

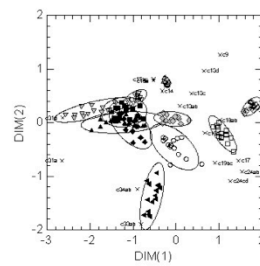
# MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND NUTRITION: AN INTRODUCTION

## 2. EXPLORING MULTIVARIATE DATA

Prof. Eugenio Parente

Scuola di Scienze Agrarie- Università della Basilicata

---



## Outline

- A few words on stat software
- Presenting multivariate data as tables.
- Scatterplots and enhanced scatterplots.
- Scatterplot matrices.
- Probability plots
- Comparing distributions: quantile-quantile plots



## Software packages

- **R** (<http://www.r-project.org/>) R is a powerful programming environment that is available as free software for a variety of platforms. Although it has a very steep learning curve and may be difficult to use for non-specialized users it offers the most comprehensive selection of graphical and statistical routines and it is continuously improved by a large scientific community. Available for Windows, MacOS, Linux/Unix
- **SAS** (<http://www.sas.com/technologies/analytics/statistics/>; available for Windows, Linux/Unix), **STATISTICA** (<http://www.statsoft.com/>; available for Windows only) and **SPSS** (<http://www.spss.com/statistics/>; available for Windows, MacOS, Linux/Unix) offer extensive data manipulation, statistical analysis, and graphing procedures; a variety of software modules for specialized applications are available.



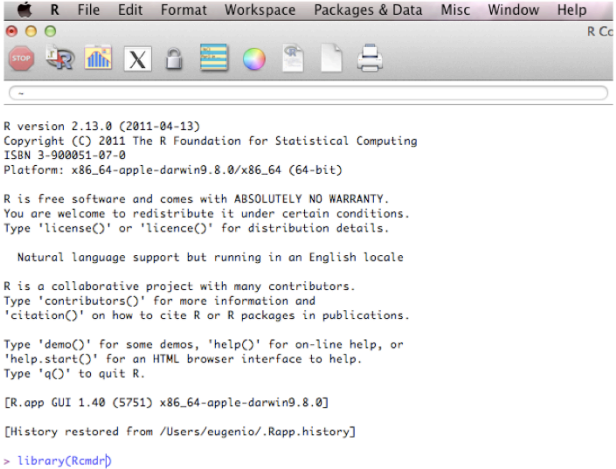
## Software packages

- **Systat** (<http://board.systat.com/Default.aspx>) is a generalist program with excellent graphing facilities and a large selection of univariate and multivariate statistical tools. Systat Inc. offers Mypstat, a (much) simplified version of Systat as free software for the use of students in academic environments. Runs only under Windows environments.
- **The Unscrambler** (<http://www.camo.com/rt/Products/Unscrambler/unscrambler.html>) is a specialized software which offers a large selection of multivariate statistical techniques and design of experiments. Runs only under Windows environments.
- **Neurosolutions** (<http://www.neurosolutions.com/>) offers a variety of packages for analyzing statistical problems by using Artificial Neural Networks, working under different environments.



11/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

## R with Rcmdr



The screenshot shows the Rcmdr application window. The title bar reads 'R Rcmdr'. The menu bar includes 'R', 'File', 'Edit', 'Format', 'Workspace', 'Packages & Data', 'Misc', 'Window', and 'Help'. The toolbar contains icons for a stop button, a search icon, a bar chart, a close button, a lock, a window icon, a globe, a document, and a printer. The main content area displays the R console output:

```
R version 2.13.0 (2011-04-13)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)


R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.40 (5751) x86_64-apple-darwin9.8.0]
[History restored from /Users/eugenio/.Rapp.history]
> library(Rcmdr)
```



11/12/2012 Multistat 3 cfu, Dec 2012 - Jan 2013

# R with Rcmdr

The screenshot displays the R Commander application window. The title bar reads "R Console" and "R Commander". The menu bar includes "File", "Edit", "Data", "Statistics", "Graphs", "Models", "Distributions", "Tools", and "Help". Below the menu bar, there are buttons for "Data set: <No active dataset>", "Edit data set", "View data set", and "Model: <No active model>".

