{bmatrix}
    a_1 \\
    a_2 \\
    a_3 \\
    a_4 \\
\end{bmatrix} = 
\begin{bmatrix}
    1 & 2 & 1 & 2 \\
    2 & 3 & 2 & 3 \\
    3 & 4 & 4 & 3 \\
    5 & 6 & 7 & 8 \\
\end{bmatrix}^{-1}
\begin{bmatrix}
    5 \\
    6 \\
    7 \\
    8 \\
\end{bmatrix}
$$

The respective *Mathematica* solution looks as follows. First we check for singularity. The determinant of $A$ is -4. Therefore our system should have a solution. Then we compute the inverse and in a last step we multiply with the vector $B$. We get $a_1 = -11/4, a_2 = 17/4, a_3 = -1/4,$ and $a_4 = -1/4$. We matrix notation makes solving a linear system an easy task.

However things are not as easy. First, the inverse matrix has to exist as in our example. If it is singular no inverse exists and the respective determinant is zero. For non-square matrices we need another feature of matrices, the rank. Take the next examples of matrices

$$
A = \begin{bmatrix} 1 & 4 & 8 & 3 \\ 2 & 5 & 10 & 5 \\ 3 & 6 & 12 & 7 \end{bmatrix},
B = \begin{bmatrix} 1 & 5 & 8 & 3 \\ 2 & 3 & 10 & 5 \\ 3 & 1 & 12 & 7 \end{bmatrix},
C = \begin{bmatrix} 1 & 4 & 8 & 3 \\ 2 & 5 & 10 & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix}
$$

In $A$ column $C$ is two times column $B$ and the both square submatrices with three rows/columns (the order of the matrix) have zero determinants. Except of two the $2 \times 2$ submatrices have determinants > 0. The rank is therefore 2. In $B$ both $3 \times 3$ submatrices have determinants > 0 and the rank is 3. In $C$ the $2 \times 2$ submatrices have determinants > 0 and the rank is 2. Hence the rank of a matrix equals the order of the largest submatrix whose determinant > 0.

We further need to know what an augmented matrix is. An augmented matrix is a simple combination of two matrices

$$
A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix};
B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}
\rightarrow A : B = \begin{bmatrix} a_{11} & a_{12} & a_{13} & b_{11} & b_{12} \\ a_{21} & a_{22} & a_{23} & b_{21} & b_{22} \\ a_{31} & a_{32} & a_{33} & b_{31} & b_{32} \end{bmatrix}
$$

The question whether a linear system has solutions is best explained from simple examples.
In the first example we have four unknown variables and four equations. The ranks of $A$ and the augmented $A:B$ are 4. The system has a single solution. In the second case the second equation is simply two times the first. Therefore we have only three independent equations but four variables. The ranks of $A$ and $A:B$ are three and therefore less than $n$, the number of variables. We have infinite solutions. In the third example the rank of $A$ is three (rows one and two are proportional) but the rank of the augmented matrix is four. In such a case no solution exists. In the fourth example we have only three equations but four variables. The ranks of $A$ and $A:B$ are three and less than $n$. An infinite number of solutions exist. In the last two examples we have more equations than variables. In the first of these examples the last equation is inconsistent to the previous. The rank of the augmented matrix is higher than the rank of $A$. In turn, in the last example the last equation is simply two times the previous. It does not contain additional information. The rank of $A$ equals the rank of $A:B$.

In general we can assess whether a linear system has solutions from the ranks of $A$ and $A:B$ and the number of variables $n$ (Fig. 5.2.3). A system has no solution (is inconsistent) if $\text{rank}(A) < \text{rank}(A:B)$. If $\text{Rank}(A) = n$ a single solution exists. Multiple solutions exist if $\text{rank}(A) < n$.
ists, if \( \text{rank}(A) < n \) infinite solutions exists. It is easy to show that a system cannot have a finite number of solutions. Assume we have two solutions. We multiply both equations with with the scalars \( r \) and \( 1-r \).

\[
A \cdot X = B \rightarrow A \cdot rX = rB \\
A \cdot Y = B \rightarrow A \cdot (1-r)Y = (1-r)B
\]

\[
\downarrow
\]

\[
A \cdot rX + A \cdot (1-r)Y = rB + (1-r)B
\]

\[
\downarrow
\]

\[
A \cdot [rX + (1-r)Y] = B
\]

We got a new equation of the structure \( A \cdot Z = B \). Hence \( [rX + (1-r)Y] = Z \) must be a solution of our initial system. Because \( r \) can be any scalar their must be infinite solutions of \( A \). Important is that a solution as in Fig. 3.3 is only possible if we can compute the inverse of \( A \cdot Z = B \). The inverse is only defined for square matrices. Hence, the number of equations must be identical to the number of variables.

Matrix algebra can also be used to solve algebraic equations. For instance logistic pollution growth follows a second order algebraic (quadratic) model

\[
\frac{dN}{dt} = -aN^2 + bN + c
\]

At three populations sizes \( N \) (1, 5, 10) we got the following rates of increase \( \frac{dN}{dt} \): 5, 25, 8 individuals generations\(^1\). Now our model looks as follows

\[
\begin{align*}
-a + b + c &= 5 \\
-a5^2 + b5 + c &= 25 \\
-a10^2 + b10 + c &= 8
\end{align*}
\]

\[
\begin{align*}
-a + b + c &= 5 \\
-a5^2 + b5 + c &= 25 \\
-a10^2 + b10 + c &= 8
\end{align*}
\]

\[
\begin{bmatrix}
-a + b + c = 5 \\
-a5^2 + b5 + c = 25 \\
-a10^2 + b10 + c = 8
\end{bmatrix}
\]

\[
\begin{bmatrix}
-a + b + c = 5 \\
-a5^2 + b5 + c = 25 \\
-a10^2 + b10 + c = 8
\end{bmatrix}
\]

\[
\text{rank}(A) = \text{rank}(A:B) = n. \text{ Hence a single solution exists. This is shown above. We see again how simple it is to solve such a system using math programs. The logistic growth model is}
\]

\[
\frac{dN}{dt} = -\frac{14}{15}N^2 + \frac{53}{5}N - \frac{14}{3}
\]

\[
\text{and the respective solution of } N(t) \text{ is obtained from the solution of the respective differential equation.}
\]

The model is a tangens hyperbolicus function (Fig. 3.4) equivalent to the logistic growth equation. \( E \) get the maximum abundance \( K \) from the Fig. \( K = 10.89 \).

Matrix notation used in this way makes it also easy to compute a linear regression. Assume you have two sets of data \( X \) and \( Y \). This corresponds to a system of linear equations

\[
\begin{align*}
Y_1 &= b_0 + b_1 X_1 \\
\vdots \\
Y_n &= b_0 + b_1 X_n
\end{align*}
\]

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix}
1 & X_1 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix} \cdot \begin{bmatrix}
b_0 \\
b_1
\end{bmatrix}
\]

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix}
1 & X_1 \\
\vdots & \vdots \\
1 & X_n
\end{bmatrix} \cdot \begin{bmatrix}
b_0 \\
b_1
\end{bmatrix}
\]

\[
(5.2.18)
\]

\[
\begin{array}{cc}
X & Y \\
1 & 2 \\
2 & 1 \\
3 & 4 \\
4 & 3 \\
5 & 6 \\
6 & 5 \\
7 & 8 \\
8 & 7 \\
9 & 10 \\
10 & 9
\end{array}
\]
To solve this system we first multiply with the transpose of \(X\), \(X'\) to get a square matrix \(X'\cdot X\) that can be inverted. We get the least squares solution of a linear regression:

\[
Y = X\cdot b \rightarrow X'\cdot Y = (X'\cdot X)\cdot b \\
b = (X'\cdot X)^{-1} \cdot (X'\cdot Y)
\]  

(5.2.19)

We test this approach with a numerical example. We have ten data pairs \(\{X,Y\}\). We first calculate the transpose \(X'\) and then the inverse of the dot product \(X'\cdot X\). We solve for \(b\) by multiplying with \((X'\cdot X)^{-1}\) and get the dot products of \((X'\cdot X)^{-1}\cdot X'\cdot Y\). The result is \(b_0 = 1/3\) and \(b_1 = 31/33\). Excel computes for the data set \(b_0 = 0.333\) and \(b_1 = 0.939\). The results are identical. Because dot products are generally not commutative we have to be careful about the ordering of matrices.

This approach can easily be extended for instance to polynomial regression. If you have a regression model of

\[
Y = b_0 + b_1 X + b_2 X^2 \ldots + b_n X^n
\]

The respective matrix notation looks as follows

\[
\begin{align*}
Y_i &= b_0 + b_1 X_i^2 + \ldots + b_n X_i^n \\
\vdots \\
Y_n &= b_0 + b_1 X_n^2 + \ldots + b_n X_n^n
\end{align*}
\]

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix} = 
\begin{bmatrix}
1 & X_1 & \ldots & X_1^n \\
\vdots \\
1 & X_n & \ldots & X_n^n
\end{bmatrix} \cdot 
\begin{bmatrix}
b_0 \\
\vdots \\
b_n
\end{bmatrix}
\]

The solution is the same as above

\[
Y = X\cdot b \rightarrow X'\cdot Y = (X'\cdot X)\cdot b \\
b = (X'\cdot X)^{-1} \cdot X'\cdot Y
\]  

(5.2.20)

We try with a slightly modified data set (squared values of \(Y\)) and a second order algebraic function \(Y = b_0 + b_1 X + b_2 X^2\).

Combining all the operations into a single Mathematica task gives \(b_0 = -26/5 = -5.2\), \(b_1 = 38/11 = 3.45\), and \(b_2 = 7/11 = 0.636\). Excel gives \(b_0 = -5.2\), \(b_1 = 3.45\), and \(b_2 = 0.636\). Again both results are identical. If the regression is forced to go through the origin \((b_0 = 0)\) the first column of \(X\) has to be set to zero which is the same as to leave it. The method can, of course, easily be extended to higher order polynomials and is easy to compute.

In the tables at the beginning of this section we had typical biological data structures. The descriptors (the variables) are often correlated. We can construct square matrices that contain information about distances,
for instance correlation matrices. But we can also construct such matrices that contain covariances.

Covariance measures the joint dispersion of two variables. The matrix $\Sigma$ is called the dispersion matrix.

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \ldots & \sigma_{1n} \\
\vdots & \ddots & \vdots \\
\sigma_{n1} & \ldots & \sigma_{nn}
\end{pmatrix}
\]

(5.2.22)

Covariance is calculated from

\[
\sigma_{ij} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \mu_j)(x_{ik} - \mu_k)
\]

(5.2.23)

$[X-\mu]$ is the centralized column vector of the data. Hence to compute the covariance this vector has to be multiplied by its transpose. The diagonal values $\sigma_{ii}$ are of course the respective variances of the variables $x$.

Another important matrix is the matrix of correlations

\[
P = \begin{pmatrix}
\rho_{11} & \ldots & \rho_{1n} \\
\vdots & \ddots & \vdots \\
\rho_{n1} & \ldots & \rho_{nn}
\end{pmatrix}
\]

(5.2.24)

Because of $\rho_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)$ we can introduce a new matrix $D$ that contains only the square roots of the diagonal values (= standard deviations) of the dispersion matrix. Hence

\[
P = D^{-1} \cdot \Sigma \cdot D^{-1} \rightarrow \Sigma = D \cdot P \cdot D
\]

(5.2.25)

The new matrix $\Sigma$ is a square matrix and contains the covariances of our variables. This is shown in the Mathematica example beside. $U$ is a correlation matrix and $S$ the dispersion matrix containing the standard deviations of the four variables included. The result gives a matrix that contains all covariances and as diagonal values the variances

Next we have to deal with eigenvalues and eigenvectors. Assume we have a matrix $A$ in two dimensional space defining data points in an $x,y$ system (Fig. 5.2.4). We can now define new orthogonal axes $u_1$ and $u_2$ that minimize the distances $\delta_1$ and $\delta_2$ to our data points. These new axes (vectors) are called principal axes of the matrix $A$. The length of the vectors $u_1$ and $u_2$ are called the eigenvalues $\lambda_1$ and $\lambda_2$. Fig. 5.2.5 and 5.2.6
also show how to interpret eigenvalues and eigenvectors. Assume our data points to be enclosed by an ellipse. The longer diagonal marks the first eigenvector. The longer this vector is the closer scatter the data points along this vector. Hence $\lambda$ is a measure of how well the principal axis describes our data matrix.

Eigenvalues $\lambda_i$ and eigenvectors $u_i$ are connected to the matrix $A$ in such a way that the dot product of $A \cdot u_i$ gives $u_i$ multiplied with $\lambda_i$. Hence

$$A \cdot u_i = \lambda_i u_i$$  \hspace{1cm} (5.2.26)

Eq. 5.26 can be transformed to

$$A \cdot u_i = \lambda_i u_i \rightarrow A \cdot u_i - \lambda_i u_i = 0 \rightarrow (A - \lambda I) \cdot u_i = 0$$  \hspace{1cm} (5.2.27)

Eq. 5.25 might be interpreted as follows. How long do I have to stretch the vector $u$ to get a new matrix $A \cdot u$ from $A$. Introducing eq. 5.25 we get

$$\Sigma_i = U \cdot \lambda U = \lambda \cdot U \cdot U = \lambda$$  \hspace{1cm} (5.2.28)

For any two principal axes vectors $u_i$ and $u_j$ the covariance $s(u_i, u_j)$ becomes

$$s(u_i, u_j) = u_i \cdot \lambda_j u_j = \lambda_j u_i \cdot u_j = 0$$  \hspace{1cm} (5.2.29)

becomes the dot product of orthogonal vectors is zero. The last two equations are of great importance. They tell that (a) principal axes are independent and (b) that the eigenvalues tell how much of total variance is explained by a given principal axis (eigenvector). Several methods of ordination and grouping for instance use this technique to reduce the number existing intercorrelated variables in such a way that new composite variables appear that are not correlated. The variables are defined by the principle axes and the goodness of fit is given by the associated eigenvalues. The general idea is to get a small number of independent variables that are easy to interpret.

---

Fig. 5.2.4

![Figure 5.2.4](image)

Fig. 5.2.5

![Figure 5.2.5](image)

Fig. 5.2.6

![Figure 5.2.6](image)
Eq. 5.2.26 is of course solved by a null vector $u$. The equation is also zero if $\det (A-\lambda I) = 0$. This property gives a solution for $\lambda_i$. In other words we are looking for such a $\lambda$ that makes $(A-\lambda I)$ to a vector that is orthogonal to $u$. An example. For the matrix $A = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}$ we need to get the associated eigenvectors from both $\lambda$ values by solving

$$\begin{pmatrix} 2-\lambda & 1 \\ 3 & 4-\lambda \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0$$

$$\begin{pmatrix} 2-\lambda & 1 \\ 3 & 4-\lambda \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0 \Rightarrow \begin{pmatrix} 2-\lambda & 1 \\ 3 & 4-\lambda \end{pmatrix} = 0 \Rightarrow (2-\lambda)(4-\lambda) = 3 \Rightarrow \lambda_1 = 1; \lambda_2 = 5$$

We get the associated eigenvectors from both $\lambda$ values by solving

$$\begin{pmatrix} 2-\lambda & 1 \\ 3 & 4-\lambda \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0 \Rightarrow \begin{pmatrix} 2-\lambda u_1 + u_2 = 0 \\ 3u_1 + (4-\lambda)u_2 = 0 \end{pmatrix} \Rightarrow u = \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix}$$

We check

$$\begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 15 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

However these vectors are not orthogonal. $u_1 \cdot u_2 \neq 0$. However, one special type of matrices gives orthogonal eigenvectors. These are symmetrical matrices. Symmetrical matrices are often used in biology. These are for instance most association and all dispersion matrices where $D_i = D_i$. The diagonal elements are 1. Look at the next Mathematica solution. A symmetrical association matrix has orthogonal eigenvectors. We check for this by multiplying both. The result is zero. The Fig. 5.2.7 shows both eigenvectors. They are indeed orthogonal. Hence symmetrical matrices have orthogonal eigenvectors.

The Mathematica example beside shows a numerical example of eq. 5.26. The dot product of a matrix $A$ with its eigenvector is identical to the multiplication of the eigenvector with its eigenvalue. Further, it holds

$$A \cdot U = U \cdot \Lambda$$

that means the matrix of eigenvectors $U$ transposes a matrix $A$ into the matrix of eigenvalues $\Lambda$.

How to calculate principal axes ? With some mathematicians one can show that

$$(\Sigma - \lambda I)[x - \mu] = 0$$

(5.2.30)

with $\Sigma$ being the dispersion matrix and $[x - \mu]$ the matrix of $x_i - \mu$ values for each variable.

This equation has the same structure as the equation of the eigenvectors (5.26). Hence $[x - \mu]$ is one of the eigenvectors of $\Sigma$. We take a simple example of two variables $x$.
and y and want to calculate the principle axes. We need the dispersion matrix that contains the covariances. The variance of X is \( \sigma_x^2 = 20 \), of y \( \sigma_y^2 = 25.03 \), and the covariance is \( \sigma_{xy} = 16.73 \). We compute the eigenvalues and eigenvectors of \( S \) and get the first eigenvector \( E = \{-0.65, -0.76\} \). The principal axis slope is therefore \( m = -0.76/-0.65 = 1.16 \). From \( y = mx + b \) we get the intercept \( b = -3.64 \) (Fig. 5.2.8). We performed a major axis regression.

The solution is given in beside. We need the dispersion matrix \( \Sigma \) and compute the eigenvectors and eigenvalues. The eigenvalue \( \lambda \) is identical to the \( \lambda \) of eq 4.12 of the major axis regression. In general we get the slope \( m \) of the major axis regression from the values \( y / x \) of the first major axis, that means the axis of the largest eigenvalue.

\[
\begin{align*}
m &= \frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 - 4\sigma_{xy}^2}} \\
\lambda &= \frac{1}{2} \left( \sigma_x^2 + \sigma_y^2 + \sqrt{(\sigma_x^2 - \sigma_y^2)^2 - 4\sigma_{xy}^2} \right) \Rightarrow 2\lambda - s_x^2 = s_y^2 + \sqrt{(s_x^2 - s_y^2)^2 - 4\sigma_{xy}^2} \\
m &= \frac{\sigma_{xy}}{\lambda - \frac{s_y^2}{s_x^2}}
\end{align*}
\]

(5.2.31)

This result is identical to eq. 4.11. We check \( m = 16.73/(39.42 - 25) = 1.16 \).

How to get the dispersion matrix \( \Sigma \)? In our case this is the matrix of covariances. The Excel example beside shows that we need the matrix of \([x-m] \) centralized vectors of X and Y. The dot product of this matrix with its transpose gives the requested dispersion matrix.

Another application of eigenvectors and eigenvalues. For Markov chains it is often necessary to calculate the power of a matrix \( A^k \). We got the diagonal matrix L from \( A \cdot U = \Lambda \cdot U = U \cdot \Lambda \). Look at the example below. We see that \( A \cdot U = U \cdot \Lambda \) with U being the matrix of eigenvectors and \( \Lambda \) the diagonal matrix of eigenvalues. We
We need the matrix of eigenvectors, its inverse and the vector of eigenvalues where each element $\lambda_i$ is raised to power $n$.

A next important property is that the determinant of a square matrix is identical to the product of its eigenvalues.

$$|P_{nn}| = \prod_{i=1}^{n} \lambda_i$$

(5.2.33)

Hence a determinant is zero if at least one of its eigenvalues is zero.
6. Multiple regression

Surely, the most often used statistical tool of biologists is regression analysis. Multiple regression is the generalization of simple linear regression where you try to predict one **dependent variable** from a second **independent variable**. In multiple regression you have again one dependent but now a set of \( n \) independent variables. In mathematical terms

\[
Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n = b_0 + \sum_{i=1}^{n} b_i X_i
\]

(6.1)

Recall that \( X \) and \( Y \) are series of observations. Hence, we can write them in vector notation

\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{pmatrix} = \begin{pmatrix}
  1 & x_{1,1} & \ldots & x_{1,k} \\
  1 & x_{2,1} & \ldots & x_{2,k} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & x_{n,1} & \ldots & x_{n,k}
\end{pmatrix} \begin{pmatrix}
  b_0 \\
  b_1 \\
  \ldots \\
  b_n
\end{pmatrix} = X \cdot b
\]

(6.2)

We solve this equation as before and take care of the ordering of elements.

\[
Y = X \cdot b \rightarrow X'Y = (X'X) \cdot b \rightarrow b = (X'X)^{-1} \cdot (X'Y)
\]

(6.3)

Let’s take a simple example. Three predictor (independent) variables \( X \) are assumed to influence the dependent variable \( Y \). The complete multiple regression needs two lines of programming and the output gives us the parameters \( b_0 = 0.54, b_1 = -0.21, b_2 = 0.23, \) and \( b_3 = -0.006 \). Our regression model looks as follows

\[
Y = 0.54 - 0.21X_1 + 0.23X_2 - 0.006X_3
\]

The previous model used raw data. Now we take a slightly different approach. We standardize our variables and use the common Z-transformation. Recall that \( Z_i = (X_i - \mu)/\sigma \) with \( Z \) having a mean of 0 and a standard deviation of 1. The regression model looks as follows

\[
\begin{array}{c|c|c|c|c}
Y & X1 & X2 & X3 \\
0.78 & 1.10 & 1.51 & 1.37 \\
0.73 & 2.84 & 3.35 & 4.24 \\
0.87 & 3.43 & 3.72 & 4.76 \\
0.55 & 1.97 & 2.90 & 3.27 \\
0.81 & 0.29 & 0.72 & 1.01 \\
0.26 & 1.52 & 1.32 & 2.04 \\
0.65 & 2.85 & 3.53 & 4.11 \\
0.76 & 2.83 & 3.46 & 4.69 \\
\end{array}
\]
Applying equation 3 we get

\[ Z_Y = Z_X \cdot b \rightarrow Z_X^\top Y = (Z_X^\top Z_X) b \rightarrow b = (Z_X^\top Z_X)^{-1} (Z_X^\top Z_Y) \]  \hfill (6.5)

How to interpret $Z_X Z_X^\top$ and $Z_X^\top Z_Y$? Look at the definition of the coefficient of correlation:

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \frac{s_X}{s_y} = \frac{1}{n-1} \sum_{i=1}^{n} Z_X Z_Y \]  \hfill (6.6)

This is a very important equation that tells that a coefficient of correlation is the sum of all pairwise $Z$-values of $X$ and $Y$. Now look at the $Z_X Z_Y$ matrix. This is nothing else that this sum calculated for all $X$ variables. For the $Z_X^\top Z_X$ matrix holds the same. It is identical to the correlation matrix for all $X$. Hence we can write

\[
\begin{bmatrix}
\frac{1}{n-1} \sum_{i=1}^{n} Z_{X1} Z_{X1} & \cdots & \frac{1}{n-1} \sum_{i=1}^{n} Z_{X1} Z_{Xk} \\
\vdots & \ddots & \vdots \\
\frac{1}{n-1} \sum_{i=1}^{n} Z_{Xk} Z_{X1} & \cdots & \frac{1}{n-1} \sum_{i=1}^{n} Z_{Xk} Z_{Xk}
\end{bmatrix}
= \begin{bmatrix}
r_{X1X1} & \cdots & r_{X1Xk} \\
\vdots & \ddots & \vdots \\
r_{XkX1} & \cdots & r_{XkXk}
\end{bmatrix}
\]  \hfill (6.7)

From this identity we get a new simple equation for the multiple regression.

\[ \beta = R_{XX}^{-1} R_{XY} \]  \hfill (6.8)

Where $R$ denotes the respective correlation matrices. The Mathematica solution for the example in the previous table is shown below. Our model is

\[ Y = -0.71 - 1.08 X_1 + 1.20 X_2 - 0.03 X_3 \]

This result based on $Z$-transformed data differs from the above that was based on raw data. In the latter case we got standardized correlation coefficients, so-called beta values.

