Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Master of Science in Geospatial Technologies

Geostatistics Exploratory Analysis

Carlos Alberto Felgueiras
cfelgueiras@isegi.unl.pt

Geostatistics – Exploratory Data Analysis

Contents

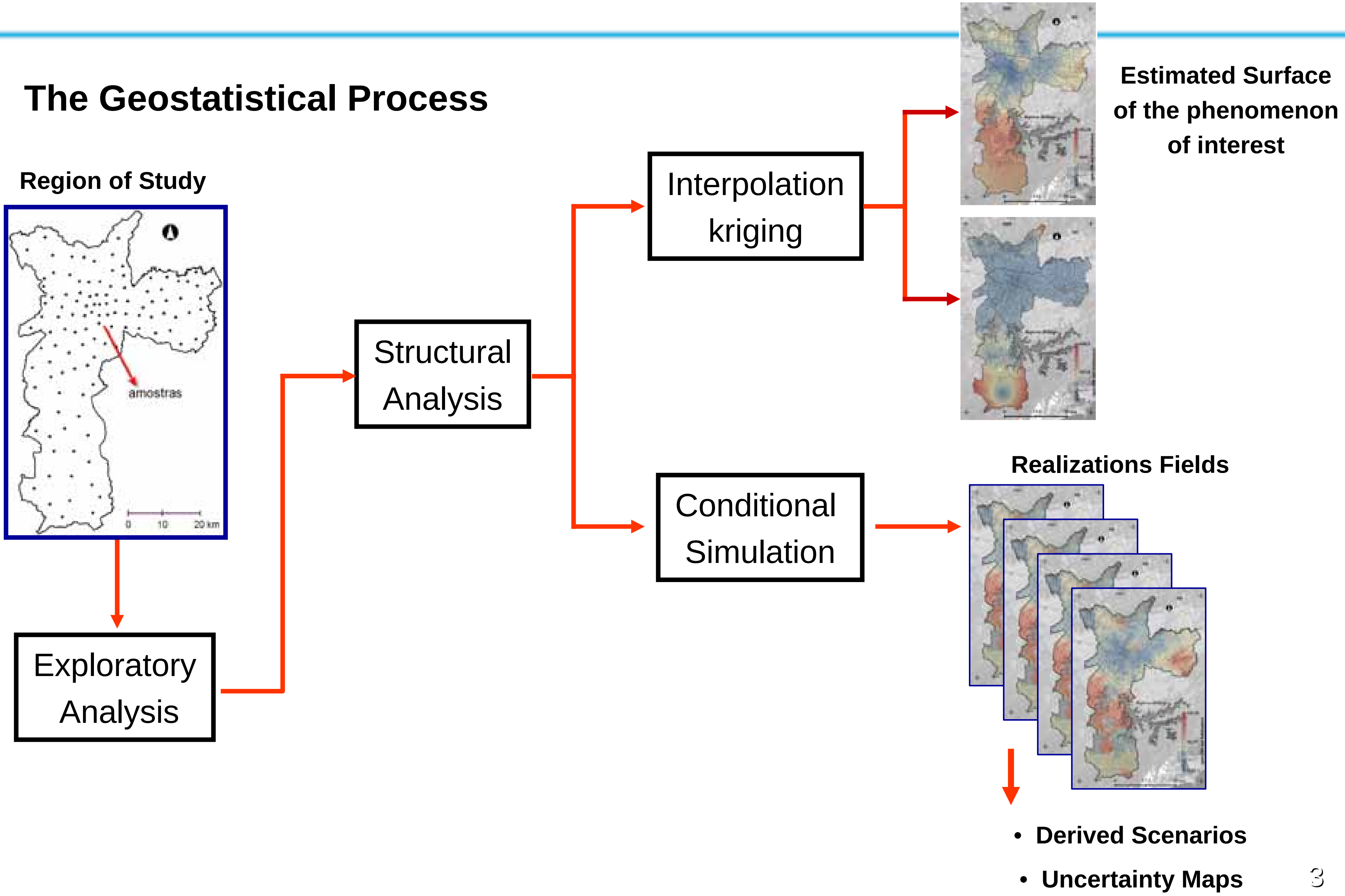
EDA and ESDA

Univariate Description

Bivariate Description

The Geostatistical Process – General View

The Geostatistical Process



Region of Study

Exploratory Analysis

Structural Analysis

Interpolation kriging

Conditional Simulation

Estimated Surface of the phenomenon of interest

Realizations Fields

- Derived Scenarios
- Uncertainty Maps

Geostatistics – Exploratory Data Analysis

http://www.ncgia.ucsb.edu/giscc/units/u128/u128_f.html

What is Exploratory Data Analysis (EDA)?