The main window is divided into three panes:

- Script Window:** Contains the R script being executed. The visible text includes:
 

```

      R is free software and comes with ABSOLUTELY NO WARRANTY.
      You are welcome to redistribute it under certain conditions.
      Type 'license()' or 'licence()' for distribution details.

      Natural language support but running in an English locale.

      R is a collaborative project with many contributors.
      Type 'contributors()' for more information and
      'citation()' on how to cite R or R packages in publications.

      Type 'demo()' for some demos, 'help()' for on-line help,
      'help.start()' for an HTML browser interface to help,
      and 'q()' to quit R.

      [R.app GUI 1.40 (5751) x86_64-apple-darwin9.8.0]

      [History restored from /Users/eugenio/.Rapp.history]

      > library(Rcmdr)
      Loading required package: tcltk
      Loading Tcl/Tk interface ... done
      Loading required package: car
      Loading required package: MASS
      Loading required package: mmet
      Loading required package: survival
      Loading required package: splines

      Rcmdr Version 1.7-0

      Attaching package: 'Rcmdr'

      The following object(s) are masked from 'package:tcltk':

        tclvalue

      Warning messages:
      1: package 'Rcmdr' was built under R version 2.13.1
      2: package 'car' was built under R version 2.13.1
      3: package 'survival' was built under R version 2.13.2
      > |
      
```
- Output Window:** Currently empty.
- Messages:** Shows a note: "[1] NOTE: R Commander Version 1.7-0: Mon Dec 10 20:14:42 2012".

## Multivariate data: is a summary table enough?

Table 2  
Technological properties of lactic acid bacteria isolated from pasta filata cheese and from natural starter cultures

Group	pH decrease				Proteolysis (mg Leu L <sup>-1</sup> ) Autolysis <sup>a</sup>						Peptidase activity, cell free extracts								
	37°C		30°C		37°C		30°C		Rate (OD min <sup>-1</sup> )		Extent (-)		μKatalmg <sup>-1</sup>						
	6h	24h	6h	24h	1d	7d	1d	7d	NaP pH 7	NaCl pH 5.5	NaCl pH 5.1	NaP pH 7	NaCl pH 5.5	NaCl pH 5.1	PeppA	PeppX	PeppNC	PeppI	
<b>Enterococci</b>																			
Min.	0.19	0.52	0.27	0.48	28.0	26.3	24.0	26.2	0.03	0.02	0.03	0.05	0.06	0.06	0.07	0.00	0.00	0.04	
Max.	1.38	2.10	1.22	1.96	45.3	48.4	46.4	45.3	0.12	0.14	0.10	0.23	0.27	0.20	0.16	1.16	0.20	0.11	
Median	0.96	1.84	0.84	1.56	35.7	33.2	31.8	33.0	0.04	0.04	0.05	0.09	0.09	0.09	0.07	0.00	0.11	0.09	
<b>Lactobacillus fermentum</b>																			
Min.	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	
Max.	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	
Median	0.94	1.00	0.69	0.94	12.4	12.6	15.7	12.6	0.00	0.00	0.00	0.00	0.00	0.00	0.38	1.94	0.97	0.40	
<b>Lactobacillus helveticus</b>																			
Min.	0.28	2.37	0.30	0.48	35.7	43.5	31.0	41.1	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.31	0.03	0.01	
Max.	0.99	3.34	0.67	2.55	46.6	47.4	47.6	47.5	0.08	0.07	0.13	0.22	0.21	0.18	0.11	1.46	3.54	0.45	
Median	0.70	3.14	0.49	2.29	44.5	45.1	45.0	46.5	0.02	0.03	0.02	0.06	0.11	0.09	0.03	0.85	1.57	0.06	
<b>Lactococcus lactis</b>																			
Min.	0.35	0.61	0.58	0.82	22.3	24.5	16.2	15.0	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.11	0.06	0.00	
Max.	2.12	2.41	1.91	2.30	43.8	46.2	46.7	47.4	0.22	0.14	0.17	0.57	0.34	0.34	0.19	7.86	2.09	0.16	
Median	1.85	2.31	1.60	2.06	33.9	37.5	30.7	35.2	0.04	0.06	0.03	0.08	0.13	0.08	0.09	1.41	0.50	0.04	
<b>NSLAB<sup>b</sup></b>																			
Min.	0.25	0.72	0.38	0.78	19.1	36.7	19.1	25.5	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.04	0.00	
Max.	1.52	3.05	0.85	2.54	46.9	49.7	47.3	48.7	0.31	0.20	0.19	0.64	0.49	0.42	0.15	2.94	8.40	0.27	
Median	0.71	2.64	0.59	1.82	37.7	46.7	38.3	44.7	0.06	0.05	0.05	0.24	0.17	0.15	0.02	0.24	2.91	0.08	
<b>Streptococcus thermophilus</b>																			
Min.	0.88	2.30	1.17	1.93	13.7	18.6	5.8	6.9	0.00	0.01	0.01	0.00	0.00	0.00	0.08	1.86	0.03	0.04	
Max.	1.46	2.47	1.31	2.13	34.6	32.9	42.5	45.9	0.05	0.03	0.09	0.07	0.07	0.01	0.10	2.25	0.87	0.16	
Median	1.03	2.45	1.22	2.01	15.5	20.5	8.9	8.3	0.02	0.01	0.01	0.06	0.00	0.00	0.09	2.06	0.45	0.10	