Our multiple regression tries to predict the values of the dependent variable on the basis of the inde-
dependent variables. However, the model does not tell anything about the goodness of fit. In ordinary regression this measure is the coefficient of determination $R^2$. $R^2$ is defined as the proportion of variance in $Y$ that is explained by the model. Using the model with Z-transformed data we get

$$R^2 = 1 - \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{Z_{\text{pred}} - Z_{\text{pred}}^i}{s_{Z(Y)}} \right)^2 = 1 - \frac{1}{n-1} \sum_{i=1}^{n} Z_{\text{pred}}^2 = \frac{1}{n} \sum_i \beta_i r_{yi}$$

(6.9)

where $k$ is the number of independent variables. Hence $R^2$ equals the sum of all Z-transformed predicted values divided through the number of cases $n$. In a last step we need information whether the grand $R^2$, the model wide coefficient of variation and the beta values of each model variable explains a significant part of total variance. In order to do this we have two possibilities. First we can use standard F and t tests. Hence

$$t(f) = \frac{\text{beta}}{\text{Standard error beta}}$$

(6.10)

with $f$ being the degree of freedom. The standard error of beta is computed from

$$\text{SE} (\beta_i) = \sqrt{\frac{r_{ii}(V^{-1})(1-R^2)}{n-k-1}}$$

(6.11)

with $r_{ii}$ ($V^{-1}$) being the element ii of the inverted association matrix $V$ and $n$ the number observations and $k$ the number of independent variables in the model. The significance of the whole model (the significance of $R^2$) is tested by an F-test.

$$F = \frac{R^2}{(1-R^2)} \frac{n-k-1}{k}$$

(6.12)

Both tests assume normally distributed errors and are only approximations. Today standard errors are most often obtained from permutation tests. Hence the values $X_i$ of the predictor variables are randomly reshuffled and after each such reshuffling a new multiple correlation is computed. This method gives a range of different beta values for each predictor variable. The distribution of values gives then the required standard error of beta.

The Table below presents a typical problem biologists are faced with. We got data about abundances of

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.572</td>
<td>0.000</td>
<td>0.000</td>
<td>0.041</td>
<td>0.690</td>
<td>0.842</td>
<td>1.841</td>
<td>2.227</td>
</tr>
<tr>
<td>0.983</td>
<td>0.304</td>
<td>0.038</td>
<td>0.646</td>
<td>1.211</td>
<td>0.443</td>
<td>1.143</td>
<td>3.245</td>
</tr>
<tr>
<td>0.654</td>
<td>0.961</td>
<td>0.875</td>
<td>1.948</td>
<td>1.294</td>
<td>0.585</td>
<td>1.456</td>
<td>2.181</td>
</tr>
<tr>
<td>0.579</td>
<td>0.066</td>
<td>0.023</td>
<td>0.230</td>
<td>1.439</td>
<td>0.680</td>
<td>0</td>
<td>2.074</td>
</tr>
<tr>
<td>0.329</td>
<td>0.291</td>
<td>0.266</td>
<td>0.623</td>
<td>1.263</td>
<td>0.471</td>
<td>1.901</td>
<td>1.908</td>
</tr>
<tr>
<td>0.335</td>
<td>0.097</td>
<td>0.061</td>
<td>0.292</td>
<td>0.836</td>
<td>0.619</td>
<td>1.257</td>
<td>2.321</td>
</tr>
<tr>
<td>0.295</td>
<td>0.139</td>
<td>0.019</td>
<td>0.278</td>
<td>0.476</td>
<td>0.015</td>
<td>1.245</td>
<td>1.476</td>
</tr>
<tr>
<td>0.404</td>
<td>0.761</td>
<td>0.433</td>
<td>1.583</td>
<td>1.240</td>
<td>0.556</td>
<td>1.669</td>
<td>1.715</td>
</tr>
<tr>
<td>0.506</td>
<td>0.129</td>
<td>0.100</td>
<td>0.307</td>
<td>0.885</td>
<td>0.444</td>
<td>1.728</td>
<td>2.527</td>
</tr>
<tr>
<td>0.127</td>
<td>0.606</td>
<td>0.452</td>
<td>6.147</td>
<td>0.860</td>
<td>0.932</td>
<td>0.192</td>
<td>0.727</td>
</tr>
</tbody>
</table>
the gall inducing hymenopteran insect Andricus curvator, a small cynid wasp, which induces large galls in the leaves of oaks. We are interested too see on which factors numbers of galls depend. We got data about two competing wasp species, Neuroterus albipes and Cynips quercusfolii, and the parasitic wasps Torymus sp. and Charips sp. Additionally, we have data about precipitation and temperature during the sampling period. At last we speculate that the height of the galls (in m) and the abundance of Andricus in the previous generation might influence the actual gall densities. From this we make a model that contains one dependent variable, \( \text{(Andricus)} \) and seven independent variables. In total we have 10 observational data sets (cases) (Tab. 1).

Now we take a common statistical package and copy our data to the package spreadsheet and run the multiple regression option with all default settings. We get Tab. 2 and are lucky. The multiple regression points to highly significant predictor variables. Only temperature seems to be unimportant.

The output contains a number of important metrics. First it gives a multiple regression coefficient of correlation \( R \), in analogy to the simple Pearson coefficient. It denotes the fraction of total variance explained by the model \( R^2 \), as well as a corrected level of variance that corrects for small sample size corr \( R^2 \) (correction for shrinkage after Carter). The significance level (the probability to make a type I error) is computed by an ordinary F-test. Next we have the predicted model parameters \( B \) and the beta-values. Note that beta and \( b \)-values are connected by the following equation.

\[
\text{beta}(X_i) = B(X_i) \frac{s(X_i)}{s(Y)}
\]  

(6.13)

where \( s \) again denotes the standard deviation of variables \( X_i \) and \( Y \).

In the case of only two variables beta-values are identical with the Pearson coefficient of correlation. The output contains also standard errors of the parameters. Be careful, standard errors not standard deviations. All multivariate statistics give standard errors. For t-tests or other pairwise comparisons instead you need standard deviations. Don’t use the errors of regression analysis unconsciously for further analysis, for instance for tests whether coefficients of correlations or slopes of two models differ. Because standard errors are always lower than standard deviations you would get highly significant results even where no such are. The t-test at the end of the table tests whether the beta-values differ significantly from zero.

We should undertake a stepwise regression analysis. That means we should step by step eliminate all variables from our regression that are not significant at a given p-level (this is in most cases the 5% error level).
Statistical advices

Stepwise regression throws out Temperature and gives us a model where all independent (predicator) variables are significant at the 5% error level. It seems that we can prepare a paper.

Really? In this short example we made a whole set of errors. We applied a method without prior checking whether the prerequisites of the method are met. In fact, the data of Tab. 1 are nothing more than simple random numbers or combinations of random numbers. How is it possible to get highly significant results with them?

A multiple regression relies on the general linear model (GLM). Hence it must be possible to formulate the whole regression model as a linear combination of variables. The variables must be connected by

6.1 How to interpret beta-values

An important but also difficult problem in multiple regression is the interpretation of beta-values. Remember that

$$\text{beta}(X_i) = B(X_i) \frac{s(X_i)}{s(Y)}$$

Remember also that the coefficient of correlation for a simple correlation between two variables is defined as

$$r = b \frac{s_Y}{s_Y}$$

It is immediately evident that beta-values are generalisations of simple coefficients of correlation. However, there is an important difference. The higher the correlation between two or more predicator variables (multicollinearity) is, the less will r depend on the correlation between X and Y. Hence other variables might have more and more influence on r and b. For high levels of multicollinearity it might therefore become more and more difficult to interpret beta-values in terms of correlations. Because beta-values are standardized b-values they should allow comparisons to be make about the relative influence of predicator variables. High levels of multicollinearity might let to misinterpretations.

Hence high levels of multicollinearity might

- reduce the exactness of beta-weight estimates
- change the probabilities of making type I and type II errors
- make it more difficult to interpret beta-values.

There is an additional parameter, the so-called coefficient of structure that we might apply. The coefficient of structure c_i is defined as

$$c_i = \frac{r_{iY}}{\sqrt{R^2}}$$

where r_{iY} denotes the simple correlation between predicator variable i and the dependent variable Y and R^2 the coefficient of determination of the multiple regression. Coefficients of structure measure therefore the fraction of total variability a given predicator variable explains. Again, the interpretation of c_i is not always unequivocal at high levels of multicollinearity.
linear regressions of the type $Y = aX + b$. We check this and don’t detect any significant deviation from linearity. In fact, however, I computed the previous generation data of *A. curvator* as a power function with random offset from the *A. curvator* data. We did not detect this because we have only ten observations and the range of values is too small to detect deviations from linearity. This leads us immediately to a second major error. We have eight variables but only ten observations. The number of cases is much too low. A general rule is that the number of data sets should be at least 2 times the number of variables included in the original model. In our case we must reformulate our initial model to include at most four variables.

There is a simple equation to compute the optimal sample size $n$ (although in reality we will seldom have the opportunity to have such large sample sizes. Instead, we will most often only deal with minimal sample sizes.

$$n = \frac{L(1 - R^2)}{R^2}$$

(6.14)

where $R$ is the desired experiment wise coefficient of determination (explained variance). This value depends on what we intend to accept as a significant difference, the effect size. We know effect sizes already from bivariate comparisons and the discussion of the t-test. $R^2$ and effect size $\varepsilon^2$ are related by the following equation

$$\varepsilon^2 = \frac{R^2}{1 - R^2}$$

(6.15)

For $\varepsilon^2 = 0.02$ (week effect) $R^2 = 0.02$; $\varepsilon^2 = 0.15$ (medium effect) $R^2 = 0.13$; for $\varepsilon^2 = 0.35$ (strong effect) $R^2 = 0.26$ The L values that can be obtained from the following Table

| Variables | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 25  |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| L         | 7.8 | 9.7 | 11.1| 12.3| 13.3| 14.3| 15.1| 15.9| 16.7| 17.4| 18.1| 18.8| 19.5| 20.1| 20.7| 22.5| 23.1| 23.7| 24.3| 25.9|

sets to see a strong effect.

If we find non-linear relationships between variables we have to linearize them, for instance by appropriate logarithmization. But be careful. Logarithmization changes the distribution of errors around the mean. The GLM but relies on the assumption that the errors are distributed normally around the mean, and that they are not correlated with the dependent variable. They have to be homoscedastic. Otherwise we speak of heteroscedasticity.

A good example is Taylor’s power law. Insects, for instance fluctuate in abundance. Assume we study ten species with different upper limits of abundance, different carrying capacities. Insight these boundary abundances fluctuate at random. This is a typical statistical process called proportional rescaling. For such a process it is easy to show that mean and variance are connected by a simple equation

$$\text{Variance } \propto \text{ mean}^2$$

(6.16)

Read: The variance is proportional to the square of the mean. Hence the variance, the distribution of errors around the mean, rises linearly with the squared mean. Mean and variance are correlated. This violates the assumptions of the GLM. Fortunately, the model is very robust against this type of correlation. The B– and beta values are nearly unaffected but the error levels p might be distorted. In non-linear relationships variances and
means are often correlated.

Next the errors around the mean must not be correlated with each others. They must not be autocorrelated. Be careful at high levels of autocorrelation. All statistical packages contain a simple test for autocorrelation, the Durbin-Watson test for serial correlation D. Without going into the details I tell that D-values around 2 indicate low levels of autocorrelation. In our example D is 2.93. This high value points to severe autocorrelation and we have to check our data which of the variables causes this problem. We have to eliminate the variable from the initial model. Another method to detect violations of the GLM assumptions is a visual inspection of the residuals. Residuals are the deviation of the data from the regression line and are often used for further analysis. Residuals should be normally distributed around the mean (the value of the regression line). Any regularities in the distribution of residuals hence point to heteroscedasticity or autocorrelation and therefore to an incorrect use of multiple regression. However, as already stressed, multiple regression is a quite robust method and only severe distortions of the residual distribution make the results unreliable.

It’s still unclear why we got highly significant results from random data. The first answer is that some of the predictor variables were highly correlated. Height and Torymus were highly correlated (R² = 0.99). This violates the assumption of the GLM. Although multiple regression is a technique designed to detect interrelations between sets of predictor variables pairwise correlations must not be too high. In other words they must not explain the same part of variance. If they were perfectly correlated a multiple regression would be impossible. Technically speaking the matrix of correlation coefficients would be singular and the inverse would not be defined. Including such variables would also give evidence of an ill designed model. That something is wrong with the variables Height and Torymus is also indicated by the values of beta. As coefficients of correlation, beta-values should range between –1 and 1. The above equation however tells that absolute values larger than one are possible. However, very high or low values of beta often indicate some violation of the GLM assumptions.

A next answer lies in the distinction between test wise and experiment wise error rates. What does this mean? If you test whether single variables are significant or compare two means you perform a test and have a probability to make a type I error (your significance level). However, if you test a whole model (the outcome of an experiment) you have to perform a series of single tests and you get a type I error level that refers to the whole experiment. In our case we got 8 test wise error levels and one experiment wise multiple R² with p(t) < 0.051. If you accept a single test at p < 0.05 and perform n tests your probability to get at least one significant result is

\[ p = 1 - (1 - 0.05)^n \]

(6.17) (why?). Hence, for n = 7 you get p = 0.30, a value much higher than 0.05. This means for our example that the probability to get at least one significant result out of seven predictor variables is nearly 30%. In a multiple regression this value is even too low because you have to consider all combinations of variables. Additionally, the lower the number of observations is, the higher is the probability for a significant result (why?). Hence to be sure to make no error you have to reduce your test wise error rate to obtain an experiment wise error rate of less than 0.05. The most easiest way to do this is to use a so-called Bonferroni correction obtained from the first element of the Taylor expansion of eq. 16. You simply divide your test wise error rate \( \alpha \) by the number of variables in the model.
In our example of the first table you should accept only variables with \( p < 0.05 / 8 = 0.006 \). Hence, only the previous generation of *A. curvator* remains in the model. This correction should always be used in multivariate statistics.

Let us at the end assume that the data of Tab. 1 were real. We can then redefine our initial model to perform a correct multiple regression. First, we have to reduce the number of variables. By this we also define a logical model that we want to test with the regression analysis. Our data contain three groups of variables. *Torymus* and *Charips* are parasitoids or inquilines of *Andricus*. We simply combine both variables and use a new variable 'parasitoid abundance' as the sum of both. By this we also reduce the degree of multicollinearity in the data set because *Torymus* and *Charips* abundances are highly correlated. We also have two variables that describe weather conditions. They are not correlated. We define the precipitation / temperature ratio P/R, the hydrothermal coefficient, as a new variable that describes moisture. Low values indicate rather dry periods, high values moist conditions. Height was highly correlated with *Torymus*. At the moment we have no good explanation for this pattern and skip the variable height. We would have to undertake further experiments to clear this point.

With this variables we undertake a new regression analysis and find only the competitors to be significant. The Durbin-Watson statistics gives a value of \( D = 2.07 \). We do not expect higher levels of autocorrelation. We drop the other variables and get at the end a simple linear regression between *Andricus* and its competitors \( (Andricus = (0.39 \pm 0.09) \text{Competitors} - (0.33 \pm 0.19); R^2_{corr.} = 0.69; p < 0.002 \). Note that the **errors are standard errors not standard deviations**. The p-value is smaller than our Bonferroni corrected acceptable error level of 0.006 and we might accept the hypothesis about a close relation between *Andricus* and *Cynips* abundances. But be careful. Both data sets consist of pure random numbers. I needed only five reshuffles to get the data matrix for this result. The example shows how easy it is to get highly significant results from nothing.

Multiple regression uses ordinary least squares and relies on coefficients of correlation as distance measure. But recall the basic model to calculate the regression

\[
\beta = R_{XX}^{-1} R_{XY}
\]

(6.19)

The model contains two association matrices. We might use other measures of distance to get different regression models. For instance look at the next table. It gives the occurrences of 12 species at five sites. This is a presence-absence table. Does the occurrence at sites two to five predict the occurrence at site one? We use the Soerensen index and get the respective association matrix. With this matrix we run a multiple regression and get the respective beta values and the inverse of R. This is enough to calculate the coefficients of determination and significance levels using equations 8 to 11. The matrix below...
Statistical advices

The results and the calculations show that none of the single predictors is statistically significant. The reason for the low t-values is that the numbers of species (cases) used is too low. The regression used in this way has a week power that means it too often favours the null hypothesis of no influence. For a sound regression model we needed at least 50 species.

The whole model, in turn, appears to be significant before Bonferroni correction. The Bonferroni corrected significance level would be 0.09. Hence despite the high $R^2$ value of 0.87 (calculated from the beta values and the respective correlation coefficients according to eq. 6.9) we infer that the predictor variables do not allow for a prediction of site occurrences at site one. Again the reason for this discrepancy is the too low number of cases.

We used a multiple regression approach of presence absence data associated with the Soerensen index. Important is that we run a multiple regression of nominally scaled data. Indeed, we can even used mixed models with nominally, ordinary and metrically scaled variables. Important is that we have to be able to use the same association measure for all of these data types. Often we have to apply a previous classification. Important is also that we do not get regression equations to predict occurrences. We only infer significant influences.

<table>
<thead>
<tr>
<th>Site</th>
<th>J</th>
<th>$r_5$</th>
<th>beta</th>
<th>ST error</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.86</td>
<td>4.06</td>
<td>0.36</td>
<td>1.217071</td>
<td>0.295792</td>
<td>0.777359</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>3.7</td>
<td>0.09</td>
<td>0.318726</td>
<td>0.282374</td>
<td>0.787145</td>
</tr>
<tr>
<td>4</td>
<td>0.77</td>
<td>1.91</td>
<td>0.36</td>
<td>1.774444</td>
<td>0.202881</td>
<td>0.845933</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>8.14</td>
<td>0.26</td>
<td>0.62078</td>
<td>0.418828</td>
<td>0.689921</td>
</tr>
</tbody>
</table>

$=AV32/(AU32*(1-$AT$27)/6)^0.5$

$8.080742 / 0.018335$

$=AT27*6/((1-AT27)^5)$
6.2 Advices for using multiple regression:

- First of all, multiple regression is a tool for testing predefined hypotheses. It is not designed for hypothesis generating. Hence the first thing you must do is to formulate a sound model of how your variables might be related. A blind pattern seeking will often lead to erroneous results because with a sufficient number of variables and small sample sizes you will very often if not always find some ‘significant’ dependencies.

- Think hard about cause and effect. In many cases it is not obvious which variable is the dependent and which the independent. In many cases a path diagram helps in formulating logical hypotheses.

- All variables must have a metric scale.

- Prior to computation check the correlation matrix of all single correlations. Eliminate independent variables that will not explain any part of total variance because they fully or nearly fully depend on another variable (multicollinearity). As a rule eliminate them if they are correlated with another independent variable with $r > 0.9$.

- Check for non-linear dependencies and try to eliminate them. Often logarithmic rescaling of some or all variables gives better results.

- In cases of heavy violations of the assumptions of the GLM use non-parametric regression. In its simplest form use ranked variables.

- The number of observations $N$ must be sufficiently large. $N$ should be at least twice the initial number of variables in the model.

- After computing the regression check for autocorrelation (using the Durbin-Watson statistic; $D$ should be between 1 and 3) and inspect visually the distribution of residuals. These should not show any regularities. Patterns in the distribution of residuals indicate violations of the general linear model.

- At small data set size use always Bonferroni corrected significance levels

- Use a stepwise analysis and eliminate in a stepwise manner all variable that do not significantly explain parts of total variance.

- Check the resulting multiple regression equation whether it is sound and logical.

- Try to check the resulting equation using other data. This leads often to an improvement of the original model.

- If you have a large number of observations (data sets) divide the data set prior to computation at random into a part with which you generate your multiple regression equation and a part with which to test this equation. If your test leads to unsatisfactory results you have to reformulate the whole model and start again. This means often that you have to include other previously not considered variables.

- A last step should be the verification of the model by other independent data sets.
6.3 Path analysis and linear structure models

Multiple regression requires a predefined model that shows relationships between variables. At least it must be possible to define dependent and independent variables. How do we know which of the variables is dependent? We need some prior knowledge or assumptions about causal relations. With four predictor variables our model looks as follows. E denotes the unexplained part of total variance.

Path analysis tries to generalize our assumptions about causal relationships in our model. Now we also try to establish the causal relations between the predictor variables. We define a whole model and try to separate correlations into **direct and indirect effects**. A path analytical model might look as follows: $X_1$ to $X_4$ are the predictor variables, $Y$ is the predicant (the dependent variable) and $e$ denote the errors (the unexplained variance) of each single variable. The aim of path analysis is now to estimate the correlation coefficients between all variables and the amounts of unexplained variances $e$. By modifying the path analytical model the method tries to infer an optimal solution where a maximum of total variance is explained. **Hence, path analysis tries to do something that is logically impossible, to derive causal relationships from sets of observations.**

Path analysis might be a powerful tool. However, it demands very thorough model development and large data sets. The smaller a data set is the higher is the probability to get a significant global solution together with significant partial correlations. Path analysis is also not unequivocal. Very often you find more than one solution that fit well to the data.

Path analysis is largely based on the computation of partial coefficients of correlation. But you have to know which variables have to be eliminated. This requires thorough prior modelling efforts. Assume the following case with 4 variables. We model the interrelationships between these variables through the following path diagram. The aim of the path analysis is now to infer the relative strength of correlations between these variables. This could be done by computing the appropriate partial correlations. Path analysis instead uses standardized variables instead of raw data. That means all variables are Z-transformed having means of zero and variances of 1. The fundamental theorem of path analysis tells now that is is always possible to compute the correlation coefficients of standardized variables from the total set of simple correlation coefficients. In path analysis these correlation coefficients of standardized variables are called **path coefficients** $p$.