- Aim is to identify data properties for purposes of:
  - pattern detection in data (homogeneity, heterogeneity,....)
  - hypothesis formulation from data (symmetry, gaussian,.....)
  - some aspects of model assessment
    (e.g.goodness of fit, identifying data effects on model fit).

- Analysis are based on:
  - the use of *graphical and visual methods* and
  - the use of *numerical techniques* that are *statistically robust*, i.e., not much affected by extreme or atypical data values.

- Emphasis on *descriptive methods* rather than formal hypothesis testing.

- Importance of “staying close to the original data” in the sense of using simple, intuitive methods.

Geostatistics – Exploratory Spatial Data Analysis

http://www.ncgia.ucsb.edu/giscc/units/u128/u128_f.html

What is Exploratory Spatial Data Analysis (ESDA)?

- extension of EDA to detect **spatial** properties of data. Need additional techniques to those found in EDA for:
  - detecting spatial patterns in data
  - formulating hypotheses based on the geography of the data
  - assessing spatial models.

- important to be able to link numerical and graphical procedures with the map
- need to be able to answer the question: “where are those cases on the map?.

  With modern graphical interfaces this is often done by 'brushing' - for example cases are identified by brushing the relevant part of a boxplot, and the related regions are identified on the map.

- This lecture only covers the case of attribute data attached to irregular areal units because we want to focus on the range of analytical tools required for ESDA and their provision using GIS.

EDA/ESDA – Data Distribution in Space

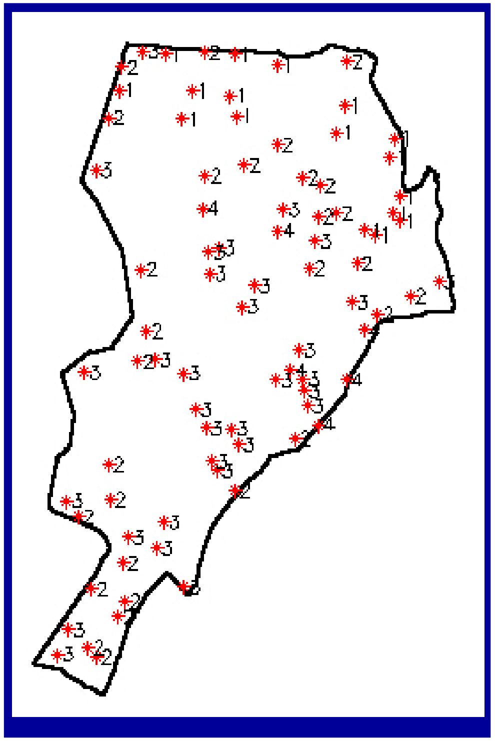
Sampling

Spatial sampling involves determining a limited number of locations in a geo-space for faithfully measuring phenomena that are subject to dependency and heterogeneity.

Dependency suggests that since one location can predict the value of another location, we do not need observations in both places.

Heterogeneity suggests that this relation can change across space, and therefore we cannot trust an observed degree of dependency beyond a region that may be small.

Basic spatial sampling schemes include random, clustered and systematic.



Each sample point α is represented by (x,y,z)

(x,y) is the 2-d space location

z is the attribute value

Exploratory Data Analysis - EDA

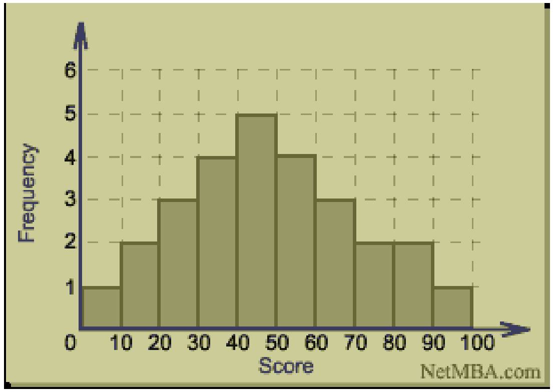
Univariate Description

- Frequency Tables and Histograms
- Cumulative Frequency Tables and Histograms
- PDFs and CDFs
- Normal Probability Plots
- Summary Statistics
    - Measures of location of the distributions
    - Measures of spread of the distributions
    - Measures of shape of the distributions
- Detection of rough components

