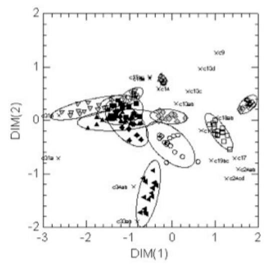
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION
1. THE BASICS

Prof. Eugenio Parente
Scuola di Scienze Agrarie- Università della Basilicata



Before we start

- This course will be delivered using vyew (http://vyew.com)
- Up to 20 students can attend on-line, but registration is needed
- Only registered students at the Università degli Studi della Basilicata are required to turn in a report for grading. I am willing to accept 4 more students who are willing to submit a report for grading
- The course timetable will be published as soon as the number of on-line students and their physical location is known

Objective

- to provide an informal and essentially practical introduction to statistical techniques used in food science and nutrition for:
- explorative data analysis
  - **principal component analysis (PCA)**
  - **multidimensional scaling**
  - correspondance analysis
- unsupervised pattern recognition
  - **hierarchical cluster analysis**
  - **non-hierarchical cluster analysis**
  - *unsupervised artificial neural networks*
- supervised pattern recognition
  - linear discriminant analysis
  - *supervised artificial neural networks*
- Inferential tools
  - **Principal Component Regression**
  - **Partial Least Square Regression**

Outline

- Introduction (a word of caution)
- Multivariate data and multivariate statistics.
- Know thy batch: bad data, good data, graphical exploration of multivariate data.
- Data transformation and standardization.
- Principal component analysis and MDS.
- Hierarchical and non-hierarchical cluster analysis.
- Neural networks
- Discriminant analysis
- Principal component regression and Partial Least Square Regression
- **Workshop**. Supervised individual and group work: analysis of multivariate data in Chemistry, Biology, Biotechnology, Food Science and Nutrition…

Prerequisites

- A BSc in Agriculture, Food Science, Chemistry, Biology, Biotechnology…
- At least 5 ECTS credits in Mathematics and 3 ETCS credits in Statistics
- Ability to use spreadsheet and statistical software packages under Windows, MacOS or Unix/Linux operating systems
- A basic knowlegde of technical English language (for speakers of English as a second language a B1 or B2 level is suggested)

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

Learning goals

At the end of the lectures the students should:

1. understand the principles of explorative multivariate data analysis and unsupervised pattern recognition
2. be able to carry out an exploratory data analysis (including PCA, MDS, CA and graphical documentation, as appropriate) on a multivariate data set
3. be able to carry out supervised and unsupervised pattern recognition on a multivariate data set
4. Understand the basics of PCR and PLS and carry out a simple analysis
5. be able to document their analysis and provide a graphical representation of their data

What I cannot teach you: how to look for relevant literature, how to use statistical software

Grading (just in case)

- At the end of the lectures, students who have registered for workshop activity must submit a report on multivariate statistical analysis of a problem pertaining to their research activity: the papers will be graded for appropriateness, completeness, originality on a A-E scale
- If no paper is turned in credits will be given for attendance only but no grade will be returned

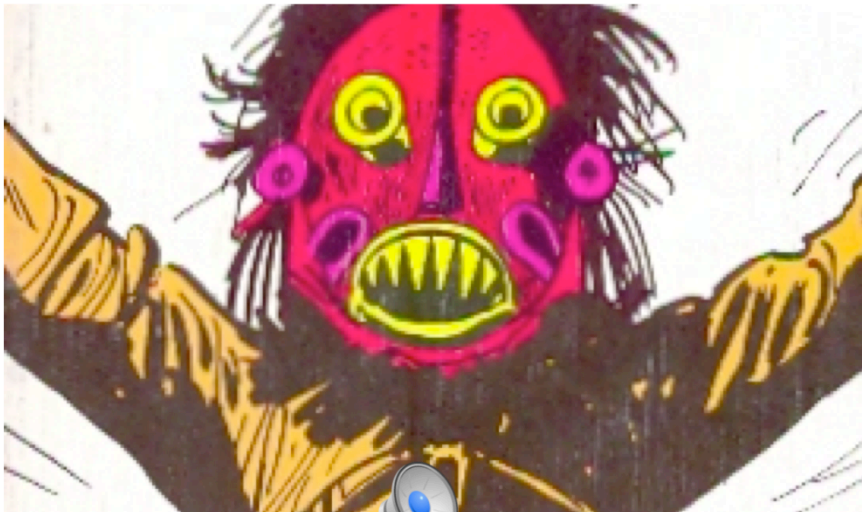
10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

A WORD OF CAUTION

What am I doing here? (you, not me)

- What this course will do for you
  - You will get a general idea of multivariate statistical methods
  - You will be able to select the right method for your problem (hopefully)
  - You will be able to discuss your problem with a statistician
  - You might get interested in multivariate statistics
- What this course will not do for you
  - It will not turn you into a statistician (whew!)
  - It will not teach you to use statistical software packages (umpf!)
  - It will not (magically) turn your raw data into publishable stuff

But then, who are you (me, not you)?



But then, who are you (me, not you)?