We see that the results of path analysis can’t be better than the underlying model. **Path analysis is a model confirmatory tool. It should not be used to generate models or even to seek for models that fit the data set.**

How to compute a path analysis? The model above contains four variables that are presumably con-
nected. We get three regression functions

\[
\begin{align*}
W &= p_{xw} X + e \\
X &= p_{xy} Y + e \\
Z &= p_{zx} X + p_{zy} Y + e
\end{align*}
\]

This is a system of linear equations. We need a set of data for each variable and the equations are then solved by linear regression. However, we can make things easier. The input values in path analysis are already Z-transformed variables. Hence

\[
\begin{align*}
Z_w &= p_{xw} Z_x + e \\
Z_x &= p_{xy} Z_y + e \\
Z_z &= p_{zx} Z_x + p_{zy} Z_y + e
\end{align*}
\]

\[
\begin{align*}
Z_w Z_y &= p_{xw} Z_x Z_y + e Z_y \\
Z_x Z_w &= p_{xy} Z_y Z_x + e Z_w \\
Z_z Z_w &= p_{zx} Z_x Z_w + p_{zy} Z_y Z_w + e Z_w \\
Z_x Z_z &= p_{xy} Z_y Z_x + e Z_x \\
Z_y Z_x &= p_{xy} Z_y Z_x + e Z_y \\
Z_z Z_y &= p_{zx} Z_x Z_y + p_{zy} Z_y Z_y + e
\end{align*}
\]

In a first step we multiplied the three equations to get all combinations of \(Z_{nm}\). Now remember equation 6.6. The sum of all \(Z Z_m\) gives \(r_{nm}\), the coefficient of correlation. Summing over all cases gives a relation between path coefficients and coefficients of correlation. Because a simple sum of all Z-values is zero the error terms vanish and we have systems of linear equations that can be solved using ordinary matrix algebra. At the end we get a path diagram as shown in the Fig. above.

Linear path analysis cannot only handle with observable or observed variables. It can also be used to deal with complex and unobservable, so-called endogenous, variables. Look at the next example. Six exogenous measured variables can be divided into two groups. \(X_1\) to \(X_4\) are related and contain a common source of variance. This variance can be identified and we define a model with the intern endogenous variable \(Y_1\). In the same way define \(X_5\) and \(X_6\) a variable \(Y_2\). \(Y_1\) and \(Y_2\) explain a part of the variance of the variable \(Z\). All relationships can be defined by path coefficients. In this case we use path analysis for a whole model with measurable and latent variables.
variables. The latter type of variable can be identified by factor analysis, with which we deal later. To do so we need linear structure models and this technique is best known under the name LISREL (Linear structure models).

In the biological sciences this technique is seldom used, mainly due to the many difficulties in interpreting the results. It needs even more thorough a priori modelling effort. In biology the number of variables is in most cases rather limited (less than 10). For \( n = 10 \) we need \( 10 \times (10-1)/2 = 45 \) correlation coefficients. For estimating these at low error levels we need at least 10 independent observations for each variable pair. In praxis for even moderate large models hundreds of data sets are necessary. Therefore biologists more often rely on multiple regression in combination with factor analysis than on LISREL. In Statistica the SEPATH module is a powerful tool for developing structured models.
6.4 Logistic regression

At the end of this lecture we have to deal with still another form of regression. Bivariate or multiple regression needs continuous variables to be run. But often our data are of a discrete type or we have mixed data. Is it possible to run a regression with such data types? Look at the next Figure 6.4.1. Four species of closely related plants were studied at five stands (the colored bars) and their aboveground biomasses compared. An analysis of variance resulted in a highly significant difference (F = 58, p(F) < 0.001). But we can also try a simple linear regression between species (coded from 1 to 4) and biomass. We get a sufficiently good regression. But this regression model would not be able to predict a species from its body weight. A body weight of 2.5 would point either to species B or C. We might try to solve the above regression equation for x and get x = (2.0 + 1.05) / 1.42 = 2.5. This is exactly between 2 and 3. From the regression we are not able to decide unequivocally whether certain biomasses point to one or another species.

But we can solve our problem. This statistical solution is the so-called logistic or logit regression. The logic behind this type of regression is very simple and best explained by the next example. In archaeological research it is sometimes difficult to separate the sexes from cranial or postcranial skeleton rudiments because for most traits sexes highly overlap. Assume, for instance, that we have a series of skeleton fragments and want to infer the sex. To do this we first have to have a model. We generate such a model from a series of skeleton samples with known sex and perform a multiple regression. But our dependent variable, the sex, has only two values, male and female. Multiple regression, however, assumes the dependent variable to be continuous. We have to transform it and we do this via the already known logistic function. The logistic function has the form

\[ Z = \frac{e^y}{1 + e^y} = \frac{1}{1 + e^{-y}} \]

Our multiple regression has the general form

\[ Y = a_0 + \sum_{i=1}^{n} a_i x_i \]

To derive an appropriate model we start with odds, that means the quotient of the probabilities O = p/(1-p). Odds have values from 0 to \( \infty \). \( \ln(O) \) in turn goes from \( -\infty \) to \( \infty \).

Now we use a simple multiple regression to estimate \( \ln(O) \) from or predictors. Hence
We got the general *logistic regression* equation where logit stands for log odds.

\[
\ln \left(\frac{p}{1-p}\right) = a_0 + \sum_{i=1}^{n} a_i x_i \rightarrow \frac{p}{1-p} = e^{a_0 + \sum_{i=1}^{n} a_i x_i} \rightarrow p = \frac{e^{a_0 + \sum_{i=1}^{n} a_i x_i}}{1+e^{a_0 + \sum_{i=1}^{n} a_i x_i}}
\]

(6.4.2)

This is a *logit regression*. The Figure 6.4.2 shows how the logit regression transforms the y value into a range between 0 and 1. We see a threshold where an unequivocal decision is not possible but for most regression values we should be able to decide whether a given set of morphological traits stems from a male or a female skeleton. The next table shows values of morphological traits together with the sex of the person from whom these values were measured. Statistica computes for this the following regression function

\[
Z = \frac{e^{a_0 + \sum_{i=1}^{n} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{n} a_i x_i}}
\]

The result is shown in Fig 6.4.3. Of 20 samples 14 allow an unequivocal decision to be make. In five cases a clear decision is impossible and there is one obviously wrong result. From this we conclude that the power of our regression model is about 95%. If we now have a set of samples from a skeleton of unknown sex, we might use our regression model to establish the sex. The probability that we choose the right sex is about 85% (17 / 20).

The logit transformation transforms the data into the range between 0 and 1. This is the same range a probability can take. We might therefore interpret the result of a logit regression as a probability that a certain set of data belongs to one of both types of Y.

Fig. 6.4.4 shows a comparison of our logit result with an ordinary multiple regression. We see that the multiple regression to a certain extent also finds out males and females. But the result is much less decisive than before. There are only seven unequivocal decisions, four results are erroneous. We also see that the errors in both models differs.
6.5 Assessing the goodness of fit

Look at the following plot. Our problem is to fit a model to a set of 15 observations. The question is what model fits best. Spreadsheet programs offer some simple build-in functions (a power, a logarithmic, an exponential, and linear and higher order algebraic functions (up to the 5th order). Often we need other models and then we have to use the non-linear regression modules of statistic packages. They use least square algorithms to fit self-defined models. However, how to assess what model fits best? Of course we might use the fraction of explained variance, the coefficient of determination $R^2$. In the example of Fig. 6.5.1 we get: linear $R^2 = 0.71$; quadratic $R^2 = 0.83$; power function $R^2 = 0.82$. Which model fits best? We can’t decide. Judged only by the $R^2$ values the quadratic might look best. In fact, the data points were generated by a linear model that fitted worst. This is an often met problem and in fact there is no general solution to the problem of finding the best model. $R^2$ is not a very sensitive measure of goodness of fit because it is easy to get high values with different models. If possible we have to enlarge the number of data points.

Even worse is the situation if we have to compare multivariate models. For instance we have to compare results of multiple regressions that contain different numbers of predictor variables. Of course the higher the number of predictors the higher will be the variance explained. Thus, the problem is to find the optimum number of predictors. One criterion in multiple regression is that we should eliminate all variables that do not significantly contribute to the overall $R^2$ value. In our case we can only leave the quadratic or the linear term in the quadratic model. To decide how many parameters to leave we can use the so-called Akaike information criterion (AIC). It gives the amount of information explained by a given model and is defined as

$$AIC = -2 \ln(L) + 2K$$

(6.5.1)

where $L$ is the maximum likelihood estimator of goodness of fit and $K$ the number of free parameters of the model. In the example above the linear and the power function models have two free parameters. The quadratic model has three free parameters. $L$ can be any appropriate measure of fit. Often $R^2$ is used. $K$ is often identical with the number of free variables (predictors) of the model. The lower AIC is, the better fits a given model to the data. For comparison of two models we can take the difference $|AIC_1 - AIC_2|$. This difference is normally distributed and we assume therefore that differences are significant at $p < 0.05$ if $\Delta AIC$ is larger than 1.96. For the example above we get $\Delta AIC$ (quadratic - power) = 1.96. $\Delta AIC$ (linear - power) = 0.27. Therefore we interpret that the power function with only two free parameters fits not significantly better than the linear model.

Another important criterion is the Bayesian information criterion (BIC) (sometimes termed Schwarz criterion)

$$SC = N \ln(L) + K \ln(N)$$

(6.5.2)

where $N$ is the number of data points. This models points in our case unequivocally against the quadratic func-
tion. ΔBIC (quadratic - power) = 2.96. ΔBIC (linear - power) = 0.27.

An important point is that AIC or BIC must not be used in parallel with $R^2$ as the measure of goodness of fit. Both methods rely on different philosophies. For instance in multiple regression variable reduction via AIC and via $R^2$ will result in different final models. $R^2$ of the AIC model is now not longer a measure of explained variance.

For detailed information about these two criteria and other less often used see http://www.aps.uoguelph.ca/~lrs/ANSC637/LRS16/.
7. Analysis of variance

Consider an experiment designed to infer the influence if light intensity on plant growth. Four experimental plots (A to D) are installed each with five replicates. The four plots differ in total light energy supply. After the experiment we measure total plant biomass in each of the 25 experimental plots. The raw data are given in the Table beside. The question is now: does light intensity influence total biomass? Of course we could take our t-test and compare all five plants separately (giving 10 single comparisons). But it is possible that none of the pairwise comparisons gives a significant result although the treatments differ due to a trend in the data set. It would be better to treat all groups simultaneously. This is the aim of an analysis of variance. The idea behind the ANOVA is very simple. It was developed by the biostatistician and genetic Sir Ronald Fisher (1890-1962) especially for the biological sciences.

The total variance of the whole data set is given by

\[
\sigma_{\text{total}}^2 = \frac{1}{N-1} \sum_{k=1}^{N} (x_k - \bar{x})^2
\]

\[
= \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2
\]

k denotes the number of groups (treatments). N is the total number of data points (the sample size) and \(n_i\) denote the numbers of data points within each treatment. For simplicity the ANOVA uses sums of squares SS = \(s^2 \cdot (n-1)\). Hence

\[
SS_{\text{total}} = \sum_{i=1}^{k} \left( \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{\text{total}})^2 \right)
\]

(7.1)

Now we divide the total variance of all 25 plots into two groups. One group contains the variance within the 4 groups, the second group the variance between the groups (computed from the grand mean). Under the assumption that the within group variability is caused by the same set of unknown variables (that operate within the groups in a similar way) any difference in the between group variability should be caused by the treatment. In a next step we compare both variances. We know already a way to compare variances. It is the F-test of Fisher.

\[
F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}
\]

(7.2)

How to calculate within and between group variances? We compute

<table>
<thead>
<tr>
<th>Plot</th>
<th>Light intensity</th>
<th>Biomass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>B</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>13</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>18</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>C</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>D</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>D</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>10</td>
</tr>
<tr>
<td>21</td>
<td>E</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>E</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>E</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>E</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>E</td>
<td>4</td>
</tr>
</tbody>
</table>
\[
SS_{\text{between}} = \sum_{i=1}^{k} n_i (x_i - x_{\text{total}})^2
\]  

(7.3)

We sum over the k groups. Why multiplying with \( n_i \), the number of data points in each treatment? To compute the between variance we assume that the total variance comes solely from the differences between the treatments. Hence, we replace the within values through the within mean. The new total sum of squares is then according to eq. 7.3

\[
SS_i = \sum_{j=1}^{n_i} (x_{i,j} - x_{\text{total}})^2 = n_i (x_i - x_{\text{total}})^2
\]

Now we need the degrees of freedom. If we would in our example divide \( SS_{\text{between}} \) through the group number (4) we would underestimate the total variance. Why? Because if we have means of three groups, the variance computed by including the fourth group is no longer a random variable. It is already determined by the other three values. We have to divide through 3, the number of variables that can fluctuate at random. In other words, we have to divide through the number of degrees of freedom. This is for the between group variance \( df_{\text{between}} = k - 1 \) if \( k \) denotes the number of groups.

The within group variance is given by

\[
SS_{\text{within}} = \sum_{i=1}^{k} \left( \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2 \right)
\]

(7.4)

\( SS_{\text{within}} \) is therefore the sum of the variances computed within the groups. This is the part of variance not influenced by the treatment. It is introduced by other (often unknown or not measured) factors and is often also called \( SS_{\text{error}} \). If \( SS_{\text{within}} \) would equal \( SS_{\text{between}} \) the treatment would not influence the dependent variable. Otherwise \( SS_{\text{between}} \) would introduce an additional part of variance. Variances of independent random variates are additive. We get a fundamental equation on which the analysis of variance is based.

\[
SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}}
\]

(7.5)

For the degrees of freedom a similar equation holds. We have \( df_{\text{within}} = N-k \) degrees of freedom. Hence.

\[
df_{\text{total}} = df_{\text{between}} + df_{\text{within}}
\]

(7.6)

In our case equation 7.6 gives \((25-1) = (4-1) + (25-4)\). This is immediately clear. If we would compute the total variance (computed from all cases) directly we would have to divide through \( n-1 = 24 \). \( SS_{\text{total}} \) has 24 degrees of freedom.

Our F-test looks therefore as follows

\[
F = \frac{SS_{\text{between}}}{SS_{\text{within}}} = \frac{N-k}{k-1} \frac{SS_{\text{between}}}{SS_{\text{within}}}
\]

(7.7)

<table>
<thead>
<tr>
<th>Variable</th>
<th>df Effect</th>
<th>MS Effect</th>
<th>df Error</th>
<th>MS Error</th>
<th>F</th>
<th>p(F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light intensity</td>
<td>4</td>
<td>91.86</td>
<td>20</td>
<td>5.84</td>
<td>15.72945</td>
<td>5.72E-06</td>
</tr>
</tbody>
</table>
Statistical programs give now the following output (in this case it would even be faster to compute F by hand).

The variable light intensity has 4 degrees of freedom. *Statistica* denotes MS (mean sums of squares) instead of SS. MS = SS / dF. MS effect is the between group variability, MS error the within group variability. F is the quotient of MS_effect / MS_error = 15.7. The corresponding probability level is far below 1%. In other words, we conclude that light intensity has a statistically significant effect on plant growth.

A numerical example how to compute an ANOVA is shown in the next table. A standard matrix contained data from 4 treatments with 5 observations. Each gave four group means. Do these group means differ?

SS_between and SS_within are computed separately and compared with SS_total. You see that the sum of SS_between and SS_within is indeed identical to SS_total.

The ANOVA can easily be extended to deal with more than one influencing variable. With the same reasoning as above we can divide the total variance into parts where each part represents the fraction of vari-

### 7.1 Advices for using ANOVA:

- You need a specific hypothesis about your variables. In particular, designs with more than one predictor level (multifactorial designs) have to be stated clearly.

- ANOVA is a hypothesis testing method. Pattern seeking will in many cases lead to erroneous results.

- Predicctor variables should really measure different things, they should not correlate too high with each other.

- The general assumptions of the GLM should be fulfilled. In particular predictors should be additive. The distribution of errors should be normal.

- As in multiple regression it is often better to use log-transformed values

- In monofactorial designs where only one predictor variable is tested it is often preferable to use the non-parametric alternative to ANOVA, the **Kruskal Wallis test**. This test does not rely on the GLM assumptions but is nearly as powerful as the classical ANOVA.

- A **non-parametric alternative** for multifactorial designs is to use ranked dependent variables. You loose information but become less dependent on the GLM assumptions.

- ANOVA as the simplest multivariate technique is quite robust against violations of its assumptions.
The next Table gives our plant data with an additional variable, nutrient availability, included.

Now, we have a \textbf{two-factorial design}. We compute the total sums of squares and the sums of squares between light intensity and between nutrient availability. This is already quite time consuming to do it by hand. But our statistic package gives immediately the following output.

We have MS effect light intensity, nutrients and the combined effect of light and nutrients. Again, we have the within variance MS error, which is the same for all three combinations of effects. The quotients of MS$_{\text{effect}}$ and MS$_{\text{error}}$ is again our F-value. We see that only light has a significant effect. Nutrient availability and the combined effect of light and nutrient do not significantly contribute to total variance. Of course, we have always to take care that sums of squares can be computed. So, for each combination of variables at least two data points must be available. Otherwise our experimental design would be incomplete.

What else has to be remembered? To answer this question we have to deal with the relationships between the variables in more detail. We saw already that variances of a set of variables can be combined in a linear manner

\begin{equation}
\sigma^2_{\text{total}} = \sum \sigma^2_i + \sum \text{cov}_{i,j}
\end{equation}

(7.8)

If we now define one of the variables Y as an dependent variable we would be able to express the variance of Y from the variances of the other independent variables

\begin{equation}
\sigma^2_Y = \sum \sigma^2_{i\neq Y} + \sum \text{cov}_{i,j\neq Y} + (\sigma^2_{\text{total}} - \sum \text{cov}_{i,j}) = \sum \sigma^2_{i\neq Y} + \sum \text{cov}_{i,j\neq Y} + \sigma^2_{\text{Error}}
\end{equation}

(7.9)

In other words, the variance of the dependent variable can be expressed by the sums of variances and covariances of the independent variables and a term that contains the variances introduced or co-introduced by Y. This is our within group variability. The aim of an ANOVA is therefore to compute these variance from the data set and to compute all respective quotients $\sigma^2_i / \sigma^2_{\text{Error}}$ and $\text{cov}_{i} / \sigma^2_{\text{Error}}$.

Additionally, we know already that if we can write the variances in such an additive form, we have the same relation for the means.

\begin{equation}
\mu_Y = \sum \mu_i + \mu_{\text{Error}}
\end{equation}

(7.10)

The last two equations are fundamental and form the basis of most multivariate tests that rely on
the general linear model.

An ANOVA can be used as an alternative to the t-test! Assume you have only two groups A and B. **Then the F-test is identical to a t-test for independent variables.** Some researchers prefer the F-test over the t-test for pairwise comparisons.

ANOVA and multiple regression are very similar. Both rely on the GLM and computations of variances and covariances. It takes therefore no wonder that necessary or optimal samples sizes are also similar. **As in the case of multiple regression you have to distinguish between test and experiment wise error rates and you should use Bonferroni corrections.**

In the previous case we dealt with only one independent variable (or predictor variable), the light intensity, plant biomass was the dependent variable. The ANOVA can easily be extended to deal with more than ONE predictor variable. In this case we speak of **multifactorial designs.** With the same reasoning as above we can divide the total variance into parts where each part represents the fraction of variance introduced by a certain variable. The next Table gives our plant data with an additional variable, nutrient availability. Now, we have a two factorial design. Equation 7.5 modifies to

\[
SS_{total} = SS_A + SS_B + SS_{AxB} + SS_{error}
\]

The degrees of freedom come from

\[
df_{total} = df_A + df_B + df_{AxB} + df_{error}
\]

\[
kmn = (k-1) + (m-1) + (k-1)(m-1) + kn(n-1)
\]

where \(k\) and \(m\) are the numbers of Categories In A and B and \(n\) is the number of cases.

We have the variance part introduced by the factor light (\(SS_A\)) only, by the additional factor nutrients (\(SS_B\)) only, by the combined effect of light and nutrients (\(SS_{AxB}\)), and of course by the within variance (\(SS_{error}\)). \(SS_A\) is computed separately for both groups of nutrients to exclude the influence of nutrients. \(SS_B\) is computed separately for the four groups of nutrients to exclude the influence of light intensity. Additionally we need the \(SS_{AxB}\). Under the hypothesis that light and nutrients are not independent the combined action of both variables might influence the plant biomass.

Be careful. For every treatment of light intensity you have to have at least two treatment cases of nutrients. Otherwise it would be impossible to compute the within variability of nutrients inside the treatment light. However in reality two cases are much too less. Even for an approximated estimate of the within variance you need at least 6 treatments for each combination of light and nutrients. Hence, you need at least \(4\times2\times6 = 48\) cases. The number of experimental data necessary for multifactorial designs raises very fast.

We computed the total sums of squares and the sums of squares between light intensity and between nutrient availability. This is already quite time consuming to do it by hand. But our statistic package gives immediately the output shown above. The *Statistica* output shows separate results for the variables.
Statistical advices

(treatments) light (effect 1) and nutrients (effects 2). These are called the main effects. Additionally the combined effect of light and nutrients is given, the secondary or combined effects. Again, we have the within variance MS error, which is the same for all three combinations of effects. The latter are computed over all groups of combinations of the main effects. The quotients of MS effect (MS_{between}) and MS error (MS_{within}) is again our F-value. For light I used the same data as in the monofactorial design. We see that including a new treatment variable changed the significance levels for light too. Now light appears to be less significant.

Nutrient availability and the combined effect of light and nutrient do not significantly contribute to total variance. Of course, we have always to take care that sums of squares can be computed. So, for each combination of variables at least two data points must be available. Otherwise our experimental design would be incomplete. The table beside shows such a typical incomplete design. We have one metrically scaled independent variable and three predictors. However, a full ANOVA including all three predictors is impossible to compute.
because not all predictor combinations occur. We will deal with incomplete designs in chapter 14.