<sup>a</sup>Non-starter lactic acid bacteria (NSLAB) include *Lb. casei*, *Lb. paracasei*, *Lb. rhamnosus* and *Lb. casei*.

<sup>b</sup>NaP pH 7: 50 mmol L<sup>-1</sup> Na-phosphate, pH 7.0; NaCl pH 5.5: 0.2 mol L<sup>-1</sup> NaCl, pH 5.5; NaCl/NaLac pH 5.1: 88.5 mmol L<sup>-1</sup> Na-lactate, 0.7 mol L<sup>-1</sup> NaCl, pH 5.1.



Available online at [www.elsevier.com/locate/journal](http://www.elsevier.com/locate/journal)  
ScienceDirect  
International Dairy Journal 26 (2012) 10–15



Acid production, proteolysis, autolytic and inhibitory properties of lactic acid bacteria isolated from pasta filata cheeses: A multivariate screening study

Paola Pizzarello<sup>a</sup>, Teresa Zanzi<sup>a</sup>, Antonietta Ricciardi<sup>b</sup>,  
Paola L.H. McCreeshy<sup>c</sup>, Eugenio Pizzarello<sup>a\*</sup>

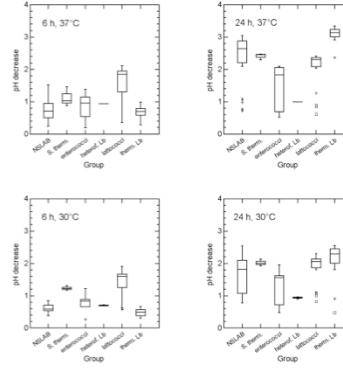
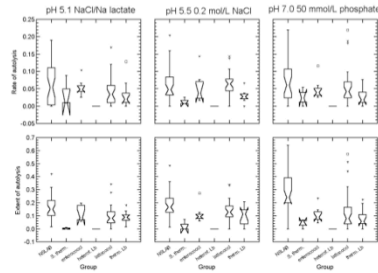
<sup>a</sup>Department of Food and Nutritional Sciences, University of Bari, Italy; <sup>b</sup>Department of Food and Nutritional Sciences, University of Bari, Italy; <sup>c</sup>Department of Food and Nutritional Sciences, University of Bari, Italy; \*Corresponding author. E-mail: [pizzarello@uniba.it](mailto:pizzarello@uniba.it)

Received 10 November 2011; accepted 10 May 2012



This set of data includes technologically relevant properties for lactic acid bacteria isolated from Pasta Filata cheeses