EDA – Univariate Description

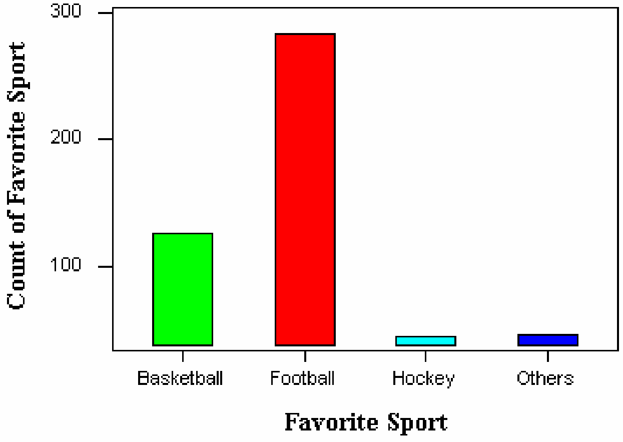
- ***Frequency Tables and Histograms***
  - Continuous Variables – Example of Exam Scores

Group	Count
0 - <10	1
10 - <20	2
20 - <30	3
30 - <40	4
40 - <50	5
50 - <60	4
60 - <70	3
70 - <80	2
80 - <90	2
90 – 100	1



  - Categorical Variables – Favorite Sports to watch

Group	Count
Basketball	123
Football	282
Hokey	15
Others	22



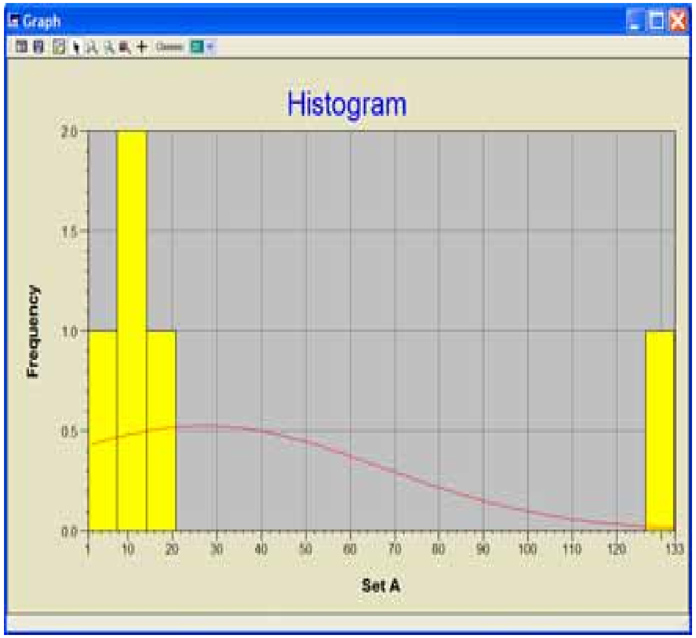
Important: The count (frequency) of each group devided by the total population give the **pdf** (**p**robability **d**istribution **f**unction) of the variable.