Now comes one example of the category how to lie with statistics. The next *Statistica* table contains the result of an ANOVA with three effects (VAR2, VAR3, and VAR4). We see that VAR2 and the combined effect of all predictors are statistically significant at the 5% error level. However, as independent variables I used simple random numbers. There shouldn’t be any significant results. The probability for a significant result should be 1 out of 20 = 5%). What has happened? We used an error level $\alpha$ of 5% for the whole ANOVA result. The probability to make no error is $1 - \alpha$. The probability to make no error in $n$ experiments is therefore $(1 - \alpha)^n$. What is then the probability that in $n$ experiments we reject erroneous $H_0$, that is that we make a type II error? This probability is $\beta = 1 - (1 - \alpha)^n$. In our above example we have 3 independent variables and their combinations. Our probability level is therefore not $\alpha = 0.05$. It should be between $\beta = 1 - (1 - \alpha)^3 = 0.14$ and $\beta = 1 - (1 - \alpha)^7 = 0.30$. Hence for single contrasts we get a new error level of 14 to 30% and we expect to make an erroneous decision in 1 out of 6 to 1 out of 3 cases. Our level of $\alpha = 0.05$ is the test wise error rate. But if we use a test several times with the same set of data or compute a multivariate test with many variables simultaneously we have to deal with the experiment wise error rate. Again we have to correct our initial significance level by the following equation

$$\alpha^* = 1 - (1 - \alpha)^{1/n} \approx \frac{\alpha}{n}$$

(7.13)

The Bonferroni correction reduces the test wise error level to reach in an acceptable experiment wise error level. You divide your initial error level through the number of independent variables, or, if you want to be on the sure side, through the number of combinations of these variables. In our above case we should therefore only accept significance levels $p(F)$ below $0.05 / 7 = 0.007$ for single comparisons to reach in an experiment wise error level of 5%.

What is the necessary sample size for the within sums of squares $SS_{error}$? In other words what is the equivalent of a power analysis in the case of an ANOVA. We can compute this by a similar t-test model as above. We define the effect size by the maximum difference of means of the groups. We get

$$t \approx \sqrt{\frac{N \text{ effect size}}{2k \sigma_{\text{within}}}}$$

Hence

$$N = 2k \left( \frac{t \sigma_{\text{within}}}{\text{effect size}} \right)^2$$

(7.15)

where $k$ is the number of groups and $N$ is the total sample size (the total number of observations). Again, we need information about the within group variability.

Now will deal with two non-parametric alternatives to the ANOVA. Assume we have an observational series with a single discrete independent variable, a monofactorial design. Then, we can use the so-called *Kruskal-Wallis test* or *Kruskal-Wallis ANOVA by ranks*. In principle, the Kruskal-Wallis is computed in the same way as the U-test. We compute the total rank sums as the base for a test statistic. Our test statistic is
where $N$ is the total number of observations, $n_i$ the number of observations of group $i$, and $k$ the number of groups. The Table shows an example how to compute a Kruskal-Wallis test with Excel. We have a series of 53 observations divided into 4 groups or treatments. The KW-value of 9.15 has again to be compared with tabulated ones. For values of $r > 5$ or larger sample sizes KW is approximately $\chi^2$ distributed and values can be taken from a $\chi^2$ table with $r-1$ degrees of freedom. In our case this value for 3 degrees of freedom and $\alpha = 0.95$ is $\chi^2 = 7.81$. Our value of 9.15 is larger than 7.81 and we accept the hypothesis that there is a difference between the treatments at an error level of 5% (the exact error level given by Statistica is 2.73%). The Kruskal-Wallis test has in monofactorial designs nearly the same power as an ordinary ANOVA and should be used in all cases where we are not sure about the underlying frequency distributions of the populations.

A second alternative is to use a rank order ANOVA. In this case we apply an ordinary ANOVA but use for the dependent metrically scaled variable ranked data. Such a rank ANOVA is applicable in monofactorial and multifactorial designs. An example is shown in the next table. We have 29 data sets grouped into four effects. Now we rank these data from the largest (0.929 = rank 1) to the smallest (0.082 = rank 29) using the Excel function `RANK`. The Excel formulas for the KW-value are shown in the example.
We compare the results of ordinary ANOVA, ANOVA with ranked data and the KW-test. The ANOVA gives a probability level for $H_0$ of $p(F) = 0.007$, the ANOVA with ranked data proposes $p(F) = 0.014$, and the Kruskal Wallis test returns $p = 0.024$. All three tests point to significant differences between the four groups. The non-parametric alternative returns a higher significance level. This is caused by the loss of information due to the ranking of the data. But the probability to make a type I error (to accept the wrong hypothesis) is also reduced.

A simple ANOVA is easy to compute. However, variance analytical techniques can be greatly extended. A first extension regards the inclusion of metric and categorical variables. In this case we perform a covariance analysis (ANCOVA). In an ANCOVA we use the residuals of a multiple regression of a dependent variable on independent metrically scaled variables (the covariates) as the test variable in a ordinary ANOVA. This is shown in the next plant example. We studied plant growth on different soils in dependence on light intensity (low/high). A simple ANOVA pointed indeed to Light as a factor that influenced plant growth. We further assume that growth is also dependent on the amount of nutrients in the soil. Fig. 7.1 shows indeed a general regression of growth on nutrients.
However for each light intensity class there is a different regression (the colored lines). Now we calculate the residuals of the single regression and the different regressions of each treatment. However using the different regressions of each treatment to calculate the residuals would include the effect of the treatment through the backdoor. Indeed the respective ANOVA gives $F = 0$ in the Statistica example.

The ANCOVA with nutrients as covariate now tells that light has only a marginal influence on growth after correcting for the effect of nutrients. Using the residuals of the grand regression between nutrients and growth as the dependent variable in a simple ANOVA gives nearly the same result. The different significance levels result from the different degrees of freedom. The simple ANOVA does not know anything about the variable nutrient and $df_{\text{error}} = 27$. In the ANCOVA one $df$ goes to nutrients and $df_{\text{error}} = 26$.

The next example shows a bifactorial ANCOVA of the same plant example. After regressing growth against nutrients light intensity remains a significant predictor of plant growth. If we first regress growth on nutrients and use the residuals (ZMN4 in the Statistica example) in a simple ANOVA we get nearly the same result. Again the different significance levels stem from the different degrees of freedom. Covariance analysis has the same limitations and prerequisites than multiple regression and ANOVA. Both rely on GLM. The covariates have to be metrically scaled.

Another special case of ANOVA regards repetitive designs. For instance in medical research we test patients before and after medical treatment to infer...
the influence of the therapy. In this case we have to divide the total variance (SS\text{total}) in a part that contains the variance between patients (SS\text{between}) and within the patient (SS\text{within}). The latter can be divided in a part that comes from the treatment (SS\text{treat}) and the error (SS\text{error}) (Fig. 7.2). Hence we have one variance more to consider, the a priori differences between the patients that are not influenced by the medical treatment. Assuming a monofactorial design we calculate the necessary sums of squares from

\begin{align*}
SS\text{total} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x})^2 \\
SS\text{between} &= k \sum_{i=1}^{n} (\bar{P}_i - \bar{x})^2 \\
SS\text{within} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{P}_i)^2 \\
SS\text{treat} &= n \sum_{j=1}^{k} (\bar{T}_i - \bar{x})^2 \\
SS\text{error} &= \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{P}_i - \bar{T}_j + \bar{x})^2
\end{align*}

(7.17)

where \(x_{ij}\) is the measurement of patient \(i\) within treatment \(j\), \(\bar{x}\) is the grand mean, \(\bar{T}\) is the mean of the \(k\) treatments, and \(\bar{P}\) the mean of the \(n\) patients. SS\text{total} is of course the sum of these sums of squares.

\[SS\text{total} = SS\text{between} + SS\text{within} = SS\text{between} + SS\text{treat} + SS\text{error}\]

The degrees of freedom come from

\[df\text{total} = df\text{between} + df\text{within} = df\text{between} + df\text{treat} + df\text{error}\]

\[kn - 1 = n - 1 + n(k - 1) = n - 1 + k - 1 + (n - 1)(k - 1)\]

To test for significance we apply again the F test with \((k-1)\) and \((n-1)(k-1)\) degrees of freedom.
The Excel example at the next page shows how to calculate a repeated measures ANOVA. The blood pressure of ten test persons were measured in the morning, in the afternoon, and in the night. The question is whether blood pressures depends on the time of day. We first calculate $SS_{within}$ from the person means and the grand mean. Next we calculate $SS_{within}$ from the single measurements and the person means. $SS_{treat}$ comes from the mean of the three day times and the grand mean. Lastly, $SS_{error}$ comes from the measurement and all three means. We see that indeed $SS_{total} = SS_{within} + SS_{between}$ and $SS_{within} = SS_{treat} + SS_{error}$.

A special technique for repetitive designs is the use of so-called “ipsative” data. Instead of using raw data the data are transformed by subtracting the case (person) means. Hence $SS_{total}$ and $SS_{between}$ become zero. Therefore eq. 7.5 reduces to

\[
SS_{total} = SS_{between} = 0
\]
\[
SS_{within} = \sum_{j=1}^{k} \sum_{i=1}^{n} x_{ij}^2
\]
\[
SS_{treat} = n \sum_{j=1}^{k} \bar{T}_{j}^2
\]
\[
SS_{error} = \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{T}_{j})^2
\]

(7.19)

Used in this way makes the repeated measure ANOVA very easy to calculate even with Excel. The significance test reduces to

\[
F = \frac{SS_{treat}}{SS_{error}} \frac{df_{treat}}{df_{error}} = \frac{n(n-1)\sum_{j=1}^{k} \bar{T}_{j}^2}{\sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{T}_{j})^2}
\]

(7.20)

Of course, the above ANOVA scheme can be extended to many treatment variables.

The validity of the $F$ test for the previous ANOVA depends on a series of prerequisites. Particularly the within variances have to be similar (homogeneous). Further, in repeated designs treatments are often correlated. In the tables above morning might correlate with evening and night because the test persons react in the same way on the time of day. These correlations have to be similar for all combinations of treatment. Otherwise the treatment would be inhomogeneous. Inhomogeneities result often in inflated type I error probabilities. To correct for this possibility we might correct the degrees of freedom to obtain lower $F$-values and significance levels. In eq. 7.18 we have $(k-1)$ degrees of freedom for the effect and $(n-1)(k-1)$ degrees of freedom for the error. To correct for violations of the AOVA prerequisites we multiply these degrees of freedom by a factor $\epsilon$. At maximum heterogeneity
In the above example we have to run the F-test with $1/2 \times 18$ degrees of freedom for the error and $1/2 \times 2$ df for the treatment. This gives $p(F=14.324; 9, 1) = 0.202$. There is no significant difference between the treatments. Another possibility is of course to use a Bonferroni correction and to divide the test wise error level through the number of treatments: $0.05 = 0.05/3 = 0.017$. A third possibility is to use the conservative ANOVA with $(n-1)$ df error and only one df treat.

Up to know we dealt with such experimental designs where for all combinations of variables at least 2 cases were available (the variances could be calculated). In this case we speak of complete designs. However, often we have only incomplete designs where not for all factor combinations within variances exist.

$$\varepsilon = \frac{1}{k-1} \quad \text{(14.5)}$$

In the above example we have to run the F-test with $1/2 \times 18$ degrees of freedom for the error and $1/2 \times 2$ df for the treatment. This gives $p(F=14.324; 9, 1) = 0.202$. There is no significant difference between the treatments. Another possibility is of course to use a Bonferroni correction and to divide the test wise error level through the number of treatments: $0.05 = 0.05/3 = 0.017$. A third possibility is to use the conservative ANOVA with $(n-1)$ df error and only one df treat.
For instance, we study our already known plants within the following design. We have three light categories (low, medium, high) and four nutrient categories (P, N, Fe, Mn). We do not expect that light and nutrients interact. That means we ignore the interaction term light x nutrients. Therefore, instead of at least 24 parallels (3*4*2) we need much less. In the best case only eight. You don’t need all treatment combinations to be realized. In other words ignoring irrelevant interactions reduces the experimental effort. In the example beside we considered only the main effects and found only light to be significant.

Essentially, the incomplete design ANOVA proceeds similar to the ANOVA for complete designs. SS error are calculated in the same way as above. SS treat contains the between treatment variances of the factors.

A special case of incomplete designs are nested designs. In nested designs occur treatments of one factor only with certain treatments of the other factor. An example gives the next table. Phosphorus occurs only in the soils of the experiments that got much light, nitrogen is associated with the medium light treatment and Fe and Mn with the low light treatments. Within the variance components module of Statistica we use the nested design option and get the output shown above. Again only light appeared to be significant.

Lastly we look at the F-test from the perspective of a multiple regression. A simple ANOVA F is defined as

\[
F = \frac{SS_{\text{treat}} (n - k)}{SS_{\text{error}} (k - 1)}
\]

R² of the multiple regression is defined as

\[
R^2 = \frac{\sigma^2_{\text{treat}}}{\sigma^2_{\text{total}}} = \frac{SS_{\text{treat}}}{SS_{\text{total}}}
\]

where \( \sigma^2_{\text{treat}} \) is of course the variance explained by the predictor variables. We further know that \( SS_{\text{total}} = SS_{\text{treat}} + SS_{\text{error}} \). Plugging this equation into the second gives

\[
SS_{\text{total}} = SS_{\text{treat}} + SS_{\text{error}} \Rightarrow SS_{\text{total}} R^2 = SS_{\text{total}} - SS_{\text{error}}
\]

\[
SS_{\text{error}} = (1 - R^2) SS_{\text{total}}
\]

Now we go back to the first equation and get

\[
F = R^2 SS_{\text{total}} (n - k) \frac{(1 - R^2)}{(k - 1)} = \frac{R^2 (n - k)}{(1 - R^2) (k - 1)}
\]

This equation is identical to eq. 6.12 (with \( m = k-1 \) and enables us to test the significance of a multiple and a bivariate regression.


7.2 Comparisons a priori and a posteriori

When planning experiments or using experimental data you have to be sure what you want to do. Do you intend to use experiments for pattern seeking to get new hypotheses or do you want to verify a priori defined hypotheses. In the first case you explain a result a posteriori with a new hypothesis, in the second case you accept or reject an a priori formulated hypothesis. Both processes have to be distinguished clearly. **Formulating and testing the same set of hypothesis within one experimental design is logically and mathematically not acceptable.**

For instance, you compare a strain of genetically modified bacteria with the wild form. You might use a t-test and find significant differences in metabolism rates. From this you formulate the hypothesis that the genetic modification influence metabolism rates. You used statistics to get an a posteriori hypothesis. But you did not prove this hypothesis using statistics. However, you might also have biochemical data at hand and predict that the genetic modification should cause different metabolism rates. You undertake an experiment and use a statistical test to verify your hypothesis. In the latter case the information gain by using the t-test is much higher then in the first case.

The distinction between a priori and a posteriori comparisons has profound implications for the use of certain test statistics. For a priori comparisons no corrections of significance values seem to be necessary. In the case of a posteriori comparisons instead our probability levels to make type I errors might be biased. We undertake pattern seeking. Our brain is constructed in such a way that it always tries to find patterns. Hence, the probability to see something is much higher than to see nothing. This is the reason for our inclination to all sorts of metaphysics or superstition.

An important method for the correction of $\alpha$-error levels in the analysis of variance is the **Scheffé test**. The Scheffé test verifies all pairwise comparisons of an ANOVA at the experiment wise $\alpha$-level. The test guarantees that no pairwise comparison can be significant at an error level higher than $\alpha$. 

8. Cluster analysis

Cluster analysis is one of the multivariate techniques that is not based on the assumptions of the general linear model. Cluster analysis is a method that does not test hypotheses. It is an intrinsically hypotheses generating method. Simply speaking a cluster analysis tries to classify a set of objects into groups (cluster or classes) according to a predefined measure of distance. However, there is no single technique. Cluster analysis is a whole set of very different methods and solutions.

Consider the following example. We take a matrix containing ten variables (objects) and 20 observations (cases). Then we start Statistica, copy the data into the cluster analysis module and perform a default analysis. The result (the Figure beside) is a dendrogram that visualizes the distances of the variables from each other according to a linkage algorithm. We are ready at hand with an interpretation. An a posteriori hypothesis generating process.

But what did Statistica really do? First of all it applied a measure of distance. The program measures case wise differences and computes from them a common distance value between two variables. It then tries to link variables according to these distances. The Jaccard measure of similarity is one measure of distance between two objects. It is defined as

\[
J = \frac{a}{b+c}
\]

(8.1)

where a is the number of common elements and b and c the total numbers of elements in object 1 and 2.

It is possible to generalize the Jaccard index to get an equation that contains nearly all measures of similarity as special cases

\[
J = \frac{a + \epsilon e}{a + \epsilon e + \delta (b+c)}
\]

(8.2)

where a is the number of common elements, b and c are the number of elements that occur only in object 1 or 2 and e is the number of elements that neither occur in 1 nor in 2.

The best known measure of distance for metrically scaled variables is of course the Euclidean distance.

\[
d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}
\]

(8.3)

where p is the number of data sets and \(x_{ik}\) and \(x_{jk}\) the values of i,j at data point k.

The most important alternative to Euclidean distance is the so-called city block or Taxi driver distance that measures the shortest way to get from point i to point j. It is defined from the general Minkowski metric

\[
d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{1/r}
\]

(8.4)
r = 1 defines the Taxi driver distance, r = 2 the Euclidean distance. Another measure of distance for metrically scaled variables is the use of pairwise correlation coefficients or the Bray-Curtis measure dealt with in Chapter 5.

More important for the outcome of a cluster analysis is the choice of the linkage algorithm. The basic procedures of such algorithms can be exemplified from the single linkage algorithm (Fig. 8.2). Starting with an association matrix we order the pairs of sites according to the degree of association. Then step by step objects are including into cluster that have a higher association than predefined. The first group is made of BD, the second group includes C because D and C are the next with respect to association. Lastly we include A. Single linkage has some undesired properties. It tends to produce a large number of clusters and has problems if three or more objects have the same distances to each other.

The Figure 8.3 shows the most important linkage algorithms. The exact form of the algorithms are given in most statistical textbooks. Important is that different algorithms frequently result in different clusters. The result of a cluster analysis relies therefore to a good deal on the choice of the linkage algorithm. In general
we have four types of algorithms

1. Sequential versus simultaneous algorithms. In simultaneous algorithms the final solution is obtained in a single step and not stepwise as in the single linkage above.

2. Agglomeration versus division algorithms. Agglomerative procedures operate bottom up, division procedures top down

3. Monothetic versus polythetic algorithms. Polythetic procedures use several descriptors of linkage, monothetic use the same at each step (for instance maximum association).

4. Hierarchical versus non-hierarchical algorithms. Hierarchical methods proceed in a non-overlapping way. During the linkage process all members of lower clusters are members of the next higher cluster. Non hierarchical methods proceed by optimization within group homogeneity. Hence they might include members not contained in higher order cluster.

There are some general properties of the linkage algorithms.

1. The **single linkage algorithm** uses the minimum distance between the members of two clusters as the measure of cluster distance. It favours chains of small clusters.

2. The **average linkage** uses average distances between clusters. It gives frequently larger clusters. The most often used average linkage algorithm is the **Unweighted Pair-Groups Method Average (UPGMA)**.

3. The **Ward algorithm** calculates the total sum of squared deviations from the mean of a cluster and assigns members as to minimize this sum. The method gives often clusters of rather equal size.

4. **Median clustering** tries to minimize within cluster variance.
The Figure 8.5 shows the same set of eight variables clustered by four different linkage algorithms. The results look different. Only two constant patterns are detectable. Variables 1, 2, 3, and 6 cluster always together and variable 8 seems to form a single cluster. It is then a question of our interpretation of the results. Most popular are **UPGMA** and the **Ward algorithm** that gives in most cases rather equally sized partitions and therefore clusters.

The Ward algorithm differs somewhat from the others by using a minimalization criterion. Using Euclidean distance the Ward algorithm tries to minimize for each cluster \( r \) the distance \( d_r^2 \) of all elements of \( r \) from the centroid (the middle) \( c \) of the cluster.

\[
d_r^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - c_r)^2 \rightarrow \min
\]

One method to assess the ‘right’ number of clusters is the so-called **elbow test**. The next Table shows the agglomeration process of ten variables. The final clustering indicates seven singles cluster. However, some of them seem to be poorly separated. We plot the Euclidean distances against the agglomeration stage as shown below. From stage seven to eight a step (an elbow) occurs, where the dissimilarity suddenly raises. At lower stages dissimilarities are rather low. The elbow criterion tells now that we should accept only those clusters above the elbow. Hence we should accept a three cluster solution with cluster one containing Var2, Var4, Var5, and Var9, cluster two containing Var6 and Var8, and cluster three containing Var1, Var3, Var7, and Var10.

However, **there is no ‘correct’ clustering**. Each method favours different aspects of distance. **Hence,**
8.1 Advices for using a cluster analysis:

- The first point is the number of data sets. Especially in the case of samples where you want to get information about the whole population, sample sizes (data sets) must not be too small.
- Screen the data for outliers. They might have overproportional influence on the outcome and should be eliminated.
- As in other multivariate techniques, an initial sound hypothesis building process helps to interpret the results.
- It is a good practice to undertake in parallel a factor analysis and to compare the results with those of the cluster analysis. Sometimes (in the case of many variables) it is better to undertake first a factor analysis and use the factor loadings instead of the raw variables for the cluster analysis. By this technique, highly correlated variables are combined and the total number of variables is reduced.
- Even without a priori factor analysis, variables that correlate highly with others ($r > 0.9$) should be eliminated. It is of course our interpretation which of the two or more correlated variables to throw out.
- If all data are metrically scaled, I advise to standardize them by Z-transformation. Large differences in absolute values might bias the outcome. Try also log-transformations.
- Try out several linkage algorithms. Compare the results and accept only clusters that have been found by all or at least by the majority of algorithms.
- Do the results make sense? Compare them with your initial hypotheses.
- If you have a large data set of independent data, you might divide them into two randomly chosen parts. Use one part in the cluster analysis for hypothesis building and the second part for verifying the result by using the same algorithm.
- Cluster analysis is a method that is open for all types of data manipulation. Therefore, you must always give basic information about how you received a given result:
  1. Distance measure and linkage algorithm
  2. Explain your method choice. Tell especially, why you used a given linkage algorithm
  3. Tell whether your results are stable.

Although cluster analysis is a genuine hypothesis generating technique, we should again start with a sound and realistic hypothesis.
8.2 K-means cluster analysis

Ordinary cluster analysis groups a set of variables into a number of clusters. This number is the result of the clustering process. A different technique is the clustering of variables into a predefined number of clusters. This is done by the k-means clustering. k-means clustering starts with a predefined number of clusters and puts the elements step by step into these clusters. It seeks for elements that have distances from the mean of the own cluster that are larger than those to another cluster. The element is then shifted to this cluster. The process stops if all elements have found ‘their’ cluster. The k-means method does not result in a dendrogram. Instead the output contains the distances of each variable from the cluster mean. The results but can be visualized by distances of elements from cluster means. This is shown in the diagram beside. 20 Elements have been grouped into five clusters. They are defined by distances of cluster centres (the red arrows) and these are computed from the elements they contain (black elements).