"Cheshire Puss," she began, rather timidly, as she did not at all know whether it would like the name: however, it only grinned a little wider. "Come, it's pleased so far," thought Alice, and she went on, "Would you tell me, please which way I ought to walk from here?”

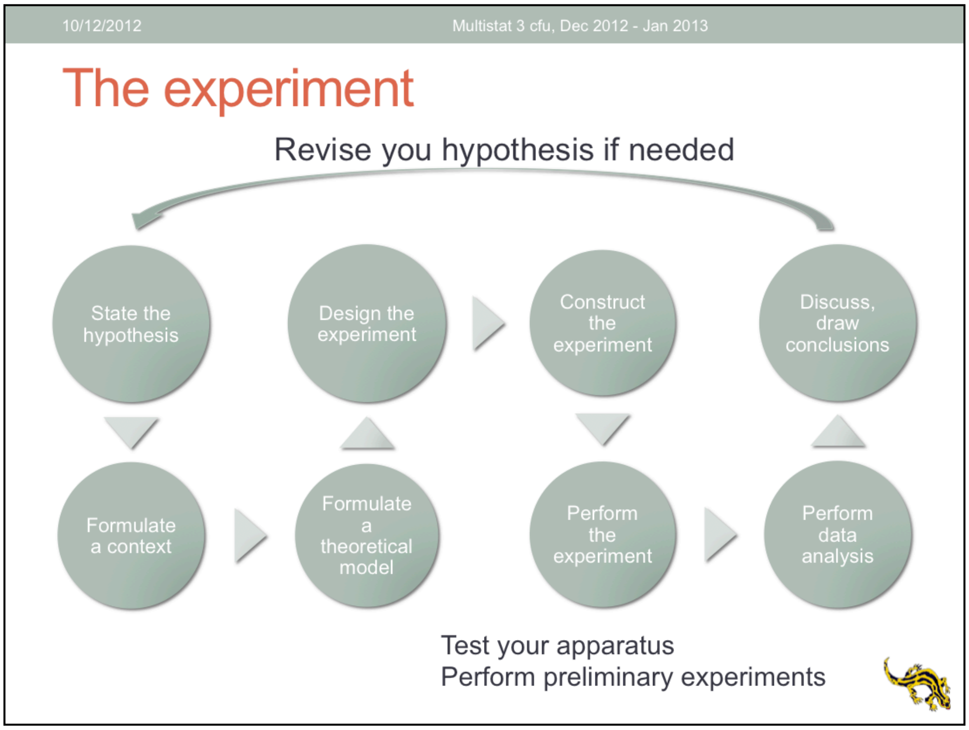
"That depends a good deal on where you want to get to," said the Cat.

"I don't much care where - " said Alice.

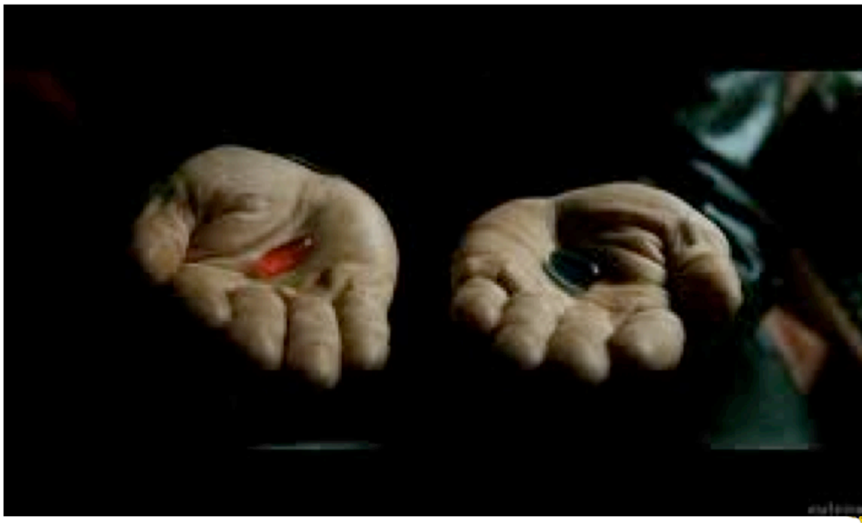
"Then it doesn't matter which way you walk," said the Cat.

" - so long as I get somewhere," Alice added as an explanation.

"Oh, you're sure to do that," said the Cat, "if you only walk long enough."


10/12/2012
Multistat 3 cfu, Dec 2012 - Jan 2013
The experiment
Revise you hypothesis if needed
State the hypothesis
Design the experiment
Construct the experiment
Discuss, draw conclusions
Formulate a context
Formulate a theoretical model
Perform the experiment
Perform data analysis
Test your apparatus
Perform preliminary experiments