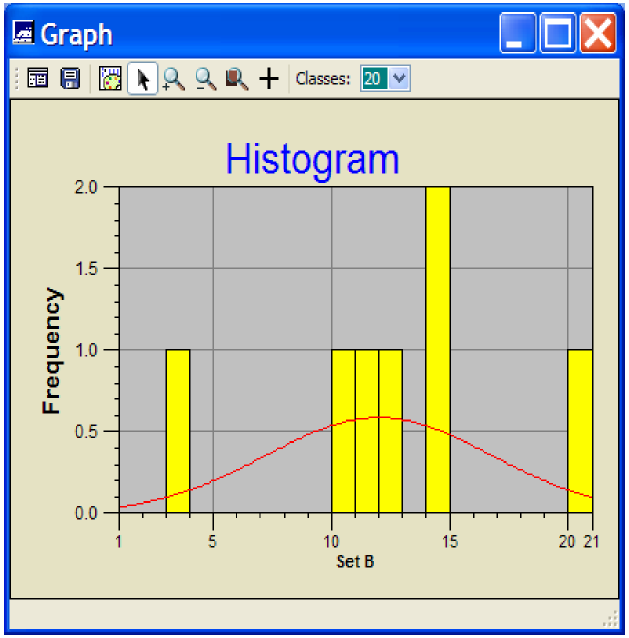
EDA – Univariate Description

- ***Frequency Tables and Histograms*** – a simple example

Ordered Sample Sets	
Set A	Set B
3	3
10	10
11	11
12	12
14	14
14	14
20	20
132	

Histograms of Sets A and B





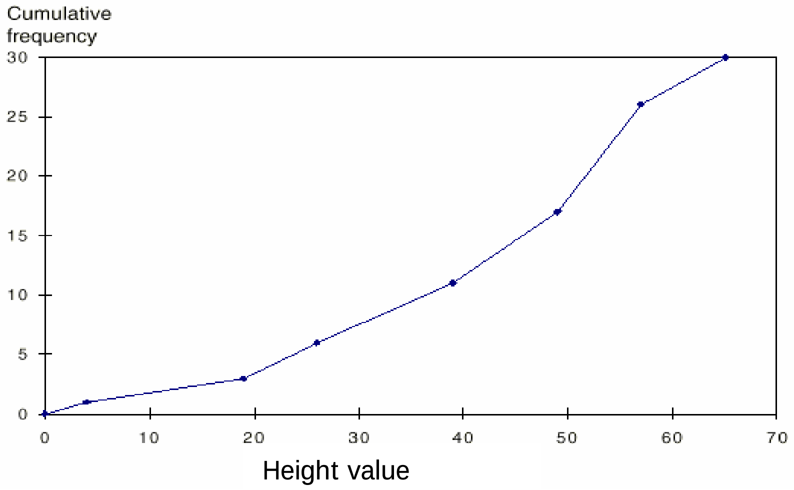
$n_A = 8$ and $n_B = 7$

EDA – Univariate Description

- ***Cumulative Frequency Tables and Histograms***

Continuous Variables

Height Values	Frequency	Cumulative Frequency
<=3	1	1
<=19	2	3
<=26	3	6
<=39.7	5	11
<=49.82	6	17
<= 57	9	26
<= 65	4	30



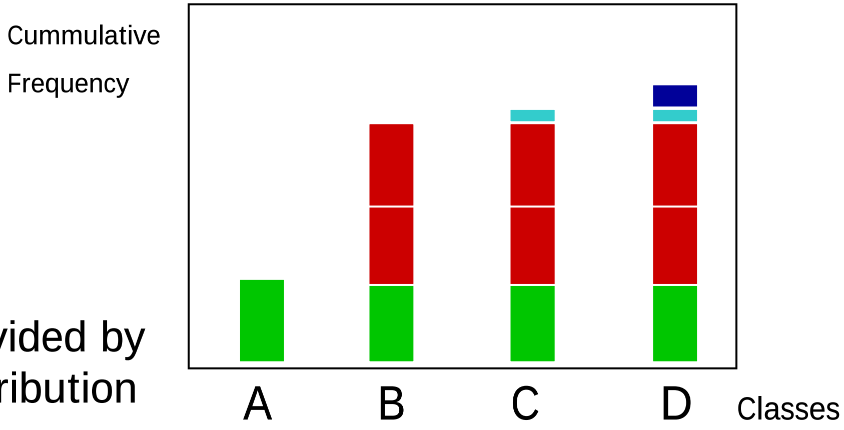
Categorical Variables

A = Basketball

B = Basketball + Football

C = Basketball + Football + Hokey

C = Basketball + Football + Hokey + Others



Important: The cummulative frequency values devided by the total population give the **cdf** (**c**umulative **d**istribution **f**unction) of the variable.