# Multivariate data: is a summary graph enough?

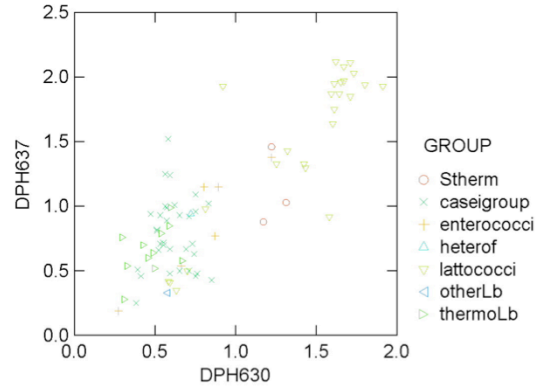


Acid production, proteolysis, autolytic and inhibitory properties of lactic acid bacteria isolated from pasta filata cheeses: A multivariate screening study  
Pablo Pinero<sup>1</sup>, Teresa Zurea<sup>2</sup>, Antonieta Ruizpérez<sup>3</sup>, Paul L.H. McSweeney<sup>2</sup>, Eugenio Ponnor<sup>2\*</sup>  
<sup>1</sup>Department of Food and Food Packaging Science, University of Burgos, Spain; <sup>2</sup>Department of Food and Food Packaging Science, University of Limerick, Ireland; <sup>3</sup>Department of Food and Food Packaging Science, University of Burgos, Spain





## Exploring data: the scatterplot



[open file technolab](#)

www.technolab.com  
ncDirect  
www.technolab.com

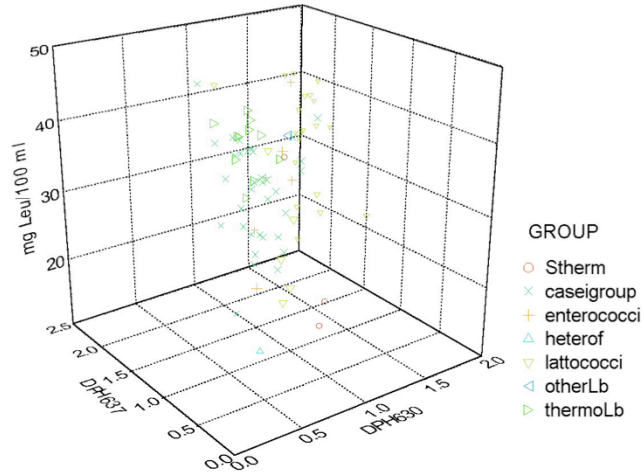
INTERNATIONAL  
DAIRY  
JOURNAL  
www.technolab.com

Acid production, proteolysis, autolytic and inhibitory properties of lactic acid bacteria isolated from pasta filata cheeses: A multivariate screening study  
Paolo Pizzoni<sup>1</sup>, Teresa Zenti<sup>2</sup>, Annamaria Ricciardi<sup>2</sup>, Paul L.H. McSwiney<sup>2</sup>, Eugenio Pizzetti<sup>2\*</sup>  
<sup>1</sup>Università di Napoli "Parthenope", Napoli, Italy; <sup>2</sup>Technolab, Dublin, Ireland; <sup>3</sup>Department of Food and Nutrition Science, University of Turin, Italy; <sup>4</sup>Department of Food and Nutrition Science, University of Turin, Italy; <sup>5</sup>Department of Food and Nutrition Science, University of Turin, Italy; <sup>6</sup>Department of Food and Nutrition Science, University of Turin, Italy



A simple 2D scatterplot may reveal grouping patterns (if you use different symbols for different group of objects), correlations, distribution, need for transformation

## Exploring data: the scatterplot

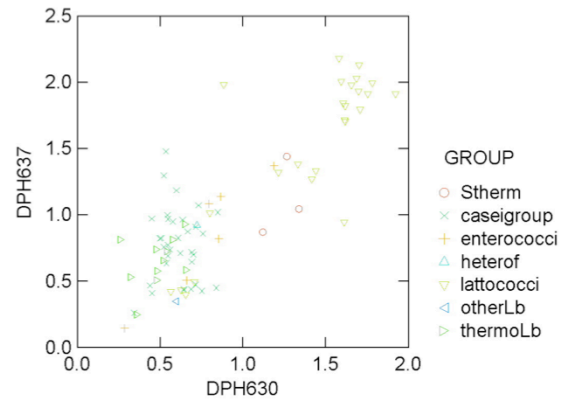


[open file technolab](#)



A 3D scatterplot may be more difficult to read (hard to interpret the position of the points because of perspective, usually points which are farther from the observer are plotted as smaller) but may be useful in detecting grouping and patterns