A shortcoming of this method is that the final clustering might depend on the entry order of the elements into the clustering process. I advise therefore to run several analyses in which the order of variables is changed. Accept the solution that was found most often.
15.3 Neighbour joining

The most often used agglomerative cluster algorithm in phylogenetic analysis is **neighbour joining** that was particularly developed for phylogenetic analysis. At the beginning the method puts all elements into different clusters (Fig. 8.3.1). It then proceeds stepwise. It starts with a dissimilarity matrix based on a certain distance measure. (shown as different branch lengths in the Fig.). First, two clusters (elements) are chosen and combined in one cluster to get a new node in the tree (X in the Fig.). From this a new dissimilarity matrix (containing the elements A, X, D, E, F) is calculated and the method proceeds at step one. The third dissimilarity matrix contains the elements (A, X, Y, F). The method depends therefore on three criteria: the algorithm to choose the pair of elements to join, the algorithm to calculate the dissimilarity matrix and an algorithm to estimate branch lengths. In the simplest case dissimilarities for any element are calculated from

\[ \delta(X,Y) \]

\[ \delta(X,Y) = \sum_{\Delta} \delta(X,Y) \]  

(8.3.1)

\[ \delta(X,Y) \] are the dissimilarities for the elements X and Y calculated from the branch lengths. To select a pair of elements to be joined the method proceeds in calculating

\[ Q_{\delta(X,Y)} = (n-2)\delta(X,Y) - \Delta(X) - \Delta(Y) \] 

(8.3.2)

for all element pairs (X,Y). The pair with the lowest value of Q is selected for joining. Given that two elements A and B have been selected and joined to a new element \( U_{AB} \) the new dissimilarities are calculated from

\[ \delta(X, U_{AB}) = \frac{\delta(X,A) + \delta(X,B) - \delta(A,B)}{2} \] 

(8.3.3)

This is called the **reduction step**. After this step the methods proceeds in a recursive way starting again from the dissimilarities. Beside I show a simple example for four species. From an initial distance matrix dissimilarities (eq. 8.3.1) were calculated. The next step involves the selection of elements (eq. 8.3.2) and at last the reduction to a new distance matrix (eq. 8.3.3). This procedure is then repeated for the new matrix. At the end the methods joins vertebrates and Protostomia.

We also need the branch lengths of U to A and B. These
are given by

\[
\begin{align*}
\delta(A, U) &= \frac{(n - 2)\delta(A, B) + \Delta(A) - \Delta(B)}{2(n - 2)} \\
\delta(B, U) &= \frac{(n - 2)\delta(A, B) - \Delta(A) + \Delta(B)}{2(n - 2)}
\end{align*}
\]

(8.3.4)

Above a more complicated example with the program Past is shown. Starting with raw data the program first calculates the distances between the species (in this case from correlation coefficients) and then proceeds as shown above.

A very good and simple introduction to neighbour joining gives http://artedi.ebc.uu.se/course/sommar/njoin/index2.html.
9. Factor analysis

In the last years factor analysis has become very popular among biologists. The reason is surely its ease in use and its ability to generate interpretable hypotheses. Additionally it generates new artificial variables (the factors) that can be used in further analysis for instance in indicator species analysis or multiple regression.

On first sight factor analysis seems to be quite similar to cluster analysis. Both methods group variables into clusters of more or less similar elements. But there is an important difference. **Factor analytical methods use correlation matrices and rely on the general linear model.**

Factor analysis is not a single method, it is a general name for a multitude of different techniques. And this large number of techniques makes factor analysis to a rather dangerous method in the hand of inexperi-
enced users who are not aware of the many pitfalls.

Starting point of any factor analysis is a data matrix of variables and cases. Factor analysis does not distinguish between dependent and independent variables. Our question is how are the variables related? We have two possibilities. Either the variables are intercorrelated or there are other hidden variables, the factors, that influence our observed variables. A regression analysis would deal with the first hypothesis. **Factor analysis assumes a priori that we have hidden factors.** This is graphically shown in Fig. 9.1. A series of variables A to J is influenced by three main factors. These factors cluster our variables. Factor 1 mainly influences A to D, factor 2 variables C to H and factor 3 variables F to J. Note that influences overlap. Aim of the analysis is to define and interpret these factors. Factor analysis groups therefore variables according to their correlations with these factors. It is then our problem to interpret these factors.

Fig. 9.2 shows the essentials of a factor analysis in more detail. We start with a data matrix containing six cases and five initial variables. These variables are for convenience Z-transformed, hence have a mean of 0 and a variance of 1. The aim of the factor analysis is now to find new variables F and its respective factor values \( f_{ij} \) for each case. Similar to multiple regression this is done from a linear combination of the z-values with known weights b.

\[
f_{ij} = \sum_{k=1}^{n} z_{ik} b_{kj}
\]

In general we have the systems of linear equations

\[
\begin{bmatrix}
  f_{11} & \cdots & f_{ij} \\
  \vdots & \ddots & \vdots \\
  f_{ni} & \cdots & f_{nj}
\end{bmatrix}
= \begin{bmatrix}
  z_{11} & \cdots & z_{in} \\
  \vdots & \ddots & \vdots \\
  z_{ki} & \cdots & z_{kn}
\end{bmatrix}
\begin{bmatrix}
  b_{11} & \cdots & b_{ij} \\
  \vdots & \ddots & \vdots \\
  b_{ni} & \cdots & b_{nj}
\end{bmatrix} = \begin{bmatrix}
  Z \\
  \cdots \\
  \cdots \\
  \cdots \\
  Z
\end{bmatrix} \cdot \begin{bmatrix}
  b_{11} & \cdots & b_{ij} \\
  \vdots & \ddots & \vdots \\
  b_{ni} & \cdots & b_{nj}
\end{bmatrix}
\]

The aim is to infer \( b \) and \( F \) is such a way that they predict \( Z \) best.

We can now use the correlation of the predicted \( f \) values with the initial z values. Because the \( f \) values are also Z-transformed we have

\[
r_{ij} = a_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} f_{kj} z_{ki}
\]

This correlation is a measure of distance and describes how close a variable is dependent on a factor. The correlation is called **factor loading.** The main output of a factor analysis is a matrix that gives the factor loading for each factor on each variable. The values \( a_{ij}^2 \) are of course coefficients of determination and give the proportion of total variance of \( i \) that is explained by the new factor \( j \). The sum of all \( a_{ij}^2 \) should ideally equal the total variance of \( i \). Because factor analysis uses Z-transformed input values this variance is one. Hence

\[
0 \leq \sum_{i=1}^{k} a_{ij}^2 \leq 1
\]

This sum is called **communality.** Because we intend to reduce the number of variables the maximum communality is frequently lower than 1, that means the new factors let part of the total variance of a variable unex-
The sum of the factor loading $a_{ij}^2$ is the same as the sum of the part of variance that is explained by a certain factor. From eq. 5.28 we learned that these are the eigenvalues $\lambda$ of the factor (eigenvector) $j$.

$$\lambda_j = \sum_{i=1}^{n} a_{ij}^2$$  \hspace{1cm} (9.3)

If we divide the eigenvalue through the number of variables we get the fraction of total variance each factor explains.

We can show the main task of a factor analysis graphically (Fig. 9.3) We have two sets of data that can be arranged in a (x,y) coordinate system. The aim is now to find new axes, the factors, in such a way that all data points are closest to these axes. If the axes are orthogonal we speak of a canonical or principal component analysis (PCA) and the factors are identical to the eigenvectors of the matrix. This is the most often met type of factor analysis and the easiest to interpret because our axes are independent. In our example we turn the system by an angle of $\alpha$ and get two new axes. Both data a close to these new axes. Hence each set is now explained by one factor instead of the previous two. Mathematically we have to minimize the variance of our data according to these new axes. In a PCA these new axes (eigenvectors) are orthogonal and therefore independent. They explain successively the maximal explainable variance. Further the higher the original variables were correlated the fewer new factors we need to explain total variance. To solve eq. 9.1 we now go back to the regression section.

$$R = \frac{1}{n-1}Z'Z$$

Further we are looking for such vectors that are closest to the initial variables. Hence we need

$$\Sigma = S \cdot R \cdot S$$

$$(\Sigma - \lambda I) \cdot b = 0$$

where $S$ is the diagonal matrix with the standard deviations $s_i$ and $s_j$. $b$ is the matrix of major axis (factors) of the system and $\lambda$ are the associated eigenvalues.

From Fig. 9.2 we see that factors loadings (the correlations between factors and initial variables), factors and variables are connected in an elementary way.

$$r_{FIA} = a_{FIA} = \sum_{k=1}^{j} Z_{Fk}Z_{Ak}$$

The basic assumption of factor analysis is that $Z$ can be expressed by a linear combination of the factor loadings $a$ and the factor values $F$

$$Z_j = a_{1j}F_1 + a_{2j}F_2 + a_{3j}F_3 + \ldots = \sum_{k=1}^{j} a_{kj}F_k$$

In matrix notation
We used the fact that the factors are not correlated. What does this mean? Factors are not correlated when they are orthogonal. Then the increase in one factor is not associated with the increase in the other. This is shown above in a simple two dimensional example. Hence that means \( F \cdot F' = cI \) with \( c \) being a constant. At the end we get

\[
\text{Eq. 9.4 gives the relation between the original correlation matrix and the matrix of factor loadings. It is called the fundamental theorem of factor analysis.}
\]

The problem is now to find such a matrix \( a \) with such a number of dimensions (factors) that is closest to \( R \).

Recall that eigenvectors are the major axis of ellipsoids. Hence these should be the new axes (factors) we are looking for. From the definition of eigenvalues and eigenvectors we also get a method to compute the total variance explained by each factor. We have to solve

\[
[R - \lambda_i I] \cdot u_i = 0
\]

As a result a factor analysis weights how much each factor contributes to the variance of each variable. These weights are the factor loadings and are nothing more than correlation coefficients of the factors with the variables. The resulting factors are independent of each other and explain successively the total variance.

A factor analysis proceeds therefore in four main steps. The first step is the computation of a correlation matrix. The next step is to arrange the variables into an \( n \)-dimensional space in such a way that the distances of all variables to the axes of the space become minimal. The problem is to find a solution where \( i \), the number of factors (dimensions), is minimal. The third step is to rotate the whole \( i \)-dimensional space around its axes in such a way that the variables range even more close around them. This is equivalent to the calculation of the eigenvectors. The loadings of each variable are than nothing more than inverse measures of distance. The closer a variable lies to an axis, the higher loads the factor to this axis.

At each step of the analysis we have to decide about principal ways of computation. In a first step we might decide whether we use raw data or Z-standardized data with mean 0 and variance 1. In many cases the latter way is preferable.

The next problem is to define how much of total variance the factors should explain. In fact eq 9.4 should be written different
The matrix of factor loadings does not explain total variance. There is rest variance contained in the matrix E. The problem to assess E is called the **communality problem**. There are three basic techniques to estimate the communality a priori. If we have no explicit hypothesis about our factors we simply assume that they together explain 100% of total variance. In this case we perform a **principal component analysis**. This is the default option of many statistic packages. We might also estimate the total variance explanation in the course of an iterative computation. In this case we perform a **major axis factor analysis**. Finally, we may take the highest correlation coefficient between two of the variables as an estimator of the maximum variance explanation of the factors. This is of course only a rough estimate. Look at the Table above. I used ten variables to conduct a default principle component analysis with **Statistica**. The program extracted five factors. If we compute the sums of the squared factor loadings for each variable we get the communalities. These values range between 0.64 and 0.84. Even a principal component analysis cannot explain the total variance. We would need more factors. But then most factors would load high only with single variables. Hence, we have to interpret that there are no hidden variables or that we don’t need to assume them. A simple regression analysis would do the same job. We also see that the first factor explains 18% of total variance, the fifth factor still 13%. In total the three factors explain 77% of variance. 23% remain unexplained.

The next step in our analysis is the rotation of axes so that the distances of the variables from the axes minimize. Again there are several techniques. Important is to differentiate between **orthogonal and oblique rotation techniques**. After orthogonal rotation factors remain independent. Otherwise they become correlated. The latter case is difficult to interpret. We should therefore prefer the **orthogonal Varimax rotation** method. Additionally, Z-transformed variables should be used. The result is a Table as above that shows a default principal component analysis with Varimax rotation.

How many factors should we accept as being significant? In our case the maximum number of factors would be ten. However, we want to find hidden variables. We need therefore a solution where at least two, better three or more variables load high with a factor. There are several criteria to find the optimal number of factors. Most often used is the **Kaiser criterion**. It says that we should only accept factors that have eigenvalues...
above 1. Alternatively we may apply a **Scree test**, similar to the elbow test of cluster analysis. This is shown in the Figure 9.4. According to Cattell factors should be considered as being significant if they are before the knee of the eigenvalue—factor rank order plot. These would be the first three factors. **Additionally, empirical data should show such a knee.** A factor analysis done with random variables should lack the knee. In our Table above a distinct knee is lacking. Indeed, the variables used were simple random numbers.

Which variables define a factor. Those that load high with a given factor. The question is therefore how to decide what is high and what not. There is no simple solution to this problem. But there are some valuable rules.

- A factor might be interpreted if more than two variables have loadings higher than 0.7.
- A factor might be interpreted if more than four variables have loadings higher than 0.6.
- A factor might be interpreted if more than 10 variables have loadings higher than 0.4.

In the Table on the previous page factor 1 loads highly positive with Var 8 and highly negative with Var2 and Var9. We can try to interpret this factor from these three variables. Factor two loads highly negative with Var3 and positively with Var7. Factor 3 might be interpreted from Var5 and Var6. However, the remaining two factors are interpretable only on the basis of the high loading single variables.

Prior to PCA we should infer whether it makes sense to do so. Is our data matrix appropriate? This can be assessed by the **Bartlett test** that tests whether the intercorrelations of the variables are sufficient to give room for common factors behind them. Particularly, the test uses the null hypothesis that the distance matrix is the identity matrix that means that all of diagonal elements are zero (no correlation or maximum distance). We calculate

\[ v = -(n - 1 - \frac{2k + 5}{6}) \ln(R) \]

(3.6)

where k is the number of variables and n the number of observations. v is \( \chi^2 \) distributed with \( k(k-1)/2 \) degrees of freedom.

The Excel example beside shows a correlation matrix of \( k = 6 \) variables. The identity matrix gives the \( \chi^2 \) probability of 1 for \( H_0 \). The second example shows how to use the Bartlett test to assess necessary sample sizes. At even moderate mean pairwise correlations you would need at least 30 cases to detect a significant structure (at \( p(H_0) < 0.05 \)). Most textbooks advise not to use factor analysis with fewer than 50 cases irrespective of the number of variables.
9.1 Advices for using factor analysis:

- Your data have to be of the metric type.
- Make sure that the assumptions of the GLM are not strongly violated. In doubtful cases use log-transformed variables.
- Try independent analyses with raw and z-transformed data sets. The results should be qualitatively the same.
- Use a principal component analysis with Varimax rotation. This makes interpretation easier.
- Use the Scree criterion for identifying the maximum number of significant factors.
- In the case of missing values whole cases should be excluded.
- The number of cases should be at least twice as large as the number of variables. Never use matrices where the number of cases is less than the number of variables.
- If you have enough data sets divide them randomly into two groups and perform for each group an analysis. Interpret your factors only when both results are identical.

Factors are new variables (Fig. 9.2) and come from \( Z = F \cdot b \). They have factor values (independent of loadings) for each case. These factors can now be used in further regression analysis. This method is especially useful when dealing with a manifold of variables which are highly correlated. In this case you should use a factor analysis to generate new artificial variables (but with meaningful interpretations) and undertake then a multiple regression. The PAST example above shows the original data, the respective eigenvalues and a diagram of factor loadings. Factors values for each case are scattered and printed.

Factor analysis can also be used to cluster cases. However, in factor analysis the number of variables to be clustered should always be larger than the number of cases. Therefore, for factorizing cases you would need many variables. Nevertheless, sometimes such a procedure is a good alternative to cluster analysis.
10. Canonical regression analysis

Multiple regression establishes linear relationships between a set of independent (predictor) variables and one dependent variable. Canonical regression analysis is a generalisation of multiple regression and deals with several dependent variables. We have therefore two sets of variables and get not a single but several regression functions. The number of functions equals the number of variables in the smaller set. The method is called canonical because only linear dependencies are assumed.

The principles of a canonical regression analysis are simple. First, you run two principal component analyses, for the set of dependent and for the independent variables. Next, you rotate the factor axes in such a way that the first factor of the independent variables set correlates maximally with the first factor of the dependent set. Hence, canonical regression analysis is a correlation analysis that uses principle component factors.

<table>
<thead>
<tr>
<th>Canonical regression analysis results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical R</td>
</tr>
<tr>
<td>CHI²(32)</td>
</tr>
<tr>
<td>p</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>Root 1</th>
<th>Root 2</th>
<th>Root 3</th>
<th>Root 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var 1</td>
<td>-0.58819</td>
<td>0.341623</td>
<td>0.710683</td>
<td>-0.49699</td>
</tr>
<tr>
<td>Var 2</td>
<td>-0.38425</td>
<td>-0.08823</td>
<td>0.40191</td>
<td>-0.73306</td>
</tr>
<tr>
<td>Var 3</td>
<td>-0.41799</td>
<td>0.653184</td>
<td>0.078462</td>
<td>0.385908</td>
</tr>
<tr>
<td>Var 4</td>
<td>-0.38081</td>
<td>-0.92348</td>
<td>-0.70002</td>
<td>0.459424</td>
</tr>
<tr>
<td>Var 5</td>
<td>0.012453</td>
<td>-0.08183</td>
<td>-1.02332</td>
<td>0.115108</td>
</tr>
<tr>
<td>Var 6</td>
<td>-0.41992</td>
<td>0.634172</td>
<td>0.044209</td>
<td>-0.79315</td>
</tr>
<tr>
<td>Var 7</td>
<td>-0.73958</td>
<td>0.320323</td>
<td>-0.41184</td>
<td>-0.07587</td>
</tr>
<tr>
<td>Var 8</td>
<td>0.022575</td>
<td>0.011094</td>
<td>-0.11499</td>
<td>-0.35833</td>
</tr>
<tr>
<td>Predicants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAR 1</td>
<td>0.057528</td>
<td>-1.15332</td>
<td>0.081897</td>
<td>0.200481</td>
</tr>
<tr>
<td>VAR 2</td>
<td>-0.38791</td>
<td>-0.78640</td>
<td>-0.70295</td>
<td>-0.47974</td>
</tr>
<tr>
<td>VAR 3</td>
<td>-0.48525</td>
<td>-0.15070</td>
<td>0.780028</td>
<td>-0.41578</td>
</tr>
<tr>
<td>VAR 4</td>
<td>0.928718</td>
<td>0.134346</td>
<td>0.21516</td>
<td>-0.4467</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eigenvalues</th>
<th>0.683835</th>
<th>0.620702</th>
<th>0.33699</th>
<th>0.033358</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicatars</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance explained</td>
<td>0.21698</td>
<td>0.147743</td>
<td>0.233063</td>
<td>0.402214</td>
</tr>
<tr>
<td>Redundancy</td>
<td>0.148378</td>
<td>0.091704</td>
<td>0.07854</td>
<td>0.013417</td>
</tr>
<tr>
<td>Predicants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance explained</td>
<td>0.109118</td>
<td>0.139378</td>
<td>0.096843</td>
<td>0.134036</td>
</tr>
<tr>
<td>Redundancy</td>
<td>0.074618</td>
<td>0.086512</td>
<td>0.032635</td>
<td>0.004471</td>
</tr>
</tbody>
</table>

The interpretation of a canonical regression analysis is difficult. The Table above shows a combined output of the Statistica software package. I used 12 variables and 30 cases. The first four variables constituted the set of dependent and the last eight variables the set of independent variables.

Statistica gives for both sets four roots, according to the number of variables in the smaller set. As in multiple regression the regression coefficients (weights) should take values between –1 and 1, but as for beta-weights values larger than 1 or smaller than –1 are possible. Such cases point always to violations of the general linear model, especially to high levels of multicollinearity. As in ordinary regression analysis high levels
Statistical advices

of multicollinearity (pairwise $r > 0.9$) makes it often very difficult to interpret the results.

It seems that in the first root the variables seven and one contribute most to the first root in the second set that loads high with VAR 4. In the second root variables 3 and 6 seem to be positively and variable 4 negatively correlated to root 2 of the dependent variables set. VAR 1 and VAR 2 contribute negatively to this root. It depends on us how to interpret this pattern (having real variables).

The eigenvalues (canonical correlations) of the first two roots are highest and we have to interpret that they contribute most to total variance explanation. However, this does not mean that they explain most of total variance. In canonical regression analysis the measure how much of total variance a root explains is called redundancy. Redundancy is computed over the sum of all squared factor loadings using $z$-normalized variables. The redundancies of the predicant roots of our example are low (0.07 and 0.08). It means that only 7 to 8% of the variance in the set of dependent variables is redundant (is explained by the predictor set). Nevertheless, our canonical correlation coefficient is 0.82 and is highly significant ($p = 0.004$). However, again I used only simple linear random numbers to generate the data matrix. Canonical regression gives very quickly highly significant results. The reason is that we use even more pairwise correlations than in other multivariate methods. The canonical correlation coefficient but is always larger than the largest pairwise correlation coefficient. Hence, with a manifold of variables in both sets it is very easy to get to highly significant canonical correlations from nothing. It is therefore important to use Bonferroni corrected experiment wise error rates. Accept only canonical correlations at $\alpha < 0.05 / n$ where $n$ is the total number of variables in the matrix.

Canonical regression analysis depends more than other multivariate techniques on our interpretation. This is surely the reason why it is relatively seldom used in the natural sciences. It is a technique designed for the social sciences where model interpretation is more important.

### 10.1 Advices for using canonical regression analysis

- Canonical regression analysis is—similar to ordinary regression analysis—based on sets of linear regressions. The basic prerequisites of the method are therefore similar to those of multiple regression. It can be seen as a generalization of all regression techniques.
- The predicant variables have to be metrically scaled.
- The sample should contain at least twice the total number of variables.
- Canonical regression analysis makes only sense if you deal in both data sets with at least three variables. If you have less variables perform ordinary regression.
- High levels of multicollinearity should be avoided. Leave out redundant variables.
- It is difficult to assess whether results of a canonical regression analysis are stable. If your data set is large you might again subdivide it at random and use one part for verification. Another technique is bootstrapping.
- Use always Bonferroni corrected error levels.
11. Discriminant analysis

Discriminant analysis is of major importance in the social sciences. In biology it is slightly neglected although it was introduced by the genetic Sir Ronald Fisher (in 1936) and might have broad application. In discriminant analysis we compare simultaneously groups of a dependent variable with sets of independent variables. Our aim is to cluster the independent variables in such a way that we can answer two basic questions: Are there significant differences between the groups of the dependent variable? How can these differences be explained. Additionally, discriminant analysis allows us to place an object of unknown group membership into a certain group.