Statistics and pills

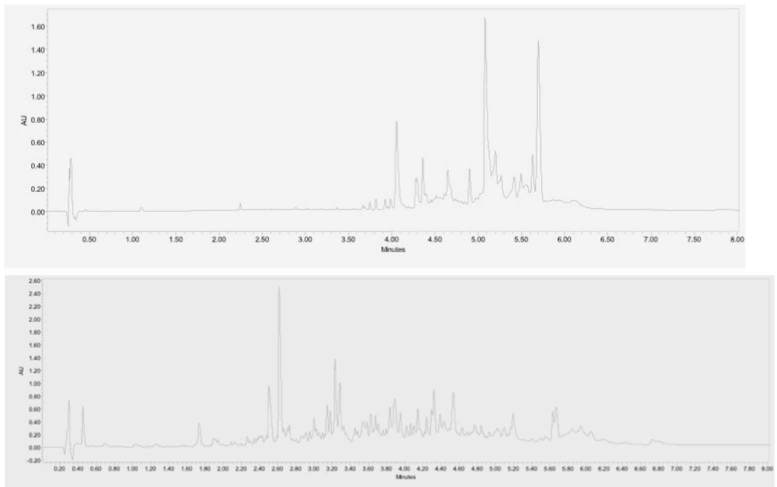


(more) words of caution

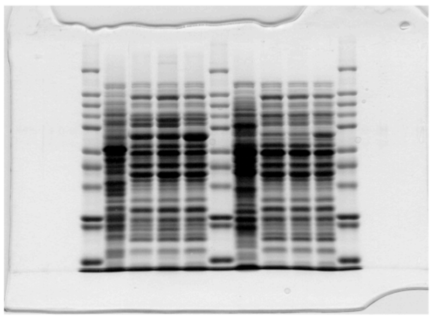
- There is no cure for a badly planned experiment
- Statistical design and analysis should be incorporated in the experiment at the planning stage:
  - be careful with the set-up of blocks replicates etc.,
  - use a design appropriate for the scope of your analysis
- There is no cure for bad data
  - be careful with sampling
  - be (extra) careful when you perform your analyses
  - never (fully) trust your instruments

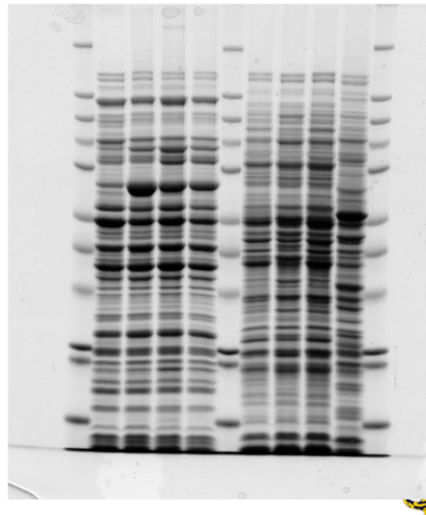
10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

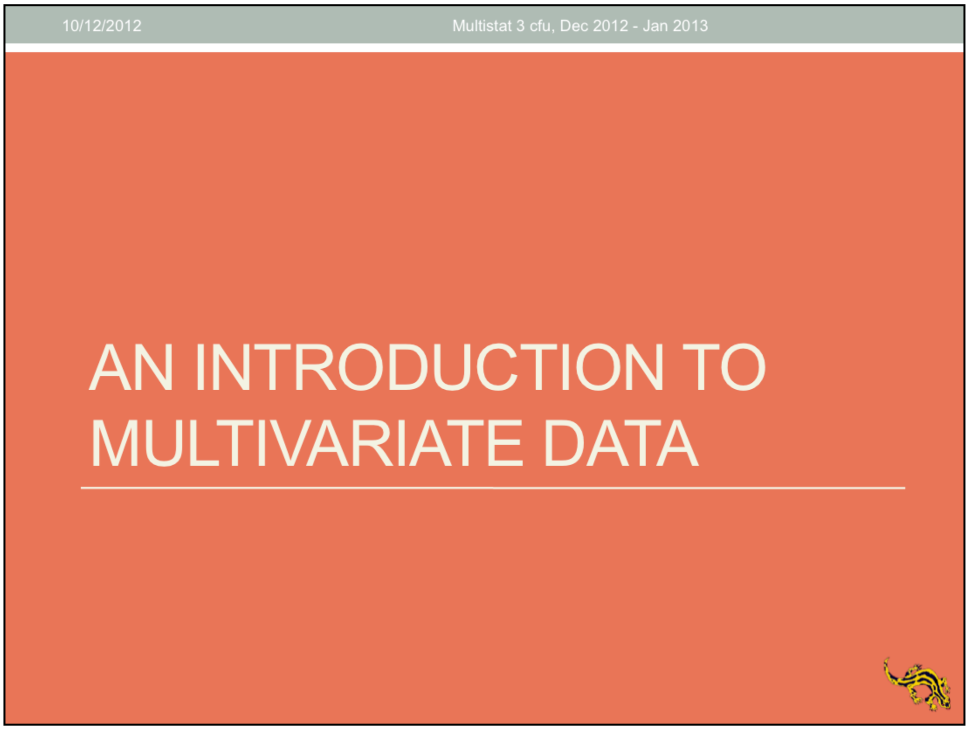
Two chromatograms



Two SDS-PAGE electropherograms






10/12/2012
Multistat 3 cfu, Dec 2012 - Jan 2013
AN INTRODUCTION TO MULTIVARIATE DATA

An introduction to multivariate data

- Types of variables
- Multivariate data: matrices and vectors
- Operations on matrices and vectors and their geometrical interpretation
- Transformations and standardization
- Missing values

Types of variables (from Steel and Torrie)

- **Qualitative (categorical)**: numerical measurement is not possible (although numerical codes can be used); categories are mutually exclusive, ordering is not meaningful (eye color, sex, marital status). Observations are classified and then enumerated
- **Quantitative**: measurement is possible because a natural order or ranking exists
  - **continuous variables**: all values in a range are possible (weight, temperature, concentration)
  - **discrete variables**: only some values are possible because of gaps in the scale of measurement (microbial counts)

Make examples of types of variables
categorical unordered: eye color, sex, marital status

Categorical with ordering: level of instruction, weight class

Quantitative continuous: weight, temperature (can we really measure on a continuous scale?)