## Exploring data: jittered scatterplot

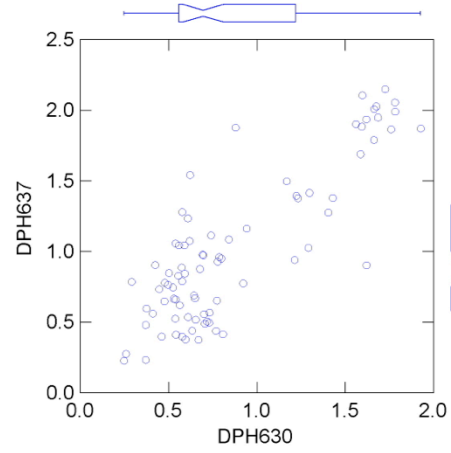


[open file technolab](#)

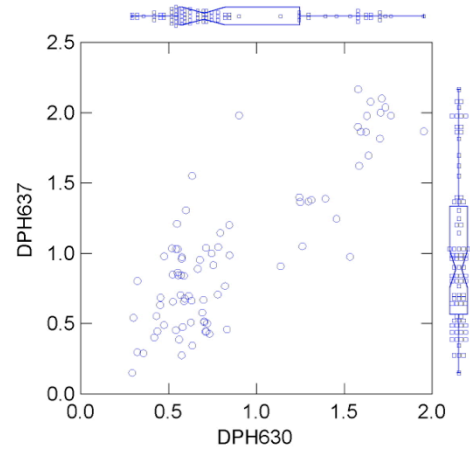


When several points overlap adding a small random jitter to points may help in visualization

## Exploring data: enhanced scatterplots with border displays (1)

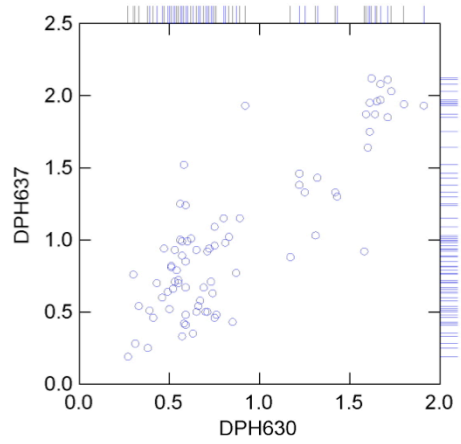


## Exploring data: enhanced scatterplots with border displays (2)



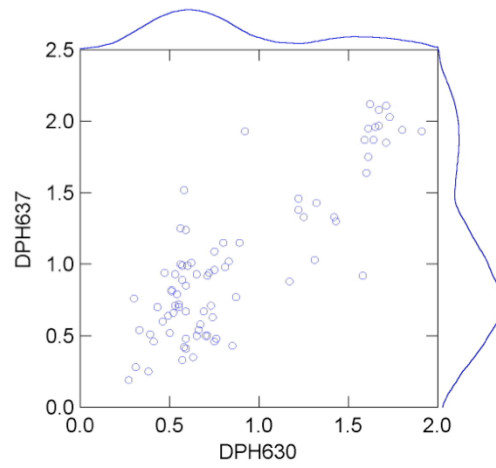
Scatterplots can be enhanced by adding border displays; here a symmetrical dot plot + box plot

## Exploring data: enhanced scatterplots with border displays (3)



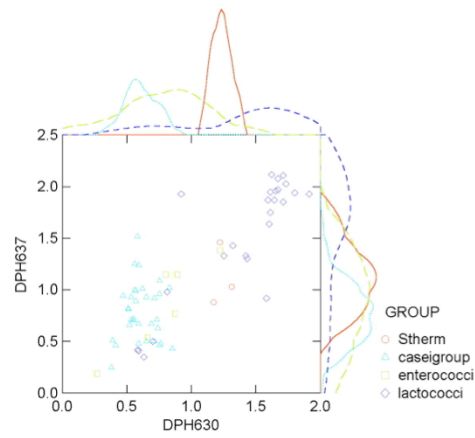
Stripe plot, useful for detecting density patterns