Ordinary discriminant analysis deals with predefined groups of a dependent variable and groups them according to a set of predictor variables. This is the major difference between this technique and cluster or factor analysis where groups are defined during analysis. These techniques are therefore complementary and can be used in parallel. Discriminant analysis is particularly important in taxonomy for classifying material of unknown species identity into taxonomic groups.

As other multivariate techniques discriminant analysis is based on the general linear model. It assumes therefore that predictor and dependent variables are linearly related. The main approach is similar to regression analysis. A discriminant function, or root, has the general form

\[ Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = a_0 + \sum_{i=1}^{n} a_i X_i \]  

In matrix notation

\[ Y = X \cdot a \]

Hence for each group of the grouping (dependent) variable Y the discriminant analysis reduces the n predictor variables via a regression analysis to one variable Y. The next step is then the use of a predefined metric to define the distances and the variances between the groups of the dependent variable. In effect discriminate analysis is an analysis of variance based on values defined by a principal component analysis. It is a combination and extension of both methods.

Consider the example of raw data given in the Table below. We have a grouping variable in which three groups have been distinguished. Our aim is to find out whether these groups can be defined in terms of six predictor variables. The first step is the computation of the discriminant function. If your dependent variable is already ordinary
scaled this function can be estimated directly. Otherwise it is advisable to undertake first a cluster analysis to define possible groups. What we need is a discriminant criterion. This criterion tries to define the discriminant function to mirror maximum distances between the group means defined by the predictor variables. Remember that we can divide total variance of a data matrix into parts of within group variance and between group variance. This is the same as to discriminate between explained (between group) and unexplained (within group) variance. As in an analysis of variance our discriminant criterion is the quotient of between and within group variance (SS = sums of squares).

\[
\Gamma = \frac{SS_{\text{between}}}{SS_{\text{within}}} = \frac{U' \cdot B \cdot U}{U' \cdot W \cdot U}
\]  

(Number 11.2)

Where B and W are the dispersion matrices containing the respective sums of squares and V is the respective eigenvector. The optimal solution for the discriminant function is found when \( \Gamma \) becomes a maximum \( \gamma \). To solve this problem we use matrix notation. We have \( Y=X\cdot a \).

We have to maximize \( SS_{\text{between}} = (Y-X\cdot a)^2 \). Hence

\[
\frac{dSS}{da} = 2X' (Y-X\cdot a) = 2X'Y - 2X'X\cdot a = 0 \rightarrow a = (X'X)^{-1} \cdot X'Y
\]

X'X gives the covariance matrix. It’s eigenvalue is therefore a measure of the variance explained by the function. To solve eq. 11.2 we rearrange

\[
(W^{-1} \cdot B - \lambda I) \cdot U = 0
\]

(Number 11.3)

There is not a single discriminant function. Each function explains only a part of the whole variance. This part is measured by the discriminant factor (the eigenvalue) defined by the loadings of each predictor variable (similar to factor analysis). The part of variance explained by a function \( u_i \) is simply defined by the quotient \( \gamma_u \) through the sum of \( \gamma \)-values. If the dependent variable consists of \( n \) groups \( n-1 \) discriminant functions (called roots) exist. To define which of the discriminant functions are statistically significant a coefficient is used that is called Wilk’s lambda. This is defined by

\[
L_i = \prod_{k=i+1}^{n} \frac{1}{1 + \gamma_k}
\]

(Number 11.4)

By transforming Wilk’s lambda to \( \chi^2 \) we can test the significance of the remaining \( g-k-1 \) eigenvalues after accepting \( k \) prior ones.

\[
\chi^2 = -(n - 1 - \frac{m + g}{2}) \ln L
\]

(Number 11.5)

where \( n \) is the number of cases, \( m \) the number of predictors and \( g \) the number of groups. The test has \( (m-k)(g-k-1) \) degrees of freedom. The result of our example shows the Table beside.

<table>
<thead>
<tr>
<th>Chi² test of roots</th>
<th>Eigenvalue</th>
<th>Canonical R</th>
<th>Wilk’s lambda</th>
<th>CHI²</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.568714</td>
<td>0.602109</td>
<td>0.637465</td>
<td>5.177945</td>
<td>5</td>
<td>0.394573</td>
</tr>
<tr>
<td>2</td>
<td>9.444256</td>
<td>0.950922</td>
<td>0.081035</td>
<td>32.15754</td>
<td>12</td>
<td>0.001314</td>
</tr>
</tbody>
</table>
verse measure with smaller values indicating higher significance. We see that only the first root is statistically significant. It explains 94% of total variance.

The next step of a discriminant analysis is the formulation of a classification function. The classification function looks similar to the discriminant function and is in the simplest case directly derived from the discriminant function.

\[ Y_i = a_0 + a_{1,j}X_{1j} + a_{2,j}X_{2j} + \ldots + a_{n,j}X_{nj} = a_0 + \sum_{j=1}^{n} a_{j}X_{ij} \]

However many other algorithms for deriving classification functions have been proposed. Best known are the generalized distance functions (based on the Mahalanobis distance) and non-parametric classification.

Statistical packages give the \( \lambda \)-values of this function as shown beside. On the basis of this function we get a value for each element of the grouping variable. But we need something that tells us which group this function predicts. This provides us the probability concept that is based on Bayes theorem of dependent probability. The probability that an element \( i \) belongs to group \( m \) is defined by

\[ p(i, m) = \frac{e^{Y_{i,m}}}{\sum_{k=1}^{n} e^{Y_{i,k}}} \]

The equation points therefore with probability \( p \) to a certain group. The sum of all probabilities adds to one. For instance a new element \( X \) has for each variable the following values: 0.15, 0.57, 0.51, 0.89, 0.61, 0.83.

We get probabilities as given in the next Table. \( X \) belongs with probability \( p = 0.76 \) to A, with probability 0.0002 to B, and with probability 0.24 to C. Probably, the new element belongs to group A. The possibility that a certain object does not belong to any group is excluded. This has to be considered when classifying new objects with a classification function of a discriminant analysis.
11.1 Advices for using discriminant analysis:

- Discriminant analysis is—similar to regression analysis—based on sets of linear regressions. The basic prerequisites of the method are therefore similar to those of multiple regression.
- The predictor variables have to be metrically scaled.
- In cases of strong deviations from linearity log-transformed variables should be used.
- Make sure that no element of the sample belongs to more than one group. Hence, your model must assure that all variables can be grouped unequivocally.
- The sample should contain at least twice the number of variable used for discrimination.
- The number of grouping variables should always be larger than the number of groups.
- Discriminant ability depends on the metrics chosen and on the grouping algorithm. If the independent variables are not or only slightly correlated a step wise procedure should be used. In this case the statistic package filters step by step those variables that best discriminate between the groups. At high levels of multicollinearity a simultaneous grouping should be preferred.
- Check your results graphically. In this case use only the first two discriminant functions.
- Use only the significant discriminate functions for interpretation.
12. Multidimensional scaling

Multidimensional scaling (MDS) is not a single statistical technique. It stands for a family of techniques designed for the analysis of hidden structures in the data set. MDS can be seen as an alternative to factor analysis. MDS techniques try to find meaningful underlying dimensions along which our variables can be placed in such a way that their distances to these dimensions become minimal. Our task is then to interpret these dimensions in terms of the variables nearest to them. This is similar to factor analysis but while factor analysis deals exclusively with correlation matrices MDS may analyze any kind of similarity or dissimilarity matrix.

MDS can be viewed as a way of rearranging objects in an efficient manner to arrive in a configuration that best approximates observed distances in a predefined n-dimensional space (Fig. 12.1). To do this we need an algorithm that tells us when a certain configuration in space gives minimal distances to the number of predefined axes. This algorithm provides a measure of the goodness of fit and is commonly termed stress. Stress is an inverse measure. The lower its value is, the better fits our configuration of axes to the data matrix. The most common measure of stress is the PHI-value defined by

\[ PHI = \sqrt{\frac{\sum \sum (d_{ij} - \delta_{ij})^2}{\sum \sum (d_{ij})^2}} \]

where \( d \) stands for the distances after rearrangement in the predefined n-dimensional space and \( \delta \) stands for the observed distances. Hence MDS uses least squares to find the best solution. The equation deals with distances or proximities. We can use any appropriate measure of proximity. Most statistical packages use the general Minkowski concept with \( z = 2 \), hence they apply Euclidean distances.

The first problem of MDS is the initial definition of the number of dimensions, hence the number of hidden variables we have to deal with and to interpret. The higher the number of dimensions is, the better will (in general) be the fit of our MDS model. If we define as many dimensions as we have variables in the data matrix, the fit would be perfect. But then we get no more information. The aim of MDS is of course to reduce the number of variables that have to be interpreted (the complexity). The problem of choosing the right number of dimensions is therefore similar to factor analysis. A common technique to decide how many dimensions are appropriate is the Scree test. We plot the final stress values against different numbers of dimensions. The right number of dimensions should be at
the point where such a plot levels off at the right side. Fig12.2 indicates a three dimensional solution. Another important criterion is the clarity of the solution. If the data points in the final plot form an unstructured cloud the number of dimensions might be inappropriate to produce an interpretable solution. In this case we should try other spaces.

Most statistical packages do not use raw data for analysis but need distance matrices. An example gives the Table above. Statistica needs a specific distance matrix containing first all pairwise distances (in this case pairwise correlations). Additionally you have to give means and standard deviations for each variable, the number of cases (they have to be identical for each variable) and the type of distance matrix. In our case the one stands for a correlation matrix. After rearrangement Statistica gives the final configuration with distances of all variables to the predefined axes. The next step is a plot as in Fig. 12.3. These scatterplots visualize the distances of all variables to the axes. Our task is then to interpret them. Fig. A shows a scatterplot of the first two dimensions from a three dimensional solution. B shows the same in the case of a four dimensional solution (data of the correlation matrix above). Note that dimension numbers are arbitrarily. Dimension 1 in A is not the same as dimension 1 in B. It seems that B is better interpretable than A while separating variables B, C, D, and
H more clearly. In A variables I, E, A, and G seem to cluster together, a view that is not supported by the four dimensional solution. But be careful. Again, the original data matrix contained simple random numbers. Scatterplots but tempt us to see patterns even when there is nothing. This pattern seeking mentality makes MDS a dangerous technique when applied to data without thorough prior model building.

At last we have to interpret our dimensions. This is a problem similar to factor analysis. Interpretation has to be done according to the variables lying very near to or very far from the focal dimension.

Statistica performs a classical metric MDS. In this case data are quantitative and proximities (correlation coefficients or other measures of distance) are treated directly as distances. In many cases it is appropriate to transform data prior to analysis. We might use log-transformed or standardized data, or we might performed a weighted MDS where variables are given weights prior to analysis.

Often, our data are not of the metric type. They might represent classes like ‘yes’ or ‘no’ or group memberships like ‘A’, ‘B’, ‘C’ etc. In this case we might perform non-metric MDS, an option that provide some statistic packages like PC Ord or CANOCO.

Non-metric MDS is based on ranked distances and tends to linearize relations. It iteratively searches for rankings and placements of n entities in k dimensions. Optimal solutions are found by comparing the stress values of the data matrix with those of randomized matrices. Randomization means that the data from the main matrix are reshuffled within columns. PC Ord reads raw data matrices formatted as shown above. You have to give the numbers of plots and variables and information whether the variables are of a metric (Q) or a categorical type (C). The example contains three metric and two categorical variables that characterize 10 species. I used the Soerensen measure of distance and an initial number of 4 dimensions. The first step of the analysis is a test which number of dimensions is appropriate. In non-metric MDS this is done via a Monte Carlo test using randomized data matrices. In our example a one dimensional solution is proposed because it is the lowest dimensional solution with a minimum number of randomized data

<table>
<thead>
<tr>
<th>Axes</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
<th>Stress in real data 15 runs</th>
<th>Stress in randomized data Monte Carlo test 30 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.193</td>
<td>25.42</td>
<td>50.895</td>
<td>11.557</td>
<td>25.42</td>
</tr>
<tr>
<td>2</td>
<td>3.082</td>
<td>6.714</td>
<td>26.443</td>
<td>8.135</td>
<td>2.031</td>
</tr>
<tr>
<td>3</td>
<td>0.377</td>
<td>3.252</td>
<td>20.956</td>
<td>2.972</td>
<td>0.28</td>
</tr>
<tr>
<td>4</td>
<td>0.001</td>
<td>1.17</td>
<td>14.569</td>
<td>1.747</td>
<td>0.018</td>
</tr>
</tbody>
</table>

p = proportion of randomized runs with stress < or = observed stress i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations)
12.1 Advices for using multidimensional scaling:

- Multidimensional scaling provides a visual representation of complex relationships. It doesn’t give any significance values for observed patterns. Hence be careful when interpreting the patterns.
- Measurement uncertainties might have a high influence on MDS results. Use only data with low standard errors.
- Axe numbers are arbitrarily. Hence the first axis is not more important than others. This is a difference to factor analysis.
- Metric MDS relies on data sets with normal distributed errors. If you aren’t sure whether this prerequisite is met transform the data (by ranking or by logarithmization) or use non-metric MDS.
- The larger the number of dimensions is, the more difficult becomes the interpretation. If two or three dimensional solutions still have high stress values abandon MDS and try factor or cluster analysis.
- Use the Scree test to establish the right number of dimensions. A rule of thumb is that stress values above 0.15 (or above 15 when transformed to a range between 0 and 100) are unacceptable, stress values below 0.1 (below 10) are very good.
- For interpreting an MDS plot concentrate on the large scale structures. The stress function accentuates (while using squared distances) discrepancies in larger distances and the algorithm tries hard to get them right. Small scale structures are difficult to interpret and might simply arise due to distortions of the distances caused by non-zero stress.
- For analysis structures inside perceived clusters you may extract the variables forming the clusters and rerun the MDS (so-called local MDS).
- Compare the MDS results (if possible) with those from a factor analysis. This helps interpreting clusters and dimensions.
13. Bootstrapping and jackknifing

Jackknife and bootstrap are general techniques to answer the question how precise estimates are. Both techniques find application when it is impossible to give standard deviations and standard errors from distribution parameters of the whole population. The basic idea behind both techniques is to use the sample as a surrogate of the population and to estimate standard errors from series of subsamples from this sample.

Assume we have five abundance data from a population of birds. This is our single sample. From these five data points we can estimate the mean abundance of the birds. But we have no estimate how precise we estimated mean abundances. We lack an estimate of the standard error. We use the jackknife procedure and take subsamples from these five data points. This is shown in the table above. The original sample is given in bold. It predicts a mean abundance of 11.6 individuals. Now we take five subsamples each time leaving out one data point. We get new mean values.

And we need so-called pseudovalues $p_i$. These pseudovalues are computed from

$$ p_i = \bar{X} - (n-1)(\bar{X} - \bar{X}_i) $$

(13.1)

where $n$ is the number of subsamples. The jackknife estimate of the standard error is then the standard error of the pseudovalues

$$ SE = \sqrt{\frac{\sum (p_i - \bar{p})^2}{n(n-1)}} $$

(13.2)

In our case SE is 4.15. Jackknifing is an especially useful technique for estimating error terms of index values like diversity or similarity metrics.

**Jackknife estimates must not be used to compute confidence intervals.** Therefore you can’t test hypothesis. You can’t use t– or U-tests for comparing means. The great advantage of the jackknife is its ease in use.

More complicated but more precise in estimating distributions is the bootstrap technique that was developed from the jackknife. The general principle is shown in Fig. 13.1. To derive distribution parameters a manifold of subsamples is taken from the single samples. These subsamples serve to approximate the unknown distribution of the population.

Although the principle of the bootstrap technique looks simple the details are quite complicated. First of all the sample has to reflect the total population. Important for the outcome is the way the subsamples are taken. The number of elements can vary at random. In this case many small samples will occur that might give rise to higher standard errors. You can also take subsamples of
The next table shows a simple example. We have ten abundance data from a species and intend to estimate the degree of aggregation. For this we use the Lloyd index of mean crowding given by

\[ L = \frac{s^2}{\bar{x}^2} - \frac{1}{\bar{x}^2} + 1 \]  

We can estimate L from the ten data points. However, this is a simple number. We do not know how precise our estimate is. We need a standard error and confidence limits. We take ten subsamples of different size and compute L for each subsample. The mean of these estimates gives us the bootstrap estimate of L and the standard deviation is our desired estimate of the standard deviation of the sample estimate. The standard error is given by SD / \sqrt{10} = 0.019.

Greater problems arise if we want to compute confidence limits. In the simplest case we might assume a normal distribution of bootstrap values around the true sample mean. In this case the standard error would be calculated as usual.
In many cases however, bootstrap distributions are not normal but skewed. In these cases we might use **studentized bootstrap values**. We use a large number of subsamples (for instance 999) and compute for each

\[ b_i = \frac{\bar{x}_i - \bar{x}}{s_i} \]

where \( n \) is the size of the subsample \( i \) and \( s_i \) its standard deviation. Now we plot the frequency distribution of these \( b_i \) values as shown in Fig. 13.2. If we rank all \( b_i \) values from smallest to largest we get the 95% confidence limits from the 2.5 and 97.5 percentile values. We need the 25\textsuperscript{th} and the 975\textsuperscript{th} \( b_i \) values and use these values instead of \( c_\alpha \). In the example of the Figure above these values are 1.8 and 2.4. Hence

\[ CL = (\bar{x} + b_{2.5 \text{ percentile}} \frac{s}{\sqrt{n}} ; \bar{x} - b_{2.5 \text{ percentile}} \frac{s}{\sqrt{n}}) = (\bar{x} + 2.4 \frac{s}{\sqrt{n}} ; \bar{x} - 1.8 \frac{s}{\sqrt{n}}) \]

The values differ from the 1.96 value of the 95% confidence limit of a normal distribution.

A key assumption behind the bootstrap and the jackknife is that the original sample values were independent. If this is not the case (for instance if the data are correlated) things become more complicated. For correlated data a **shifting block sample procedure** has been proposed where not random samples but many data blocks are used to retain at least part of the correlation.

For estimating standard errors via the bootstrap at least 1000 subsamples should be taken. For having not too much replicates (samples that contain the same elements) the original sample size should not be smaller than 30. For small sample sizes below 10 bootstrap estimates become more and more unreliable. In these cases we should enlarge our confidence limits and reduce type 1 error probabilities to 1\% or even 0.1\%.
13.1 Reshuffling methods

Closely related to bootstrapping are methods that infer statistical significance from a random reshuffling of data. Consider the coefficient of correlation. Statistical significance of \( r > 0 \) (H₁) is tested against the null hypothesis \( H_0 \) of \( r = 0 \). Most statistics programs do this using Fisher’s Z-transformation

\[
Z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)
\]

that is approximately normally distributed and has a variance of \( \sigma^2 = \frac{1}{n-3} \). For instance a coefficient of correlation of \( r = 0.60 \) obtained from 10 data points has \( Z = 0.69 \). \( \sigma = 0.38 \). Hence \(-2\sigma < Z < 2\sigma = 0.76\) and we conclude that at the 5% error level \( r \) is not significantly different from 0.

We can test in a different way and use a reshuffling method. The table above shows the necessary calculations. \( X \) and \( Y \) are correlated by \( r = 0.6 \). Now we reshuffle \( Y \) 10 times (in reality at least 100 reshufflings would be necessary for a good estimate of the standard deviation \( s_r \) of \( r \). For simple inference of statistical significance however in most cases even 50 reshufflings are sufficient. If we estimate \( s_r \) sufficiently precisely we can use the Z-transformation (not Fisher’s Z) and compare \( Z \) with the standard normal distribution. For our 10 reshuffles this comparison points to a significant difference at \( p < 0.05 \).

Another way is to use a t-test

\[
t = \frac{0.6}{0.26} = 2.31
\]

Note that the standard error of a distribution is identical to the standard deviation of the sample. You have to run the t-test with 1 degree of freedom. We get \( p = 0.13 \) n.s.. We see that each of the three methods gives a different level of significance. Each test has a different power and uses a different approximation to the standard normal.

Another example. The following Excel table contains two variables \( X \) and \( Y \). We wish to compare the means. A t-test is problematic because both distributions are highly skewed. The alternative, an U-test might be influenced by the high number of tied ranks. The t-test points to a significant difference at \( p < 0.05 \). Again we use a reshuffling of data. We compare the mean significance level of (in theory 1000) randomizations with the observed one. In this case we can do this without programming using a very good Excel add in, PopTools that
is freely available at www.cse.csiro.au/poptools. This add in contains many advanced level matrix applications and statistical tests. One module contains a **Monte Carlo** simulation. Monte Carlo simulations provide us wish resampling procedures with desired confidence limits. In our example 500 resamplings with replacement of the data (using the PopTools module *shuffle*) gives a mean significance level of 0.67 with a lower confidence limit (at p < 0.01) of 0.03. We infer that there are no differences in the mean of both variables because the observed significance level is still within the confidence limit range. But look at the t-test. From a mean of 0.67 and a standard deviation of \( s = \sqrt{0.06} = 0.24 \) we would infer a significant difference of the observed and the reshuffled significance levels \( Z = (0.67 - 0.05) / 0.24 = 2.58; \ p = 0.01 \). The reason is that the skew in our data biases the type I error level of the parametric test that assumes a normal random distribution of errors around the mean. Indeed, the upper 95% confidence limit of \( r = 0.67 + 2 \times 0.24 = 1.15 \) is already impossible to achieve. Such skews are quite frequent in biological data and if we are unsure about the data structure a Monte Carlo approach is desirable.

A third example shows how to infer regularities. Assume 20 species of a genus occur within a given habitat. Classical ecological theory tells that species to co-occur must differ to a certain degree to minimize interspecific competition. We want to test this assumption. We have data on species body weights. Our question is whether species body weights are regularly spaced within the observed range. The aim is to compare the observed variance with the variance expected if body sizes were randomly distributed along the body weight axis. Hence we rank our data according to body size and calculate the variance of the distances in body weight. This variance would be minimal at an equal spacing of body weight. The more species body sizes are aggregated (clumped) the higher the variance should be.