Quantitative discrete: microbial counts

Types of variables (from Everitt and Dunn)

- **Qualitative**
  - **Nominal**: unordered categorical variables (eye colour)
  - **Ordinal**: a categorical variable for which there is ordering but no implication of distance (preference on an hedonic scale)
- **Quantitative**
  - **Interval**: equal differences between the points of the scale but position of 0 is arbitrary (°C, °F, pH)
  - **Ratio:** equal differences between the points of the scale and the position of 0 is not arbitraty (°K, mol H^+)

Ordinal variables are frequent in sensory science and psychometics: people may not use a linear scale and distance between points on a scale (i.e. an hedonic scale 0-9) might not be constant

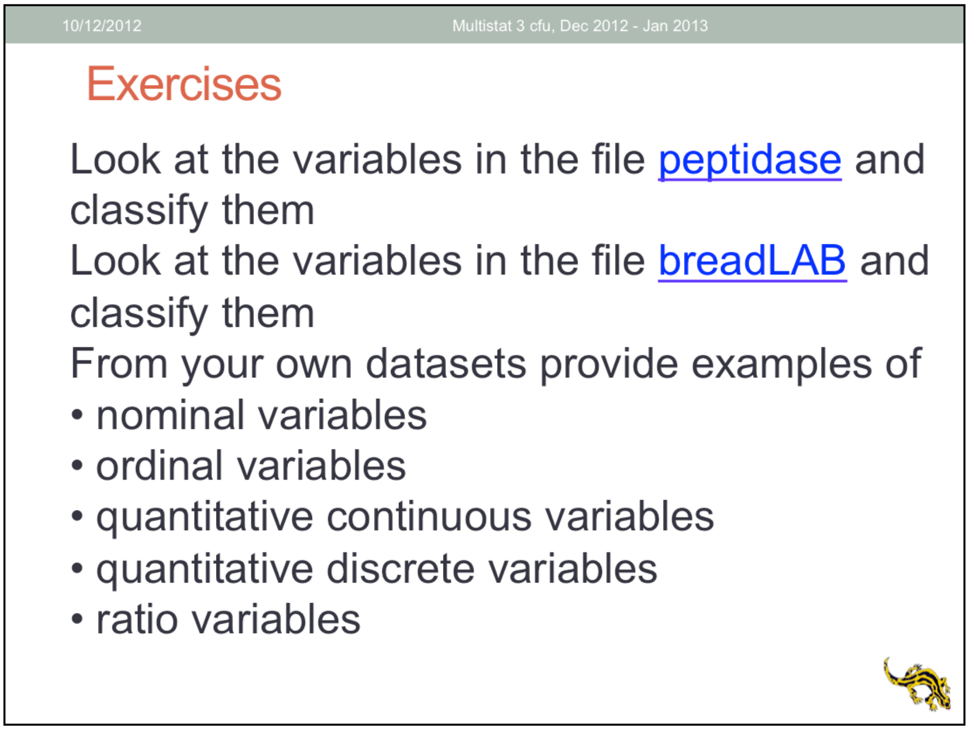
The difference between interval and ratio variables is important: although they both use scales with constant intervals, the fact that the 0 in interval scales is arbitrary has implications in comparing values. For example, temperature in °C is an interval scale with an arbitrary 0, while temperature in °K is a ratio scale. A body at 40°C is not twice as hot as a body at 20°C as can be easily seen by converting in °K.