## Exploring data: enhanced scatterplots with border displays (4)



Multivariate kernel for distribution free density evaluation

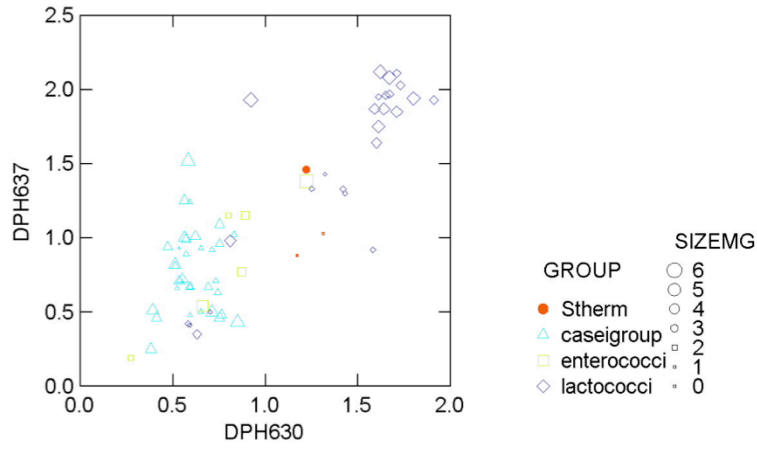
## Exploring data: enhanced scatterplots with border displays (5)



Here density kernels are by group

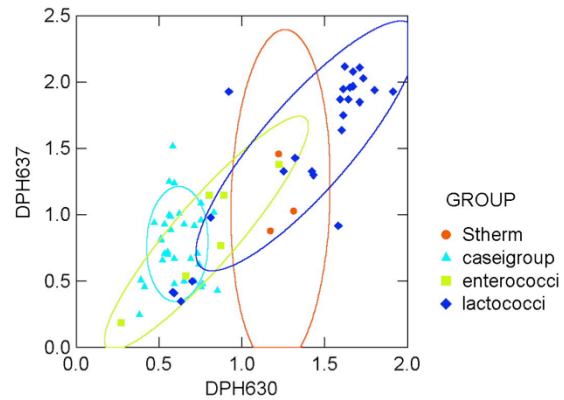


## Exploring data: bubble plots



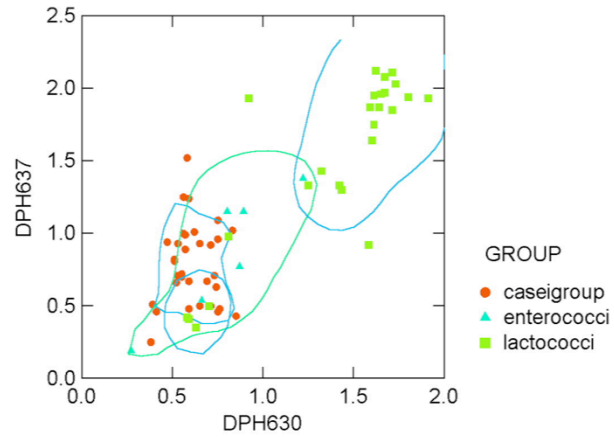
A third variable, free amino acids, may be added as size of the points; be careful; area is perceived in a different way than length; usually a square root transformation helps

## Exploring data: scatterplots with bivariate confidence ellipses



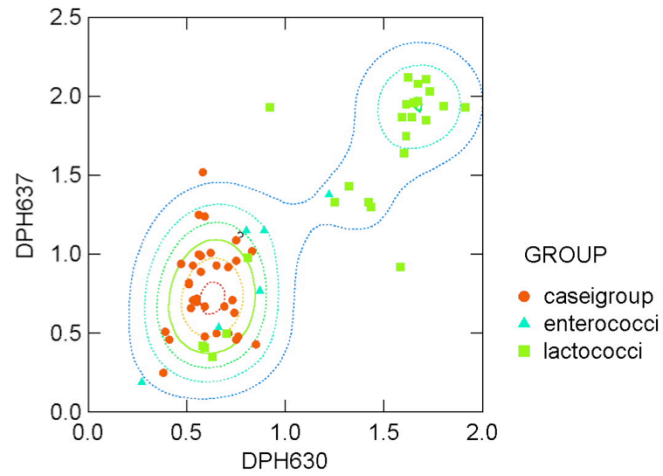
Bivariate ellipses can be drawn around group of points; they may either be confidence ellipses for the centroid (the coordinates of the mean of the two variables) or for the observations; be careful, they assume bivariate normal distribution; a transformation may help in making your distribution closer to normal. Orientation of the ellipse is either the covariance (sample confidence) or the correlation (centroid confidence)