EDA – Univariate Description

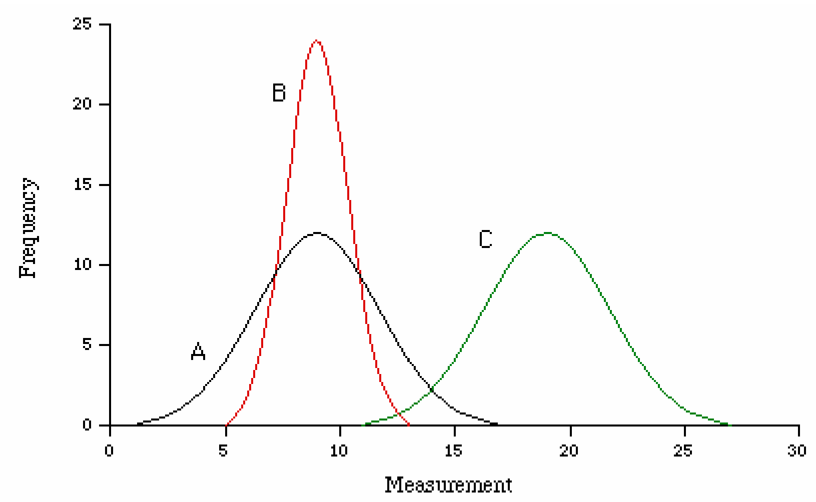
- **PDFs and CDFs - The Normal or Gaussian Distribution**

Graphically the normal distribution is best described by a 'bell-shaped' curve.

This curve is described in terms of the point at which its height is maximum (its 'mean') and how wide it is (its 'standard deviation').

It is a **parametric distribution** because it can be totally described by the two parameters, the mean and the standard deviation

It is mostly used in equations development and hypothesis (to be confirmed) for theoretical developments.

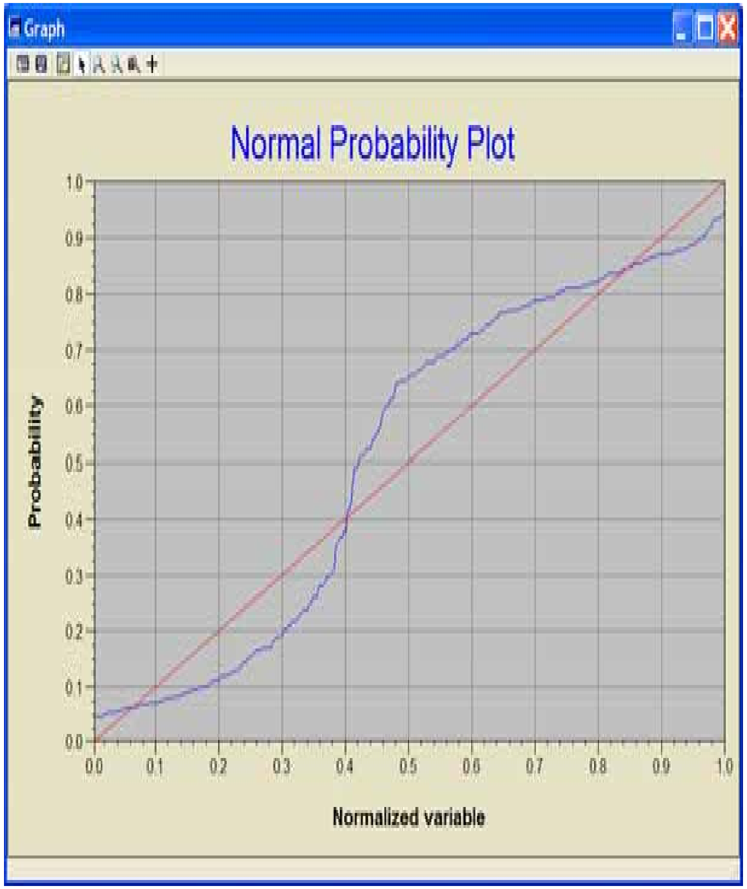


It can be described by:

$$Z = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left[(Y-\mu)/\sigma\right]^2}$$

EDA – Univariate Description

- **Normal Probability Plots**
  - Some of statistical tools work better, or only work, if the distribution of the data values is close to a Gaussian or Normal distribution.
  - On a Normal Probability Plot the y-axis is scaled in such a way that the cumulative frequencies will plot as a straight line if the distribution is Gaussian.