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

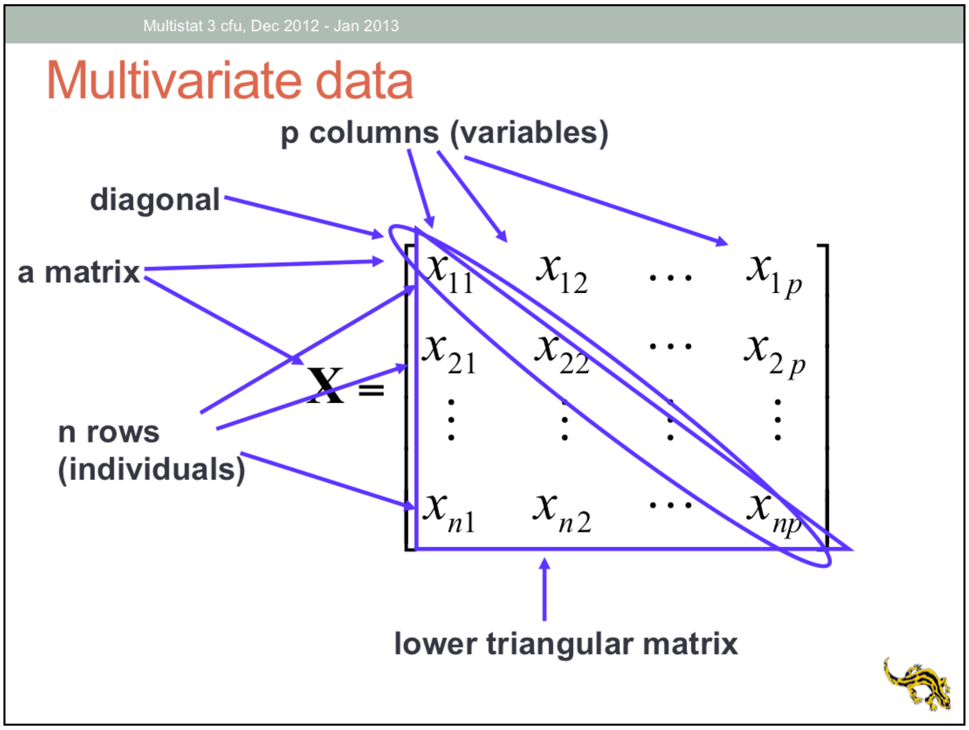
What you can do (an cannot do)

OK to compute	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Y	Y	Y	Y
Median and percentiles	N	Y	Y	Y
Add or subtract	N	N	Y	Y
Mean, standard deviation, standard error of the mean	N	N	Y	Y
Ratio or coefficient of variation	N	N	N	Y

From http://www.graphpad.com



Link Systat or Excel data files



A data matrix may often include both categorical and quantitative variables, although most often only the latter are used

Trace of a matrix is the sum of its diagonal elements. Addressing of the elements in the matrix is relevant, since a matrix is an ordered table of elements. Usually the index i is used for rows and j for columns.

A symmetric matrix is a square matrix which is equal to its transpose: distance and similarity matrices are usually symmetric

Vectors

$$\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$$

open file simpledata

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

A row vector specifies coordinates in an Euclidean space witn p dimensions, i.e. specifies the value of all variables for a single observation. x' is a traposed vector (usually the notation is used to indicate row vectors anyway). A matrix is actually made by one or more row vectors (or column vectors). Therefore transposition of vectors or matrices is simply exchanging rows for columns (first row becomes first column and so on and a nxp matrix becomes a pxn matrix)

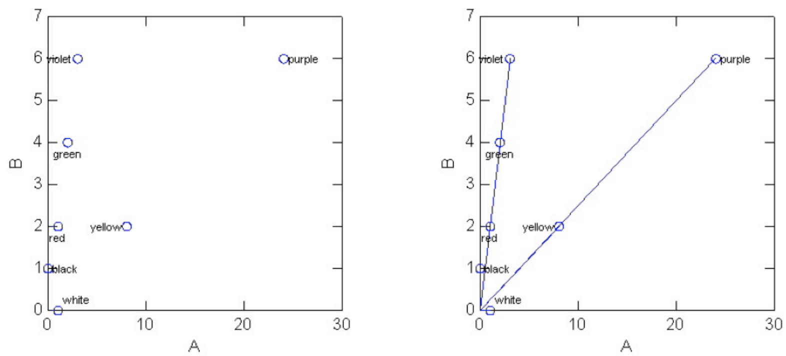
The rank of a matrix is the number of the row vectors which are linearly independent. A matrix is full rank if all the row vectors are linearly independent

A simple dataset

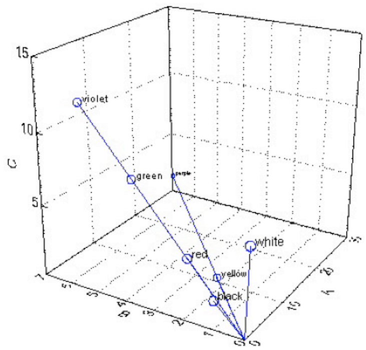
A	B	C	LABEL$
0	1	2	black
1	2	4	red
1	0	6	white
2	4	8	green
8	2	1	yellow
24	6	3	purple
3	6	12	violet

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

A simple data set



A simple data set



Special matrices

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

AI=A, AA^{-1}=I

$$\mathbf{E} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Special matrices:

I (identity) all 0 except the diagonal, (1). It plays the same role as the number 1 in multiplication in arithmetics

0 (null) all elements are 0,

E, all elements are 1.

A square matrix which is equal to its transposed matrix is symmetrical.

Diagonal matrix, all elements are 0 except the diagonal.