The next table shows how to perform the test. Our raw data are first ranked. The next step is to compute the quotients \( q = W_n / W_{n-1} \) for all consecutive species pairs. Hence, we use an equal spacing at a log scale (\( \ln [q] = \ln[W_n] - \ln[W_{n-1}] \)). This is our test variable. If species weight were regularly spaced the variance of \( s_q \) would be low. Now we test this value against an appropriate null model. We fix the upper and lower values of body weight and assign \( S - 2 \) body weights at random within this range. From 1000 or more such randomizations we infer each time \( s_q \) and get also the standard error of \( s_q \). From this we obtain upper and lower 95% confidence limits. If the observed \( s_q \) value is within this range we conclude a random distribution of body size. In the previous table I show only 4 such randomizations and use the Z-transformation. This procedure is valid.
### Statistical advices

<table>
<thead>
<tr>
<th>Species</th>
<th>Body weights</th>
<th>Ranked body weights</th>
<th>W/W₁, Body weights</th>
<th>Randomized body weights</th>
<th>Ranked randomizations</th>
<th>W/W₁, Randomized</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>2.00</td>
<td>0.90</td>
<td>0.92</td>
<td>1.05</td>
<td>0.40</td>
<td>0.43</td>
<td>0.33</td>
</tr>
<tr>
<td>3.00</td>
<td>1.02</td>
<td>1.06</td>
<td>1.73</td>
<td>0.45</td>
<td>0.47</td>
<td>0.39</td>
</tr>
<tr>
<td>4.00</td>
<td>1.07</td>
<td>1.12</td>
<td>2.13</td>
<td>0.50</td>
<td>0.67</td>
<td>0.41</td>
</tr>
<tr>
<td>5.00</td>
<td>1.07</td>
<td>1.14</td>
<td>3.25</td>
<td>1.10</td>
<td>1.36</td>
<td>1.16</td>
</tr>
<tr>
<td>6.00</td>
<td>1.06</td>
<td>1.19</td>
<td>2.86</td>
<td>1.22</td>
<td>1.58</td>
<td>1.25</td>
</tr>
<tr>
<td>7.00</td>
<td>1.04</td>
<td>1.21</td>
<td>3.15</td>
<td>1.10</td>
<td>1.45</td>
<td>1.17</td>
</tr>
<tr>
<td>8.00</td>
<td>0.91</td>
<td>1.98</td>
<td>3.87</td>
<td>1.23</td>
<td>1.65</td>
<td>1.27</td>
</tr>
<tr>
<td>9.00</td>
<td>0.78</td>
<td>3.16</td>
<td>4.58</td>
<td>1.01</td>
<td>1.40</td>
<td>1.19</td>
</tr>
<tr>
<td>10.00</td>
<td>0.83</td>
<td>3.08</td>
<td>4.38</td>
<td>1.10</td>
<td>1.52</td>
<td>1.20</td>
</tr>
<tr>
<td>11.00</td>
<td>0.81</td>
<td>3.13</td>
<td>4.24</td>
<td>1.04</td>
<td>1.47</td>
<td>1.18</td>
</tr>
<tr>
<td>12.00</td>
<td>0.84</td>
<td>3.17</td>
<td>4.14</td>
<td>1.08</td>
<td>1.51</td>
<td>1.21</td>
</tr>
<tr>
<td>13.00</td>
<td>0.80</td>
<td>3.26</td>
<td>4.05</td>
<td>1.23</td>
<td>1.60</td>
<td>1.23</td>
</tr>
<tr>
<td>14.00</td>
<td>0.86</td>
<td>3.18</td>
<td>3.91</td>
<td>1.11</td>
<td>1.48</td>
<td>1.20</td>
</tr>
<tr>
<td>15.00</td>
<td>0.88</td>
<td>3.24</td>
<td>3.73</td>
<td>1.10</td>
<td>1.47</td>
<td>1.21</td>
</tr>
<tr>
<td>16.00</td>
<td>0.90</td>
<td>3.22</td>
<td>3.60</td>
<td>1.09</td>
<td>1.45</td>
<td>1.20</td>
</tr>
<tr>
<td>17.00</td>
<td>0.91</td>
<td>3.29</td>
<td>3.48</td>
<td>1.06</td>
<td>1.42</td>
<td>1.19</td>
</tr>
<tr>
<td>18.00</td>
<td>0.91</td>
<td>3.31</td>
<td>3.40</td>
<td>1.07</td>
<td>1.43</td>
<td>1.21</td>
</tr>
<tr>
<td>19.00</td>
<td>0.93</td>
<td>3.35</td>
<td>3.40</td>
<td>1.07</td>
<td>1.43</td>
<td>1.21</td>
</tr>
<tr>
<td>20.00</td>
<td>0.94</td>
<td>3.39</td>
<td>3.41</td>
<td>1.07</td>
<td>1.43</td>
<td>1.21</td>
</tr>
</tbody>
</table>

**Variance**

-0.49

### Standard deviation

<table>
<thead>
<tr>
<th>Species</th>
<th>Body length</th>
<th>Ranked length</th>
<th>Wn/Wn-1</th>
<th>Randomized</th>
<th>Wn/Wn-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 3 5 2</td>
<td>4.61602</td>
<td>3.33436</td>
<td>4.73984</td>
<td>3.52019</td>
</tr>
<tr>
<td>2</td>
<td>2 3 5 0</td>
<td>5.08762</td>
<td>4.34326</td>
<td>5.66870</td>
<td>4.73984</td>
</tr>
<tr>
<td>3</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>4</td>
<td>2 3 5 0</td>
<td>5.34882</td>
<td>6.77584</td>
<td>6.38778</td>
<td>5.01155</td>
</tr>
<tr>
<td>5</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>6</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>7</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>8</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>9</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>10</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>11</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>12</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>13</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>14</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>15</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>16</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>17</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>18</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>19</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
<tr>
<td>20</td>
<td>2 3 5 0</td>
<td>5.10971</td>
<td>4.79115</td>
<td>6.09281</td>
<td>4.92775</td>
</tr>
</tbody>
</table>

**Variance**

0.003

**Mean**

0.04

**Standard deviation**

0.03

**Z**

-0.49
under the assumption of a normal random distribution of errors around the mean. $Z = -0.49$ indicates a random distribution of body weights.

The next example shows an often made error with this resampling approach. 20 ground beetle body length were measured in mm and the approach is the same as before. Because we used lengths instead of weights we use $q = W_n - W_{n-1}$ instead of $W_n / W_{n-1}$. Now we randomize and get $Z = 2.65$. The observed variance is higher than the randomized and we infer a significant aggregated distribution of body sizes. The error in this example is that we measured length in mm but calculated our random numbers at five decimals. Hence our observed $q$ values take only natural numbers. The randomized $q$ values take all real numbers between 3 and 23. They are less clumped and this small difference produced our significant result. Indeed, the table above shows the same calculations using randomized data at the mm level. Now our significant result vanishes. Indeed the original data were nothing more than random numbers.

The error occurred not only due to the difference in precision. The observed $q$ values were small natural numbers with a lot of tied ranks (identical values). Such data are always prone to Poisson errors.

Good programs that do resamplings or reshuflings are *PopTools, Sample, or Resampling procedures*. Of course *R* and *MatLab* allow for an easy programming of own algorithms.
13.2 Mantel test and Moran’s I

The Mantel test (after the America biostatistician Nathan Mantel) is a test of spatial autocorrelation. Assume an guild of ground dwelling species. We measured species composition at different sites. We have the suspicion that he change in species composition is related to changes in soil type. We also measured several variables of soil chemistry and soil type. Now we have two matrices. One of sites and soil property. The question is whether these two matrices are correlated. Hence, apart from a simple correlation between species richness and soil type we wish to include the spatial aspect. In order to do so we first construct two distance matrices containing the site distances of species composition and soil type. The next step is to calculate the coefficient of correlation between the two matrices. The procedure is shown in the next table. We use the two distance matrices. For convenience these contain normalized (Z-transformed) data. The Mantel test statistic is now the cross product term (the coefficient of correlation)

\[ r = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} Z(1)_{ij} Z(2)_{ij} \]

(13.2.1)

where \( n \) is the number of elements in the lower part (or upper) of the distance matrix \( = \frac{m(m-1)}{2} \) with \( m \) sites. In the example below the \( r \)-values is the sum of all xy products of the elements \( x \) and \( y \) of both lower parts of the distance matrices.

The elements of the distance matrices are not independent. Therefore it is not possible to test for significance directly. We test therefore via a permutation procedure and reshuffled the elements of one of the table randomly. At least 5000 such reshuffles are then used to compute mean coefficients and its confidence limits. For site numbers larger than 30 \( r \) might be tested directly using Fisher’s Z-transformation.

Of course the Mantel test can also be used to test for simple spatial autocorrelation. In this case the second matrix contains only the (Z-transformed) distances of the sites. An alternative is to rank the matrix entries prior to computing \( r \). This procedure is equivalent to Spearman’s rank order correlation. It should be used in cases when \( X \) and \( Y \) are non-linearly related.

Another often used statistic in spatial analysis is Moran’s \( I \). Moran’s \( I \) is similar to a correlation coefficient all applied to pairwise cells of a spatial matrix. It differs by weighting the covariance to account for

<table>
<thead>
<tr>
<th>Sites</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>11.586</td>
<td>-1.72</td>
<td>0.253</td>
<td>0.465</td>
<td>1.118</td>
<td>0.475</td>
<td>0.63</td>
<td>1.307</td>
<td>0.695</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.023</td>
<td>0.443</td>
<td>0.45</td>
<td>-0.5</td>
<td>1.096</td>
<td>-0.63</td>
<td>0.604</td>
<td>-0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.4</td>
<td>0.303</td>
<td>1.745</td>
<td>0.131</td>
<td>-0.56</td>
<td>0.63</td>
<td>-0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.253</td>
<td>0.443</td>
<td>-1.4</td>
<td>1</td>
<td>-0.71</td>
<td>0.546</td>
<td>0.557</td>
<td>0.079</td>
<td>0.173</td>
<td>0.169</td>
</tr>
<tr>
<td>5</td>
<td>0.465</td>
<td>-0.5</td>
<td>0.303</td>
<td>-0.71</td>
<td>1</td>
<td>-0.46</td>
<td>0.562</td>
<td>1.314</td>
<td>-0.44</td>
<td>-0.45</td>
</tr>
<tr>
<td>6</td>
<td>1.118</td>
<td>-0.5</td>
<td>1.745</td>
<td>0.546</td>
<td>-0.46</td>
<td>1</td>
<td>0.579</td>
<td>0.103</td>
<td>0.221</td>
<td>-0.48</td>
</tr>
<tr>
<td>7</td>
<td>0.475</td>
<td>1.096</td>
<td>0.131</td>
<td>0.557</td>
<td>0.562</td>
<td>0.579</td>
<td>1</td>
<td>0.264</td>
<td>0.766</td>
<td>1.004</td>
</tr>
<tr>
<td>8</td>
<td>0.63</td>
<td>-0.63</td>
<td>-0.56</td>
<td>0.079</td>
<td>1.314</td>
<td>0.103</td>
<td>0.264</td>
<td>1</td>
<td>-0.17</td>
<td>1.202</td>
</tr>
<tr>
<td>9</td>
<td>1.307</td>
<td>0.604</td>
<td>0.638</td>
<td>0.173</td>
<td>-0.44</td>
<td>0.221</td>
<td>0.766</td>
<td>-0.17</td>
<td>1</td>
<td>1.143</td>
</tr>
<tr>
<td>10</td>
<td>0.695</td>
<td>-0.59</td>
<td>-0.82</td>
<td>0.169</td>
<td>-0.45</td>
<td>-0.48</td>
<td>1.004</td>
<td>1.202</td>
<td>1.143</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sites</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>11.586</td>
<td>-1.72</td>
<td>0.253</td>
<td>0.465</td>
<td>1.118</td>
<td>0.475</td>
<td>0.63</td>
<td>1.307</td>
<td>0.695</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.023</td>
<td>0.443</td>
<td>0.45</td>
<td>-0.5</td>
<td>1.096</td>
<td>-0.63</td>
<td>0.604</td>
<td>-0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.4</td>
<td>0.303</td>
<td>1.745</td>
<td>0.131</td>
<td>-0.56</td>
<td>0.63</td>
<td>-0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.253</td>
<td>0.443</td>
<td>-1.4</td>
<td>1</td>
<td>-0.71</td>
<td>0.546</td>
<td>0.557</td>
<td>0.079</td>
<td>0.173</td>
<td>0.169</td>
</tr>
<tr>
<td>5</td>
<td>0.465</td>
<td>-0.5</td>
<td>0.303</td>
<td>-0.71</td>
<td>1</td>
<td>-0.46</td>
<td>0.562</td>
<td>1.314</td>
<td>-0.44</td>
<td>-0.45</td>
</tr>
<tr>
<td>6</td>
<td>1.118</td>
<td>-0.5</td>
<td>1.745</td>
<td>0.546</td>
<td>-0.46</td>
<td>1</td>
<td>0.579</td>
<td>0.103</td>
<td>0.221</td>
<td>-0.48</td>
</tr>
<tr>
<td>7</td>
<td>0.475</td>
<td>1.096</td>
<td>0.131</td>
<td>0.557</td>
<td>0.562</td>
<td>0.579</td>
<td>1</td>
<td>0.264</td>
<td>0.766</td>
<td>1.004</td>
</tr>
<tr>
<td>8</td>
<td>0.63</td>
<td>-0.63</td>
<td>-0.56</td>
<td>0.079</td>
<td>1.314</td>
<td>0.103</td>
<td>0.264</td>
<td>1</td>
<td>-0.17</td>
<td>1.202</td>
</tr>
<tr>
<td>9</td>
<td>1.307</td>
<td>0.604</td>
<td>0.638</td>
<td>0.173</td>
<td>-0.44</td>
<td>0.221</td>
<td>0.766</td>
<td>-0.17</td>
<td>1</td>
<td>1.143</td>
</tr>
<tr>
<td>10</td>
<td>0.695</td>
<td>-0.59</td>
<td>-0.82</td>
<td>0.169</td>
<td>-0.45</td>
<td>-0.48</td>
<td>1.004</td>
<td>1.202</td>
<td>1.143</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sites</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>11.586</td>
<td>-1.72</td>
<td>0.253</td>
<td>0.465</td>
<td>1.118</td>
<td>0.475</td>
<td>0.63</td>
<td>1.307</td>
<td>0.695</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.023</td>
<td>0.443</td>
<td>0.45</td>
<td>-0.5</td>
<td>1.096</td>
<td>-0.63</td>
<td>0.604</td>
<td>-0.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-1.4</td>
<td>0.303</td>
<td>1.745</td>
<td>0.131</td>
<td>-0.56</td>
<td>0.63</td>
<td>-0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.253</td>
<td>0.443</td>
<td>-1.4</td>
<td>1</td>
<td>-0.71</td>
<td>0.546</td>
<td>0.557</td>
<td>0.079</td>
<td>0.173</td>
<td>0.169</td>
</tr>
<tr>
<td>5</td>
<td>0.465</td>
<td>-0.5</td>
<td>0.303</td>
<td>-0.71</td>
<td>1</td>
<td>-0.46</td>
<td>0.562</td>
<td>1.314</td>
<td>-0.44</td>
<td>-0.45</td>
</tr>
<tr>
<td>6</td>
<td>1.118</td>
<td>-0.5</td>
<td>1.745</td>
<td>0.546</td>
<td>-0.46</td>
<td>1</td>
<td>0.579</td>
<td>0.103</td>
<td>0.221</td>
<td>-0.48</td>
</tr>
<tr>
<td>7</td>
<td>0.475</td>
<td>1.096</td>
<td>0.131</td>
<td>0.557</td>
<td>0.562</td>
<td>0.579</td>
<td>1</td>
<td>0.264</td>
<td>0.766</td>
<td>1.004</td>
</tr>
<tr>
<td>8</td>
<td>0.63</td>
<td>-0.63</td>
<td>-0.56</td>
<td>0.079</td>
<td>1.314</td>
<td>0.103</td>
<td>0.264</td>
<td>1</td>
<td>-0.17</td>
<td>1.202</td>
</tr>
<tr>
<td>9</td>
<td>1.307</td>
<td>0.604</td>
<td>0.638</td>
<td>0.173</td>
<td>-0.44</td>
<td>0.221</td>
<td>0.766</td>
<td>-0.17</td>
<td>1</td>
<td>1.143</td>
</tr>
<tr>
<td>10</td>
<td>0.695</td>
<td>-0.59</td>
<td>-0.82</td>
<td>0.169</td>
<td>-0.45</td>
<td>-0.48</td>
<td>1.004</td>
<td>1.202</td>
<td>1.143</td>
<td>1</td>
</tr>
</tbody>
</table>
spatial non-independence of cells with respect to distance. The measure is defined by

\[ I = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} z_i^2} \] 

(13.2.2)

where \( N \) is the number of matrix cells, and \( z \) the Z-transformed cell values and \( w \) the weighing coefficient \( w \) is frequently defined as

\[ w_{ij} = \frac{1}{(1 + d_{ij})^2} \] 

(13.2.3)

d is the difference between the cells \( i \) and \( j \). \( w \) is therefore an inverse distance function. Nearer cells therefore more influence on Moran’s I than cells spatially apart.

Statistic programs frequently calculate the statistical significance of I from a reshuffling of cells (a Monte Carlo approach). If cell values were randomly distributed (not spatially autocorrelated) the expected I is

\[ E_0(I) = \frac{-1}{N-1} \] 

(13.2.4)

The statistical significance of the difference of \( I - E_0(I) \) can be estimated from the variance of I under the assumption that the distribution of values is linearly random. The variance of I is then

\[ \sigma^2 = \frac{N[(N^2 - 3N + 3)S_1 - NS_2 + 3S_0^2] - b[(N^2 - N)S_1 - 2NS_2 + 6S_0^2]}{(N-1)^3S_0^2} - \frac{1}{(N-1)^2} \]

\[ S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_{ij} + w_{ji})^2 \]

\[ S_2 = \frac{1}{2} \sum_{i=1}^{N} \left( \sum_{j=1}^{N} w_{ij} + \sum_{j=1}^{N} w_{ji} \right)^2 \]

\[ K = \frac{\sum_{i=1}^{N} (x_i - \mu)^4}{\left( \sum_{i=1}^{N} (x_i - \mu)^2 \right)^2} = \frac{\sum_{i=1}^{N} z_i^4}{\left( \sum_{i=1}^{N} z_i^2 \right)^2} \]
14. Markov chains

Beside equation solving matrices have by far more biological applications. Assume you are studying a contagious disease. You identified as small group of 4 persons infected by the disease. These 4 persons contacted in a given time another group of 5 persons. The latter 5 persons had contact with other persons, say with 6, and so on. How fast does the disease spread in the population? To answer this question we first define a matrix describing the first contacts. You have four infected persons and 5 contact persons of the second group.

\[
A = \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
0 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 \\
\end{pmatrix}
\]

Hence person 1 of the first group contacted with person 2 of the second group. No. 2 of the first group contacted with No. 1, 2, and 4 of the second group and so on. Now you describe the second order contacts of group three with group two.

\[
B = \begin{pmatrix}
0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 \\
\end{pmatrix}
\]

To find the number of persons in group three that had (via group two) contact with infected persons of group one you have to multiply both matrices. We get as the result

\[
C = B \cdot A = \begin{pmatrix}
1 & 1 & 2 & 2 \\
0 & 0 & 0 & 1 \\
1 & 2 & 1 & 1 \\
0 & 1 & 0 & 0 \\
1 & 1 & 2 & 1 \\
1 & 2 & 1 & 1 \\
\end{pmatrix}
\]

How to interpret this result? From the computational scheme of a dot product of two matrices follows that the new elements of C result from all combinations of respective rows and columns of A and C. Hence the ones and twos denote indirect contacts of a person in the third group with a person of the first group. Person 1 of the third group had 6 indirect contacts (1+1+2+2), the persons 2 and 4 only one.

However, we can also use probabilities of infection instead of contacts. Say that any contact gives a probability of 0.3 that a person will be infected. We have to replace the one with 0.3 to get the probability that persons in contact with infected persons get infected. Our model becomes

\[
\begin{pmatrix}
0 & 0.3 & 0.3 & 0.3 \\
0 & 0 & 0.3 & 0.3 \\
0.3 & 0.3 & 0 & 0.3 \\
0.3 & 0 & 0 & 0.3 \\
0.3 & 0.3 & 0.3 & 0 \\
0.3 & 0 & 0 & 0.3 \\
\end{pmatrix}
\]
Hence person 1 of the third group has a probability of 0.54 of being infected. Of course this method can be applied to further groups or generations (if we interpret the groups as generations). By this we get probabilities of occurrences of initial events in subsequent time windows. The matrix multiplication allows for the prediction of infections during epidemics.

**Markov chains**

The above discussion leads immediately to the concept of Markov chains (after the Russian mathematician Andrei Markov, 1856-1922). A Markov chain is a sequence of random variables in which the future variable is determined by the present variable but in independent of the way in which the present state arose from its predecessors. Hence if we have a series of process states the value of state n is determined by only two things. The value of state n-1 and by a rule that tells how state n-1 might transform into state n. Most often these rules contain probabilities. Then state n-1 goes with probability $p_i$ into any state $i$ of n. Hence in a Markov chain states prior than the previous do not influence the future fate of the chain. This is why Markov chains are often said to be without memory.

Take for instance a gene that has three alleles A, B, and C. These can mutate into each other with probabilities that are given in Fig. 14.1. A mutates into B with probability 0.12 and into C with probability 0.2. Hence with probability $1 - 0.12 - 0.2 = 0.68$ nothing happens. We can these so-called transition probabilities write in a matrix form.

$$P = \begin{pmatrix} 0.68 & 0.07 & 0.1 \\ 0.12 & 0.78 & 0.05 \\ 0.2 & 0.15 & 0.85 \end{pmatrix}$$

This matrix that gives the transition probabilities is called the transition matrix. The sum of all matrix rows must add to 1, the sum of all probabilities. This is a general feature of all probability (stochastic) matrices. If we now take the initial allele frequencies we can compute the frequencies of the alleles in the next generation. Assume we have initial frequencies of $A = 0.2$, $B = 0.5$, and $C = 0.3$. This gives a vector of the form $X_0 = \{0.2, 0.5, 0.3\}$. The frequencies in the next generation are computed from

$$X_1 = P \cdot X_0 = \begin{pmatrix} 0.68 & 0.07 & 0.1 \\ 0.12 & 0.78 & 0.05 \\ 0.2 & 0.15 & 0.85 \end{pmatrix} \begin{pmatrix} 0.2 \\ 0.5 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.201 \\ 0.429 \\ 0.37 \end{pmatrix}$$

Again, the new frequencies of $A = 0.201$, $B = 0.429$, and $C = 0.37$ add up to zero. If we multiply two probability matrices the resulting matrix is again a probability matrix.
If we continue the process we get \( X_3 = PX_2, X_4 = PX_3 \). In general the frequencies of a Markov chain after \( n \) states starting from the initial conditions \( X_0 \) and determined by the transition matrix \( P \) is given by

\[
X_n = P^n \cdot X_0 = U \cdot A^n \cdot U^{-1} \cdot X_0
\]

(14.1)

Of course this law is very similar to recursive processes leading to exponential distributions. It is a generalization. Equation 14.1 defines the simplest form of a Markov chain process.