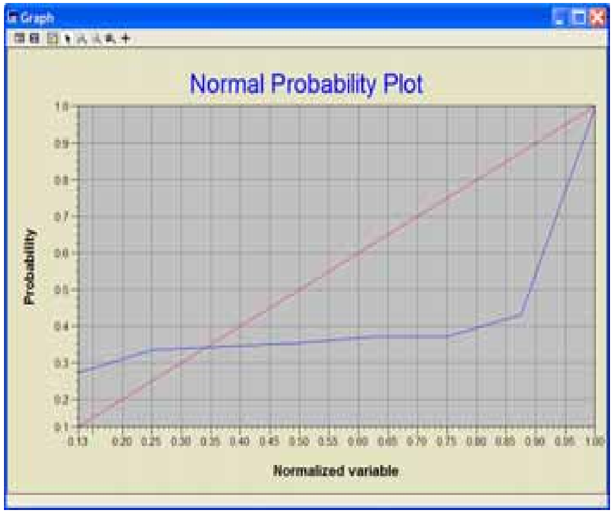


EDA – Univariate Description

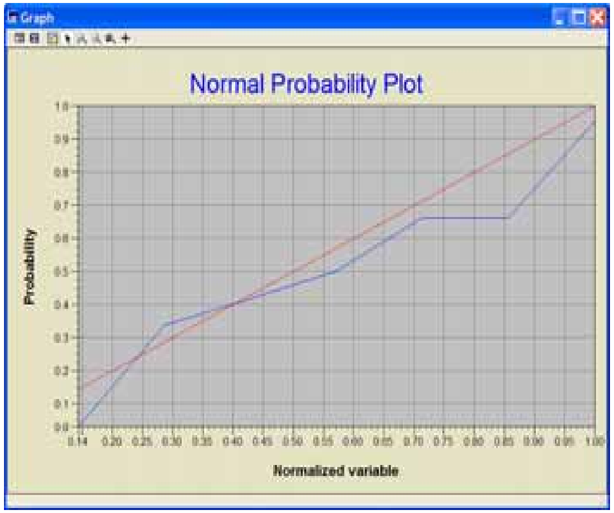
- **Normal Probability Plots** – a simple example

Ordered Sample Sets	
Set A	Set B
3	3
10	10
11	11
12	12
14	14
14	14
20	20
132	

Normal Probability Plots of Sets A and B



Distribution too far from the Gaussian behaviour

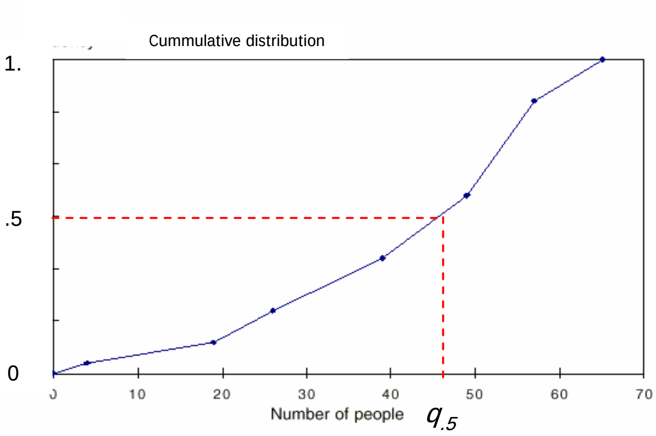


Distribution more closer to the Gaussian behaviour

EDA – Univariate Description

- **Summary Statistics – Measures of Location of the distributions**
  - ***Centers of the distribution***
    - **Mean** is the arithmetic average of α data values.

$$m = \mu = \bar{z} = \frac{1}{n}\sum_{\alpha=1}^{n} z(\alpha)$$

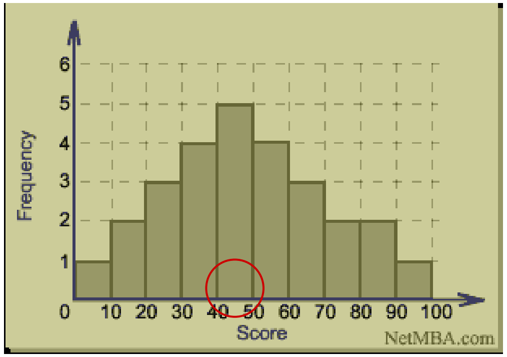


- **Median** Once the data is ordered: $z_1 \le z_2 \le .... \le z_n$

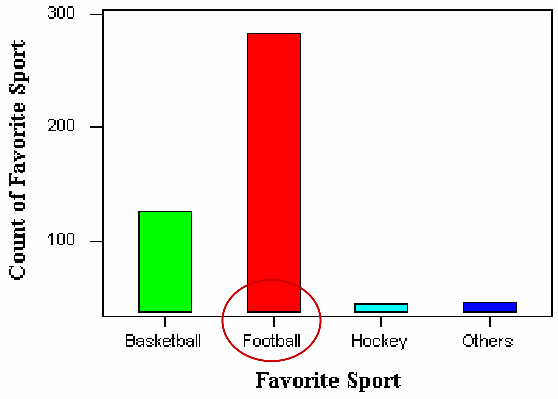
M = $z_{(n+1)/2}$ if n is odd or M = $(z_{n/2}+z_{(n/2)+1})/2$ if n is even.