The inverse of a matrix is the matrix which, when multiplied for the original matrix, gives the matrix I. Multiplication by the inverse of a matrix is the equivalent of division

Sum of matrices

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

$$c_{ij} = a_{ij} + b_{ij}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix} + \begin{bmatrix} 3 & 4 \\ 7 & -1 \end{bmatrix} = \begin{bmatrix} 5 & 7 \\ 8 & 3 \end{bmatrix}$$

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

Sum and difference of matrices can be carried out only if the two matries have the same dimensions and the results is the ordered sum (difference) of elements

The following properties hold:

Commutative A+B=B+A

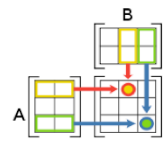
Associative A+(B+C)=(A+B)+C

A-(B-C)=(A-B)+C

Matrix multiplication

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2\times 2 + 3\times 1 + 1\times 0 & 2\times 1 + 3\times -1 + 1\times 1 \\ 1\times 2 + 1\times 1 + 0\times 0 & 1\times 1 + 1\times -1 + 0\times 1 \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 0 \\ 3 & 0 \end{bmatrix}$$



$$c_{ij} = \mathbf{a'}_{(i)} \mathbf{b}_j$$

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \qquad \mathbf{A}(\mathbf{BC}) = \mathbf{AB}(\mathbf{C})$$

Scalar multiplication: all elements are multiplied by a scalar. Matix multiplication is possible only if the number of columns n of the first matrix is equal to the number of rows of the second matrix. Commutative property does not apply to matrix multiplication, whiel associative (ABC=(AB) C) and distributive (A(B+C)=AB+AC) properties apply. If two matrices have dimensions (m,n) and (n,p) the result of their multiplication has dimensions (m,p). The product of any matrix by the identity matrix (I) is the matrix itself. The internal product of two vectors is a scalar x'x which is the sum of squares of the x elements of the vector and also is the square of the distance of the point represented by the vector from the origin. While if holds (AB)'=B'A' in general BA=AB does not hold

The image is from: http://en.wikipedia.org/wiki/Matrix_(mathematics)

Matrix multiplication

$$\mathbf{C} = \mathbf{D}(d_i)\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 & 7 \\ -1 & 3 & -2 \\ 4 & 3 & -5 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 14 \\ -3 & 9 & -6 \\ 16 & 12 & -20 \end{bmatrix}$$

$$\mathbf{AI} = \mathbf{A}$$

$$\mathbf{AA^{-1}=A^{-1}A=I}$$

Premultiplication of a matrix by a diagonal matrix gives a second matrix in which each row of the second matrix is multiplied by the corresponding diagonal element of the first matrix. if you premultiply by a diagonal matrix, the rows of the matrix are multiplied by the element of the diagonal; if you postmultiply, the columns are multiplied

The inverse B of a square matrix A is a matrix such as:

AB=BA=I

A matrix is singular or degenerate if it cannot be inverted; this happens if and only if the determinant is 0.

A matrix is invertible only if it is full rank.

Vectors: scalar product

$$\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$$

$$\mathbf{cx}' = \begin{bmatrix} cx_1 & cx_2 & \cdots & cx_p \end{bmatrix}$$

The scale of the vector changes by multiplication by the scalar c. Scalar multiplication of a matrix works exactly in the same way. From an Euclidean point of view, multiplication of a vector by a scalar changes its length by the factor c.

Vector multiplication

$$\mathbf{xx'} = \begin{bmatrix} 2 \\ 5 \\ 6 \\ 3 \end{bmatrix} \begin{bmatrix} 2 & 5 & 6 & 3 \end{bmatrix} = \begin{bmatrix} 2\times 2 & 2\times 5 & 2\times 6 & 2\times 3 \\ 5\times 2 & 5\times 5 & 5\times 6 & 5\times 3 \\ 6\times 2 & 6\times 5 & 6\times 6 & 6\times 3 \\ 3\times 2 & 3\times 5 & 3\times 6 & 3\times 3 \end{bmatrix}$$

The matrix obtained by this product is the SSCP (sum of squares and cross products) matrix, a simmetrical matrix.

Vectors: internal product

$$\mathbf{x}'\,\mathbf{x} = \begin{bmatrix} 2 & 5 & 6 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 6 \\ 3 \end{bmatrix} = 4 + 25 + 36 + 9 = 74$$

It is the sum of squares of the vector. By convention vectors are written as row vectors. Column vectors are obtained by transposing row vectors