## Exploring data: scatterplots with bivariate confidence kernels



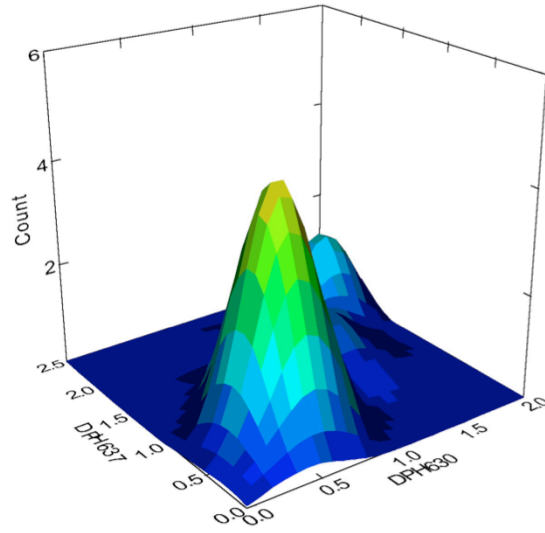
Confidence kernels work in the same way but do not assume a particular form for the distribution

## Exploring data: contour density plots

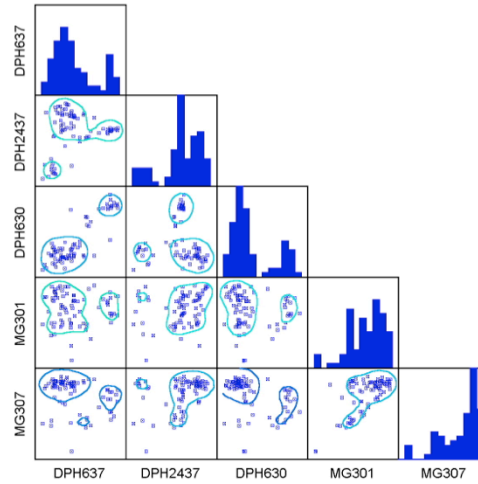


If you want to detect “natural group”s of observations a kernel density contour plot may help; this is actually the same as projecting on the x-y axis the cuts of the z axis at various heights

## Exploring data: 3-D density scatterplots

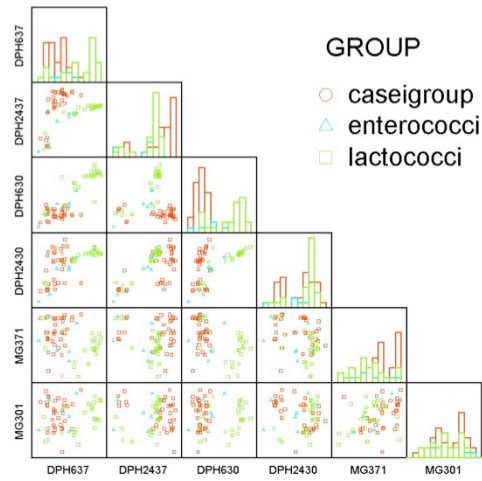


## Exploring data: SPLOM (scatterplot matrix) with density displays

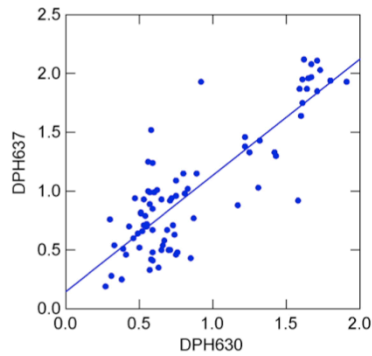


Scatterplot matrices are very useful for exploring the relationships between several variables; the usual format is a triangular matrix; but you can also select which variables should go on x and y

# Exploring data: SPLOM (scatterplot matrix) with density displays



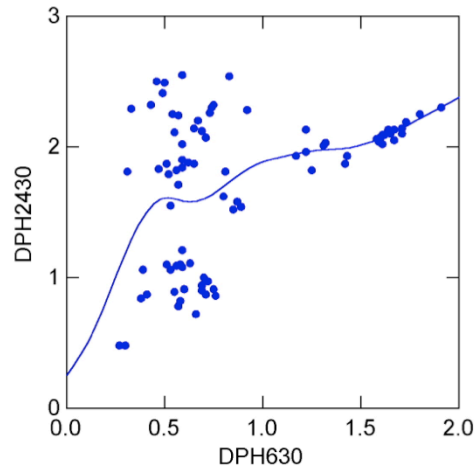
## Exploring data: scatterplots with smoothers (linear)