Half of the values are below the median and half are above. M = $q_{.5}$

- **Mode** The value that occurs most frequently.



40 - <50



EDA – Univariate Description

- **Summary Statistics - Measures of Location of the distributions**
  - **Minimum and Maximum**

    The smallest and the largest values, respectively, in the data set.
  - **Lower and Upper Quartile**

    In the same way that the median splits the ordered data into halves, the quartis split the increasing sorted data into quarters.

    The median is the second quartil $q_{.5}$ or the Q_2
  - **Deciles, Percentiles, and Quantiles**

    Deciles split the data into tenths, percentiles split the data into hundredths. Quantiles are the generalization of this idea to any fraction. We represent as q_p the p-quantile of the data. Example: $q_{.1}$ is the first decil and $q_{.25}$ = Q_1 is the first quartil, and so on.

EDA – Univariate Description

- **Summary Statistics - Measures of Location of the distributions**

A simple example

Sample Sets	
Set A	Set B
3	3
10	10
11	11
12	12
14	14
14	14
20	20
132	

n_A = 8 and n_B = 7

	Set A	Set B
Minimum	3	3
Maximum	132	20
Mean	27	12
Mode	14	14
Median	13	12
Quartil 1	10,75	10,5
Quartil 2	13	12
Quartil 3	15,5	14

→ Obs. The Mean value is sensitive to erratic or extreme values

EDA – Univariate Description

- **Summary Statistics - Measures of Spread of the distributions**
  - **Variance** (second moment) is the average squared difference of the observed values from their mean value.

$$\sigma^2 = \frac{1}{n-1}\sum_{\alpha=1}^{n}\left(z(\alpha)-\bar{z}\right)^2$$

  - **Standard Deviation** The square root of the variance

$$\sigma = \sqrt{\sigma^2}$$

  - **Interquartile Range** - the difference between the upper and lower quartiles

$$IQR = Q_3 - Q_1 = q_{.75} - q_{.25}$$

Obs. The interquartile range does not depend on the mean value.

EDA – Univariate Description

- **Summary Statistics - Measures of Spread of the distributions**

A simple example

Sample Sets	
Set A	Set B
3	3
10	10
11	11
12	12
14	14
14	14
20	20
132	

$n_A = 8$ and $n_B = 7$

	Set A	Set B
Mean	27	12
Variance	1822,57	26,33
Std Var	42,69	5,13
Quartil 1	10,75	10,5
Quartil 3	15,5	14
Q_3-Q_1	4,75	3,5

→ Obs. Variance and Standard Deviation are also sensitive to erratic or extreme values

→ Obs. The interquartil value does not depend on the mean value.

EDA – Univariate Description

- **Summary Statistics - Measures of Shape**
  - Coefficient of Variation – determines if the distribution is large or narrow. It is unit free. CV >> 1 indicate Outliers.