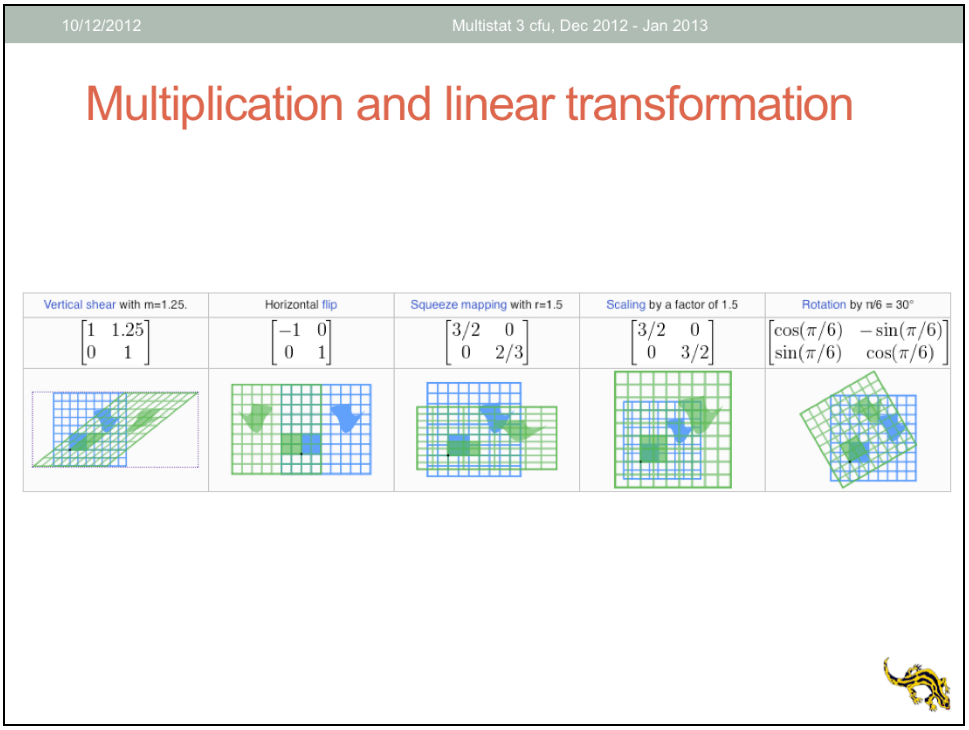


Image from http://en.wikipedia.org/wiki/Matrix_(mathematics).

Multiplication of a mxn matrix by a nxn matrix can be used to obtain linear transformations of the first matrix. Examples are provided for Euclidean space in two dimensions. Check the article on linear transformations on wikipedia for further details.

Euclidean distance and internal product

$$\mathrm{d} = \sum_{i}^{p}\left[(x_i - y_i)^2\right]^{\frac{1}{2}}$$

$$\cos\theta = \frac{\mathbf{x'y}}{(\mathbf{x'x})^{1/2}(\mathbf{y'y})^{1/2}}$$

This slide shows the difinition of Euclidean distance between two points with coordinates x and y (x and y are vectors). θ è is the angle between the two vectors; In cos θ the vectors are divided by their length which is equivalent to normalizing them to unit length. If the two vectors are coincident then cos θ =1; if they are at right angle cos θ =0. When **y=0, d is also called the L2-norm (norm, for short) of the vector and is its length**

Matrix and vector multiplication

$$\mathbf{A}_{\mathbf{mn}}\mathbf{x}_n = \mathbf{y}_m = \begin{bmatrix} 0.6 & 0.2 \\ 0.3 & 1.6 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$$

Postmultiplication of a mxn matrix by a vector with dimension n linearly transforms the group of variables represented by the vector in a new space with n dimensions

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

Eigenvalues and eigenvectors

The eigenvalues of a matrix **A** with dimension *pxp* are the solutions of the equation

$$\left|\mathbf{A} - \lambda \mathbf{I}\right| = 0$$

a. the product of the eigenvalues of **A** is equal to the determinant of **A**
b. the sum of the eigenvalues of **A** is equal to trace(**A**)

The determinant of the equation is a polynomial of grade p.

From: http://en.wikipedia.org/wiki/Matrix_(mathematics)

A number λ and a non-zero vector v satisfying

$$Av = \lambda v$$

are called eigenvalue and eigenvector of A, respectively. The number λ is an eigenvalue of A if and only if $A-\lambda I$ is not invertible, which is equivalent to

$$\det(A-\lambda I) = 0$$

The function $p_A(t) = \det(A-tI)$ is called the characteristic polynomial of A, the degree of this polynomial is n. Therefore $p_A(t)$ has at most n (possibly complex) different roots, i.e. eigenvalues of the matrix.

The trace of a square matrix is the sum of its diagonal entries. It equals the sum of its n eigenvalues.

Eigenvalues and eigenvectors

Each of the eigenvalues λ_i is associated to a vector $\mathbf{x}_i$ whose elements satisfy the system of equations:

$$|\mathbf{A} - \lambda\mathbf{I}|\mathbf{x}_i = 0$$

the eigenvectors $\mathbf{x}_i$ and $\mathbf{x}_j$ associated to the eigenvalues λ_i and λ_j of a simmetrical matrix are orthogonal

Transformation and standardization

Transformations/standardizations may be necessary:

- to make scales of measurements of variables comparable
- to correct for departures from assumptions related to hypothesis testing:
  - indipendence of mean and variance
  - experimental errors indipendently and normally distributed with a common variance
  - other (improving correlation)

More on transformations

Method	Transformation(s)	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	Dependent variable = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = log(x)	$y = b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = log(y) Independent variable = log(x)	$\log(y) = b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

From stat trek web site

Standardization

- ***z* scores**: for each variable the sample mean is subtracted and the result is divided by the standard deviation. The values are standardized with 0 mean and unit variance.
- **range**: the minimum value of the sample is subtracted and the results are divided by the range (max-min). The values are standardized between 0 and 1.