Smoothers help in seeing trends in the data but you should not abuse them; if you use a parametric smoother, like linear, log, power, etc, you actually fit a regression and, just by magic points get closer to the regression line

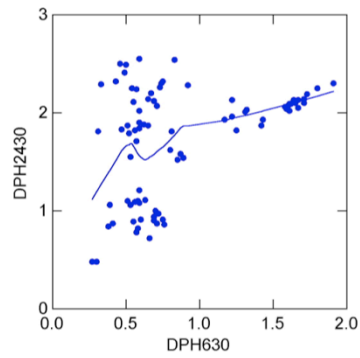


## Exploring data: scatterplots with smoothers (DWLS)



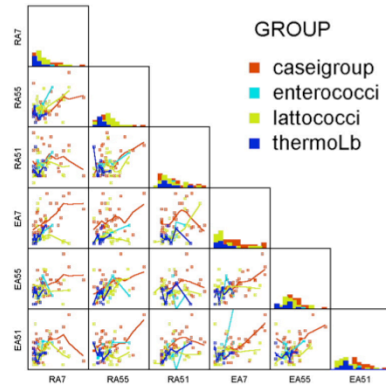
Distance Weighted least Square Smoothing is a non parametric smoother which fits a curve to the data; all the data influence the curve at any given point but their influence depends on their distance; a tension parameter can be used to adjust the curve (the lower the tension the more the curve is affected by closer rather than farther points)

## Exploring data: scatterplots with smoothers (LOWESS)



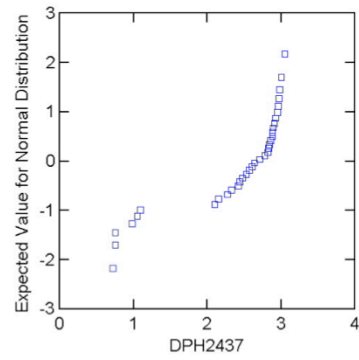
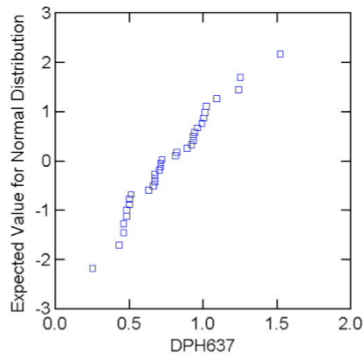
LOWESS Locally Weighted Scatterplot Smoothing

## Exploring data: SPLOM (scatterplot matrix) with smoothers



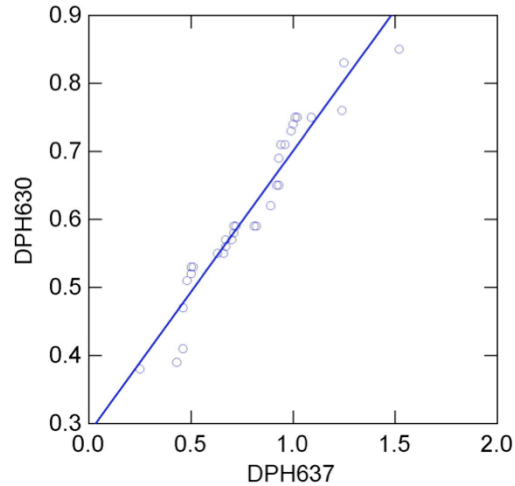
You can use smoothers in SPLOM but things may become a wee bit messy

## Exploring data: probability plots



probability plots can be used for single variables to test if their distribution fits a particular model (normal, binomial, Poisson, etc.) and evaluate the need for transformation or check if actually two populations are included in the data

## Exploring data: quantile-quantile plots



It is often interesting to see if two variables share the same distribution; a quantile-quantile plot compares the quantiles of the two distributions; if they are similar you get a straight line with points neatly arranged; probability plots can be used for single variables to test if their distribution fits a particular model (normal, binomial, Poisson, etc.) and evaluate the need for transformation or check if actually two populations are included in the data

## Some rights reserved

This presentation was created by Eugenio Parente, 2008, revised in 2012. With the exception of figures and tables taken from published articles the material included in this presentation is covered by Creative Commons Public License “by-nc-sa” (<http://creativecommons.org/licenses/by-nc-sa/2.5/deed.en>).