We also see that state \( n+1 \) is only dependent on state \( n \). This property serves even as the general definition of a Markov process. The probability \( i \) of a state \( X_n \) with respect to the previous states \( X_1 \) to \( X_{n-1} \) is the same as the probability of \( X_n \) with respect to state \( X_{n-1} \) only. The previous states have no influence any more. Mathematically written

\[
p(X_n = i | X_{n-1}, X_{n-2}, X_{n-3}, \ldots X_{n-1}) = p(X_n = i | X_{n-1})
\]

(14.2)

Does our mutation process above reach in stable allele frequencies or do they change forever? This question can be answered twofold. Does the frequency distribution remains constant or does the process eliminate one or more alleles? The first question is whether the frequencies of the alleles remain constant. In this case the following condition must hold

\[
X_{n+1} = X_n = P \cdot X_n
\]

This can be written in terms of eigenvectors

\[
P \cdot X_n = I X_n \rightarrow (P - I) \cdot X_n = 0
\]

(14.3)

\( P_n \) is called the stationary state. This state is defined by the eigenvector \( U \) of the transition matrix \( P \) with the largest eigenvalue. This is scaled to \( \lambda = 1 \). \( X_n \) is called the steady-state or equilibrium vector. The Excel example beside shows the transition matrix of three alleles. \( \lambda_3 = 1 \) and the third eigenvector \( U_3 \) defines the stationary state, that is the frequency distribu-
tion the process will end in. Note that the eigenvectors are normalized to have the length one. To get the frequencies (probabilities) we have to divide $U_3$ through the sum of it’s values. Fig. 14.2 shows that in our case the allele frequencies quickly converge to the steady state.

Do all Markov chains converge? We look at several important special cases. Fig. 14.3 shows a graphical representation of a Markov chain with four states. Given are the transition probabilities. The missing probabilities can be inferred from the scheme. We see that state D cannot be reached from any other state. It forms a closed part of the whole chain. If a chain does not contain closed subsystems it is called irreducible. In such a system all states can be reached. Fig. 14.4 now shows a simple example of a periodic chain. The whole chain forms a circle. Fig. 14.3 shows an aperiodic chain. Fig. 14.3 shows also two other important concept. First the states A, B, and C are recurrent that means it is sure that a finite time (even a very long time) the process returns to the initial state. State D is not recurrent. In some chains there is only a certain probability that the chain returns to a previous state. These chains are called transient. An important class of finite Markov chains are now recurrent and aperiodic chains. These are called ergodic. The chain of Fig. 14.3 is ergodic (except of state D), chain of Fig. 14.4 not (it is periodic). The next table shows the transition matrices, the eigenvalues, and the eigenvectors of both chains. For both chains $\lambda = 1$ exists, but only chain 14.3 converges. The probability matrix theorem now tells that every irreducible ergodic transition matrix (that is the matrix containing only probabilities) has a steady state vector $T$ to which the process converges.

$$\lim_{k \to \infty} P^k X_0 = T$$

Fig. 14.2 points also to another problem. How fast do Markov processes converge to the steady state.
The time to convergence is obviously connected to the probabilities in the matrix. The recurrence time of a state $i$ is now defined as the mean time at which the process returns to $i$. It can be shown that the recurrence times $T_i$ of any state $i$ are inversely related to the stationary probabilities $\pi_i$.

$$T_i = \frac{1}{\pi_i} \quad (14.5)$$

The mean time to stay in any state is of course the inverse of the probability not to leave the state. Hence in Fig. 14.3 the recurrence time of state A is $T = (0.38 + 0.51 + 0.77)/0.38 = 4.33$ steps.

The question how long it will take to reach the stationary state is identical to the question what function describes the Fig. 14.3 and how to calculate the parameter values of the function. With some mathematics one can show that it is an exponential function of the type

$$p = ae^{-bt} \quad (14.6)$$

There are no simple solutions for the parameters.

A typical application of Markov chains in biology is succession. For instance gravel pits have a distinct mosaic plant community structure. Abandoned pits go through series of successional stages. If we now map the plant distribution of the gravel pit immediately after abandonment we get a matrix of initial states. From other studies it is known with what frequency certain structural elements transform into others. Hence we have a transition matrix. We can now describe the whole process of succession by a Markov chain model. In this case we have a matrix of the initial stage and the transition matrix. The model looks as follows

$$X_t = P^t \cdot X_0 \quad (14.7)$$

We consider six different plant community classes and have the following transition matrix. Our initial stage is given by Fig. 14.6 where the six communities are represented by different colours. After $t = 100$ states (Fig. 14.7) our map changed totally. Community types 1 and 2 dominate, 5 and 6 vanished. After even 1000 states (Fig. 14.8) not much had changed. However, very slowly the frequency of community 3 raises. The proportion of
community 4 remains stable but type 2, which dominated the intermediate stage of succession, decreases.

How to compute these pictures. Either you apply a commercial program that computes Monte Carlo simulations and Markov chains or you write a program for your own. In our case I used a self written program that iterates equation 14.2. For shorter series you can run a math program iteratively. Above a simple Fortran solution is shown with which I computed the matrices on the left side.

Markov chains find application in probability theory. Assume for instance you have a virus with N strains. Assume further that at each generation a strain mutates to another strain with probabilities \( a_{i\rightarrow j} \). The probability to stay is therefore \( 1-\sum a_{i\rightarrow j} \). What is the probability that the virus is after k generations the same as at the beginning. This can be modelled by a Markov chain with the following transition matrix

\[
p = \begin{pmatrix}
1 - \sum a_{1,i\rightarrow 1} & \cdots & a_{1,N}

\vdots & \ddots & \vdots

a_{N,1} & \cdots & 1 - \sum a_{N,i\rightarrow 1}
\end{pmatrix}
\]

We get the desired probability from the matrix element \( p_{11} \) of \( P^k \). Hence

\[
P^k = U \cdot \lambda^k \cdot U^{-1}
\]

The next table shows the respective Excel solution for a given transition matrix using the Matrix add in for \( k = 5 \). The requested probability is \( p_{11} \) 0.23. Markov chains are therefore ideal tools for calculating probabilities if we have multiple pathways to reach certain states. Particularly, they describe the probability to get in k steps from state A to state B if the transition probabilities can be described using a transition matrix.

Random walk models

A special example of Markov chains are random walks

Wee know already that random walks are defined by the general state equation

\[
N_{t+1} = N_t + \text{ran}
\]

The state \( N_t \) is only defined by the previous state and a probability function of change. Typical examples of such random walks are for instance animal movements. Let’s consider an animal A being at place \( x_0 \). In a next step it might turn to left with probability \( p_l \), turn to right with probability \( p_r \) or walk straight on with probability \( p_s \). Our random walk model looks at follows

\[
\begin{pmatrix}
P^A \\
P^B \\
P^C
\end{pmatrix}
= \begin{pmatrix}
P^A \\
P^B \\
P^C
\end{pmatrix}
\begin{pmatrix}
0.230675 \\
0.47613 \\
0.293195
\end{pmatrix}
+ \begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
+ \begin{pmatrix}
0.3 \\
0.15 \\
0.15
\end{pmatrix}
\begin{pmatrix}
0.338197 \\
0.561803 \\
1
\end{pmatrix}
\begin{pmatrix}
0.230675 \\
0.47613 \\
0.293195
\end{pmatrix}
\begin{pmatrix}
0.338197 \\
0.561803 \\
1
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.814984 \\
-0.450512 \\
-0.364472
\end{pmatrix}
\begin{pmatrix}
0.004424 \\
0.055966 \\
0
\end{pmatrix}
\begin{pmatrix}
0.20048 \\
0.51785 \\
0.28167
\end{pmatrix}
\begin{pmatrix}
0.258105 \\
0.43003 \\
0.311865
\end{pmatrix}
This is a recursive equation that describes a directional process. It’s two dimensional equivalent would have the form

\[
\begin{pmatrix}
  p_1 \\
p_s \\
p_r
\end{pmatrix}
\]

\[x_n = x_{n-1}
\]

\[
\begin{pmatrix}
  p_1 & p_2 \\
p_3 & p_4
\end{pmatrix}
\]

where the columns define forward or backward walk.

Recursive probability functions are also special cases of Markov chains. We can’t know, where the animals ends his walk. But we might use a model of 5000 animals and try to give probabilities of the outcome. Such a Monte Carlo simulation provides us with a frequency distribution of end points of the random walk. Then we can tell that a typical animal ends his walk there or there. To model this we need the possible area into our animal can walk, the number of possible states. This is indicated by the green area in Fig. 14.9. If this number is finite we speak of a bounded random walk. What is if the animal reaches the lower or upper boundary? In the Figure the animal is reflected from the barrier.
15. Time series analysis

This last lecture deals with the analysis of biological time series. Such time series can be changes in population sizes of bacteria, animals, or plants in time, biological rhythms like the circadian rhythm, patterns of brain activity in time, or diversity patterns during evolution. Fig. 15.1 shows a hypothetical time series, the change of abundance of a species in time. Such a time series can be described by four basic parameters. The mean abundance ($\mu = 14.6$), the standard deviation ($\sigma^2 = 4.7$), and the upper and lower boundaries. The lower boundary of such an abundance time series is of course 0, the upper boundary is often not defined. A first important measure of variability in time series is the coefficient of variation. Because the standard deviation still depends on the mean ($\sigma^2 \propto \mu$) CV is defined as $\sigma / \mu$ and is therefore a standardized measure of variability. In our case $CV = \sqrt{4.7 / 14.6} = 0.32$.

A general description of such a time series is

$$N_{t+1} = f(N_{t-\tau}) + b$$  \hspace{1cm} (15.1)

This is a very general recursive function. b is some random variate. If the time delay $\tau$ is zero, the next generation depends solely on the state of the previous. Let’s study eq. 15.1 in more detail. In the simplest case $f(N) = N$. Hence

$$N_{t+1} = N_t + b$$

This is a random walk and a typical time series looks as in Fig. 15.1. b is often defined as $b = ran(-c,c)$. The coefficient of variation of such a random walk has typically values between 0 and 1. The higher $N_0$ is in relation to the range of b, the lower CV will be.

A time series might also arise from a multiplicative process

$$N_{t+1} = a N_t$$

where a is typically defined as $a = ran(1-c, 1+c)$. The multiplicative process leads frequently to a higher variability with CV-values around or above 1. Of course the multiplicative random walk can be transformed into an additive one by taking the logarithms of abundance N (Fig. 15.2).

$$\ln(N_{t+1}) = \ln(N_t) + \ln(a)$$

If a is small with respect to N the variability will be small, a large factor a results instead in a large variability that appears to be chaotic. In chapter 7 of part A we saw that a proportional rescaling process has the form

$$\sigma^2 = a \mu^2$$

This is a power function. We can apply this equation to the analysis of time series. Fig. 15.3 shows such series. A time
series can be seen as a complex wave that consists of many composite waves of different wavelengths (and frequencies). It has therefore much in common with self-similar processes that are also described by power functions. We apply a ruler length technique. We assume that this time series was formed by a superposition of waves.

\[ N(t) = \sum (a_i \sin(\psi_i t) + b_i \cos(\psi_i t)) \]

We now lay series of time windows onto this series and measure for each window length (ruler length) \( l \) the difference \( \Delta N \). \( \Delta N \) is the difference between the element \( N_t \) and \( N_{t+l} \). For each windows length \( l \) we measure \( \mu_{\Delta N} \) and \( \sigma^2_{\Delta N} \). Then we plot variance (often also called spectral density) against \( \mu \). Additionally we plot \( \sigma^2_{\Delta N} \) against ruler length \( l \). Variance and ruler length \( l \) of the time series should be connected by an equation equivalent to Taylor’s power law.

\[
\sigma^2 \propto l^\beta = (1/f)^H = (1/f)^{2H}
\]  

(15.2)

with \( H \) being the so-called Hurst exponent. Figs. 15.3 and 15.4 show that variance and mean and variance and ruler length are both related by power functions. Our exponent \( \beta \) gives us an impression how irregular our time series is. \( \beta \) (Fig. 15.5) is in our case approximately 1 and we conclude that we deal with a Poisson process.

\( H \) is related to the fractal dimension of the curve and is given by \( H = 2 - D \). In our case the fractal dimension \( D \) of the curve would be \( H = 0.86/2 = 2 - D \rightarrow D = 1.57 \). Calculating a fractal dimension \( D \) by this method is called the second moment technique and is frequently used in the study of time series. Fig. 15.5 is called the variogram.

Most often only half of the variance is used to estimate \( H \). Now we deal with a semivariogram. We use again the window technique of Fig. 15.3 and compute

\[
\gamma(l) = \frac{1}{2} \sigma^2(l) = \frac{1}{2} \left( \frac{\sum_{i=1}^{N} (x_{i+l} - x_i)^2}{N} \right)
\]

(15.3)

For a fractal wave \( \gamma(l) \) and \( l \) (the wavelength considered) and \( \gamma(l) \) and \( 1/l = f \) (the frequency considered) should be related by a power function. The exponent \( 2H \) is again related to the fractal dimension by \( H = 2 - D \).

Now look at Fig. 15.6. A periodic process (a time series) is step by step disturbed by irregularities, so called noise. I used the function
We might think of an audio wave that becomes less and less clear. At the end it looks totally chaotic. Each time we compute the variogram and determine the Hurst exponent. For $a = 0$ (2.6 A) the function is smooth. The slope of the variogram is approximately 2. From this we infer that the fractal dimension of this time series is $D = 1$. Of course, it is a smooth line. Our process is similar to the random walk above. This type of noise is called brown noise (after the Brownian motion in physics that can be described by a random walk model). Only the long wavelength contribute to the noise. In B to D the influence of the random variate is step

$$N(t) = \sum_{i=1}^{4} (\sin(i^2 t)^2 + ran(-a,a))$$

Fig. 15.6
by step enlarged. The Hurst exponents decrease. In B the exponent is approximately 1, hence the fractal dimension is \( D = 1.5 \). In this case variance is a linear function of wavelength. In other words all wavelengths take proportionally equally part in total variance. The light equivalent is pink and we speak of pink noise (or 1/f noise). Pink noise reflects linear scale dependence of noise and is often found in nature. In C the exponent decreased to appr. 0.5, \( D = 1.75 \). The lower wavelengths (higher frequencies) are responsible for a large part of total variance, but variance increases still with wavelength. This is the case in red light and for \( \beta \)-values between 0 and 2 we generally speak of red noise. Lastly in D the slope is zero hence \( D = 2 \). All wavelengths (frequencies) contribute equally to total variance. In the light analogy this would be white and we speak of white noise. Note that even black noise (\( H = 1.5 \)) exists (try to model this). In rare cases \( \beta \) becomes negative. We speak of blue noise.

Why is spectral analysis important? Recent investigations in fractal geometry showed that most natural phenomena show some mix of long term and short term variability. They can be described by \((1/f)^\beta\) power laws to detect whether long term or short term processes dominate. In biological processes Hurst exponents were commonly found to range between 0.5 and 1 leading to fractal dimensions of the time series between 1 and 1.5. This is not white noise and we should be able to detect long term structures in existing time series.

Look at Figs. 15.6 B and C. I did not take all of the data points but only the last six window sizes of the time series. Too large window sizes don’t see small wavelengths, Too small window sizes pay too much emphasis on small scale patterns. The model used to construct the time series had high small scale variation. Hence the series exhibits at smallest window sizes another pattern than at larger sizes. Such scale inhomogeneities (scaling regions) have to be considered in every time series analysis. Nearly all natural series do not show the same pattern over its entire length (for instance are not constantly fractal, or have different fractal properties at different scales). The aim of any time series analysis is also to detect such scaling regions were patterns shift. In our case we infer that our time series changes its properties above a wavelength of appr. 20. Below its white noise, above red to pink noise.
Now we look back to Fig. 15.6. We plot for the same four time series $N_t$ against $N_{t+1}$. Fig. 15.7 shows the results and we see that both values are correlated except for the white noise case. Therefore we can study time series by an autoregression analysis. In general we try to find a function of the type

$$N_t = aN_{t-\tau} + b$$  \hspace{1cm} (15.4)

or even more general

$$N_t = \sum_{i=1}^{\tau-1} a_i N_{t-i} + b$$  \hspace{1cm} (15.5)

Hence autocorrelation depends on the assumption that nearby objects (data points) are more closely related than distant ones. We try to model our time series by an algebraic function. Autoregression analysis also tells us whether there are long term regularities in our data. The above computations are easily done with Excel. A Statistica solution for the pink noise data (Fig. 15.6, 7 B) and three time lags $\tau$ shows the next table. The data are highly autocorrelated, the slope $r$ is nearly identical with the Excel solution.

Another example. Fig. 15.8 shows bird counts per hour of the monitoring programme for North American birds between 1957 and 2001 (data from Santa Cruz bird club organization). Do bird numbers fluctuate at random or are they regulated? In particular are they regulated by their own abundance? In such a case we speak of density dependence. Our aim is to detect whether high abundances are followed by lower abundances in the next year and whether low abundances are followed by high abundances in the next year.

In this case abundance itself would be the regulating agent. To infer this we undertake an autoregression analysis. Figure 15.9 shows a plot of $N_{t+1}$ against $N_t$. We detect no significant autocorrelation. But be carefully. Maybe there are larger time lags and we should better use the general autocorrelation model of eq. 2.5. In this case we have to use a statistic program. Statistica computes with the present data set a model with 5 time lags. None of the autoregression functions is statistically significant. We are not able to detect any density dependent regulation in our bird data set.
17. Statistics and the Internet

Here are several useful links to internet pages that contain information about statistics

**Handbooks**

Hyperstat (large online statistic handbook and glossary): http://www.davidmlane.com/hyperstat/glossary.html
Statistica online handbook (the complete Statistica handbook and glossary for download) http://www.statsoft.com/textbook/stathome.html
StatPrimer (an online statistic handbook by Bud Gerstman) http://www.sjsu.edu/faculty/gerstman/StatPrimer/
Introductory statistics and multivariate statistics (Two very extensive online statistical textbook covering all sorts of useful techniques with many animations) http://www.psychstat.missouristate.edu/multibook/mlt00.htm;
Statistik für Psychologinnen und Psychologen (a very well designed online script of statistics with many tables and examples) http://www.psychologie.uni-freiburg.de/signatures/leonhart/skript/
Non-parametric tests. Many informations at http://statpages.org/.

**Software sites**

Statpac (freeware programs for basic calculations) http://www.davidmlane.com/hyperstat/glossary.html
XLStat (statistic module for Excel, free trials) http://www.davidmlane.com/hyperstat/glossary.html
Virtual Library in Statistics (Many links to interesting web pages and programs) http://www.stat.ufl.edu/vlib/statistics.html
UCLA distribution calculator (very good probability and table calculator for many basic statistical distributions) http://www.stat.ucla.edu/~dinov/courses_students.dir/Applets.dir/OnlineResources.html.
UCLA statistics calculators (a large collections of basic statistics calculators and tables, easy to use) http://calculators.stat.ucla.edu/
Neville Hunt’s homepage (if you want to produce statistic tables by yourself, here you find a very good instruction how to make Excel tables for important statistical distributions) http://www.mis.coventry.ac.uk/~nhunt/tables.htm
Past. (a very good simple to use statistic package) http://folk.uio.no/ohammer/past/.
PopTools. (a very useful free Excel add in for spatial analysis and more) http://www.cse.csiro.au/poptools/.
Free statistic packages. (a collection of software sites) http://statpages.org/javasta2.html.
SAM (a very good program for spatial analysis) http://www.ecoevol.ufg.br/sam/.
Matrix (a very good matrix algebra add in for excel) http://digilander.libero.it/foxes/index.htm

**Libraries**

Virtual Library in Statistics (Many links to interesting web pages and programs)
http://www.stat.ufl.edu/vlib/statistics.html
18. Links to multivariate techniques

**GLM:** A nice introductory description at http://trochim.human.cornell.edu/kb/genlin.htm

**Analysis of variance:** All about at http://faculty.vassar.edu/lowry/vsanova.html and http://www.statsoftinc.com/textbook/stanman.html

**Multiple regression:** all about at http://www.statsoftinc.com/textbook/stmulreg.html and http://www2.chass.ncsu.edu/garson/pa765/regress.htm
A short course at http://cs.gmu.edu/ene

**Path analysis:** All about at http://luna.cas.usf.edu/~mbrannic/files/regression/Pathan.html

**Cluster analysis:** All about at http://www.clustan.com/what_is_cluster_analysis.html
A nice introduction at http://149.170.199.144/multivar/hc.htm

**Discriminant analysis:** All about at http://www.statsoftinc.com/textbook/stdiscan.html and http://www2.chass.ncsu.edu/garson/pa765/discrim.htm
Look also at http://www.doe-mbi.ucla.edu/~parag/multivar/da.htm

**Factor analysis:** All about at http://www2.chass.ncsu.edu/garson/pa765/factor.htm and http://www.statsoftinc.com/textbook/stfacan.html
Look also at http://www.doe-mbi.ucla.edu/~parag/multivar/pca.htm. And http://www-psychology.concordia.ca/fac/bukowski/psy732/related%20files/Lecture%208/Lect%208.ppt


19. Some mathematical sites

**Libraries**
Mathematics Virtual Library (Many links to interesting web pages and programs) http://www.math.fsu.edu/Science/math.html
Math on the web (Search engine for all sorts of mathematics) http://www.ams.org/mathweb/mi-mathinfo07.html
The Math Archive (Many links to interesting web pages and programs) http://archives.math.utk.edu/
Eric Weisstein’s Mathematics (an online mathematics dictionary) http://mathworld.wolfram.com/
The Internet Mathematics library (a large collections of topics for pupils and students, math-beginners) http://mathforum.org/library/
Mathematic resources (a large compilation of math internet pages) http://www.clifton.k12.nj.us/cliftonhs/chsmath.html

**Software sites**
The Windows software collection (public domain and freeware) http://archives.math.utk.edu/software/msdos.directory.html (contains many very nice programs)
Derivative calculator (a nice small but quite effective program to computing derivatives) http://cs.jsu.edu/mcis/faculty/leathrum/Mathlets/derivcalc.html
JAVA Mathlets for Math Explorations (a nice collection of small math programs for everybody) http://cs.jsu.edu/mcis/faculty/leathrum/Mathlets/
The integrator (a small but effective integration program) http://integrals.wolfram.com/index.jsp.
Modelowanie reczwistości (a nice Polish page with a program collection and many further links) http://www.wiw.pl/modelowanie/
Maxima. The oldest but very good free ware math program. http://maxima.sourceforge.net/