Transformations

- **square root**: useful when small whole numbers (following a Poisson distribution) or percentages in the range of 0-20 or 80-100 (use sqr(100-x)) are involved. When very small numbers are involved use sqr(Y+1/2).
- **log**: equalizes variances when mean and standard deviation are proportional; it also makes multiplicative effects additive. Useful with positive integers covering a wide range (for example microbial counts). An alternative is log(Y+1) (if 0 is possible)
- **angular transformation**: arcsin(sqr(Y)), useful with binomial data expressed as percentages or decimal fractions

Basic multivariate statistics: mean and variance

$$\mu' = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_p \end{bmatrix}$$ **where $\mu_i = E(x_i)$**

$$\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix}$$ **is an estimate of μ'**

$$\boldsymbol{\sigma}' = \begin{bmatrix} \sigma_1^2 & \sigma_2^2 & \cdots & \sigma_p^2 \end{bmatrix}$$ **where $s^2_i = E(x_i - \mu_i)$**

$$\mathbf{s}' = \begin{bmatrix} s_1^2 & s_2^2 & \cdots & s_p^2 \end{bmatrix}$$ **is an estimate of σ'**

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

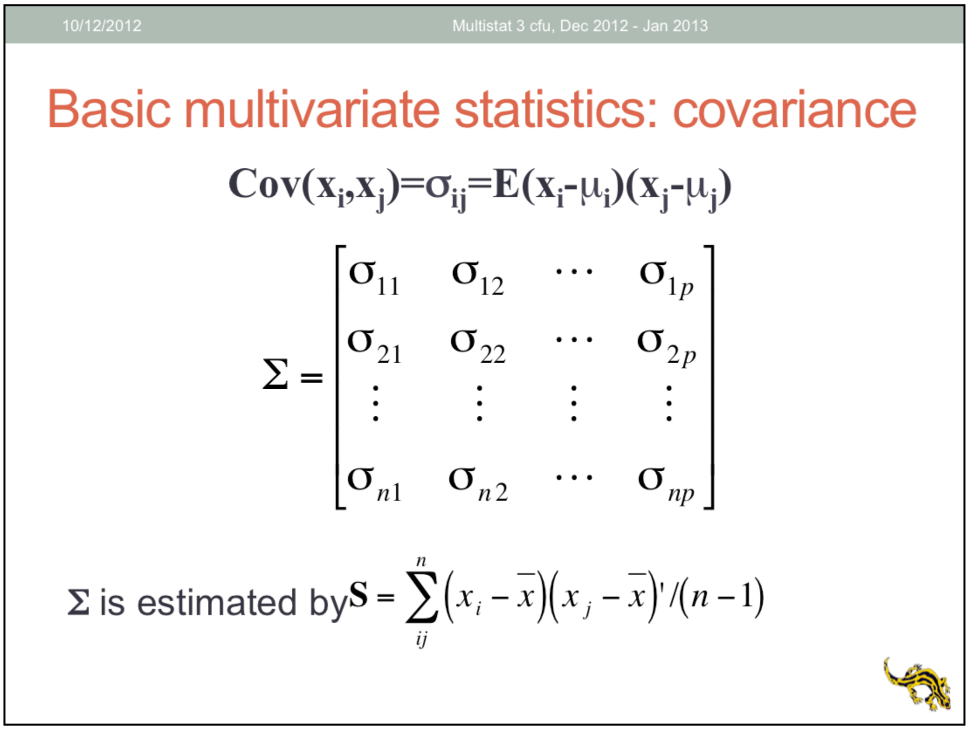
Example

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

n=2,p=2
$\mathbf{x}_1'$ =[1,2] with mean 1.5
$\mathbf{x}_2'$ =[3,4] with mean 3.5
$\mathbf{x}_1'\mathbf{x}_1$=1x1+2x2=5 (a scalar)

$$\mathbf{x}_1\mathbf{x}_1' = \begin{bmatrix} 1x1 & 1x2 \\ 2x1 & 2x2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

a simmetrical *pxp* SSCP matrix



if i=j the covariance is the variance; with p variables a symmetrical pxp matrix is obtained and its diagonal are the variances (p variances) while there are p(p-1)/2 covariances. The matrix can also be called variance-covariance matrix. Covariance is highly sensitive to scale and therefore difficult to interpret

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

Basic multivariate statistics: correlation

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

with p variables there are p(p-1)/2 correlations, arranged in a simmetrical *pxp* matrix $\mathbf{R}=\mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$, where:

$$\mathbf{D}^{-\frac{1}{2}} = diag\left(1/\sqrt{\sigma_{ii}}\right)$$

correlation is covariance standardized by the standard deviation of the variables. It varies between –1 and 1 and is a measure of linear relationship of the variables. With p variables there are p(p-1)/2 corelations arranged in a pxp matrix, with diagonal elements=1. Usually only the lower triangular matrix is used.

10/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

Missing values

- Missing values can seriously impair the possibility of carrying out a multivariate analysis because of difficulties in calculating statistics
- Missing value analysis can help in finding patterns (or lack thereof) in missing data
- A number of methods can be used to use matrices including missing data:
  - Imputation (mean, median, etc.)
  - Deletion (pairwise, listwise)
  - Estimation

If a limited number of missing values is present it is oK to perform the analysis or to carry out imputation; if the number of missing values is high and especially if it is not at random the analysis should be used only to obtain preliminary information.