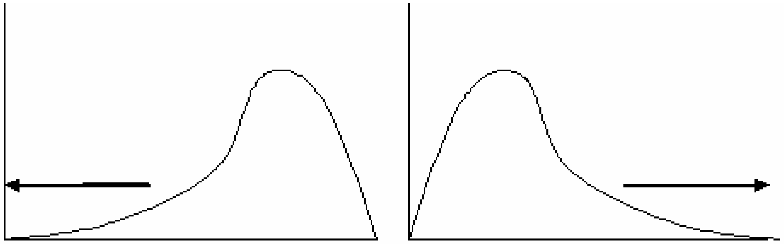
$$CV = \sigma / m$$

  - Skweness (coefficient of asymmetry) is a measure of symmetry

$$\gamma = \frac{\sum (Z - \bar{z})^3}{n\sigma^3}$$

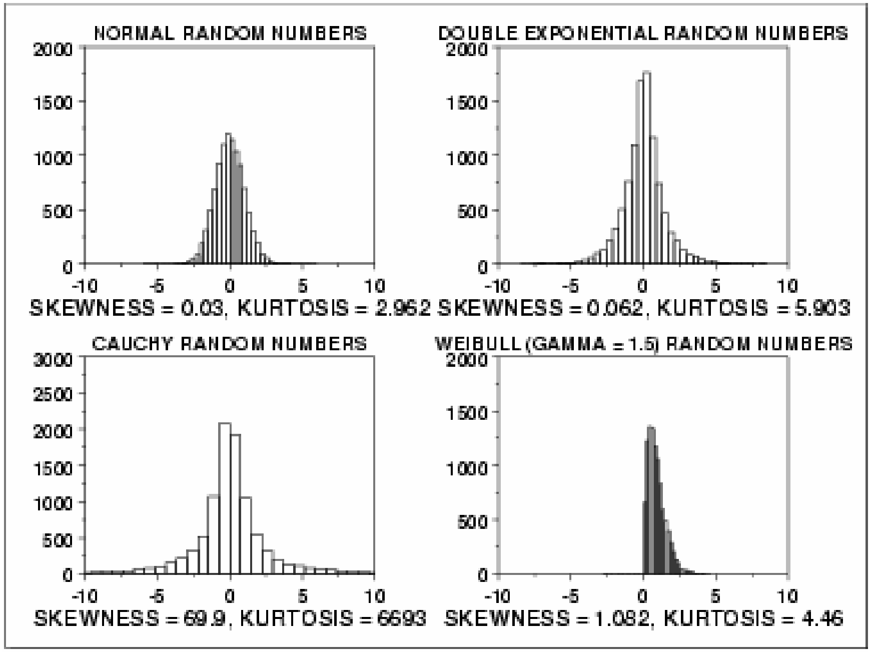
  - Curtosis is a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. Coefficient of flatness.

$$\beta = \frac{\sum (Z - \bar{z})^4}{n\sigma^4}$$



Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution

Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution



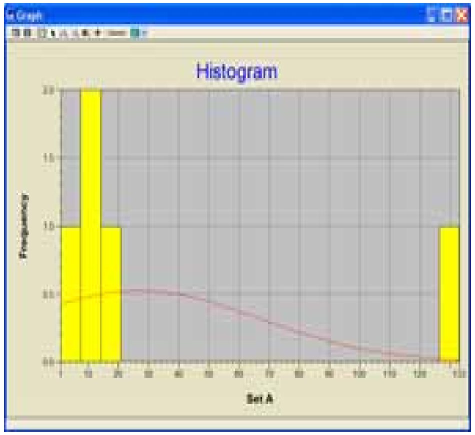
EDA – Univariate Description

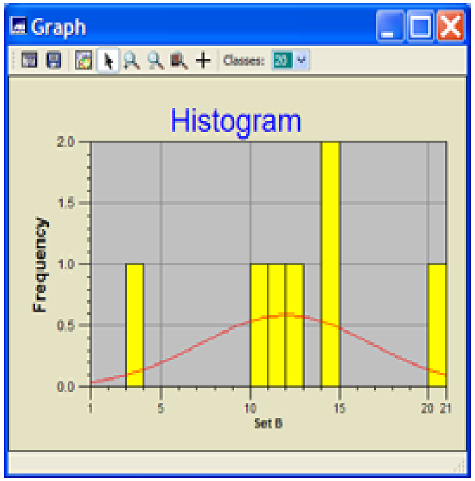
- **Summary Statistics - Measures of Shape of the distributions**

A simple example

Sample Sets	
Set A	Set B
3	3
10	10
11	11
12	12
14	14
14	14
20	20
132	

	Set A	Set B
Mean	27	12
Std Var	42,69	5,13
Coef. of Variation	1,58	0,43
Skewness	1,81	-0,22
Curtosis	4.60	2,21





EDA – Univariate Description

- **Detection of rough components**
  - Any data value can be thought of as comprising two components : one deriving from some *summary measure* (**smooth**) and the other a *residual component* (**rough**)

DATA = **smooth** PLUS **rough**

Rough example

Outliers: values more than a certain distance above (below) the upper (lower) quartile of the distribution (Haining 1993,201). (Errors or Atypical values)

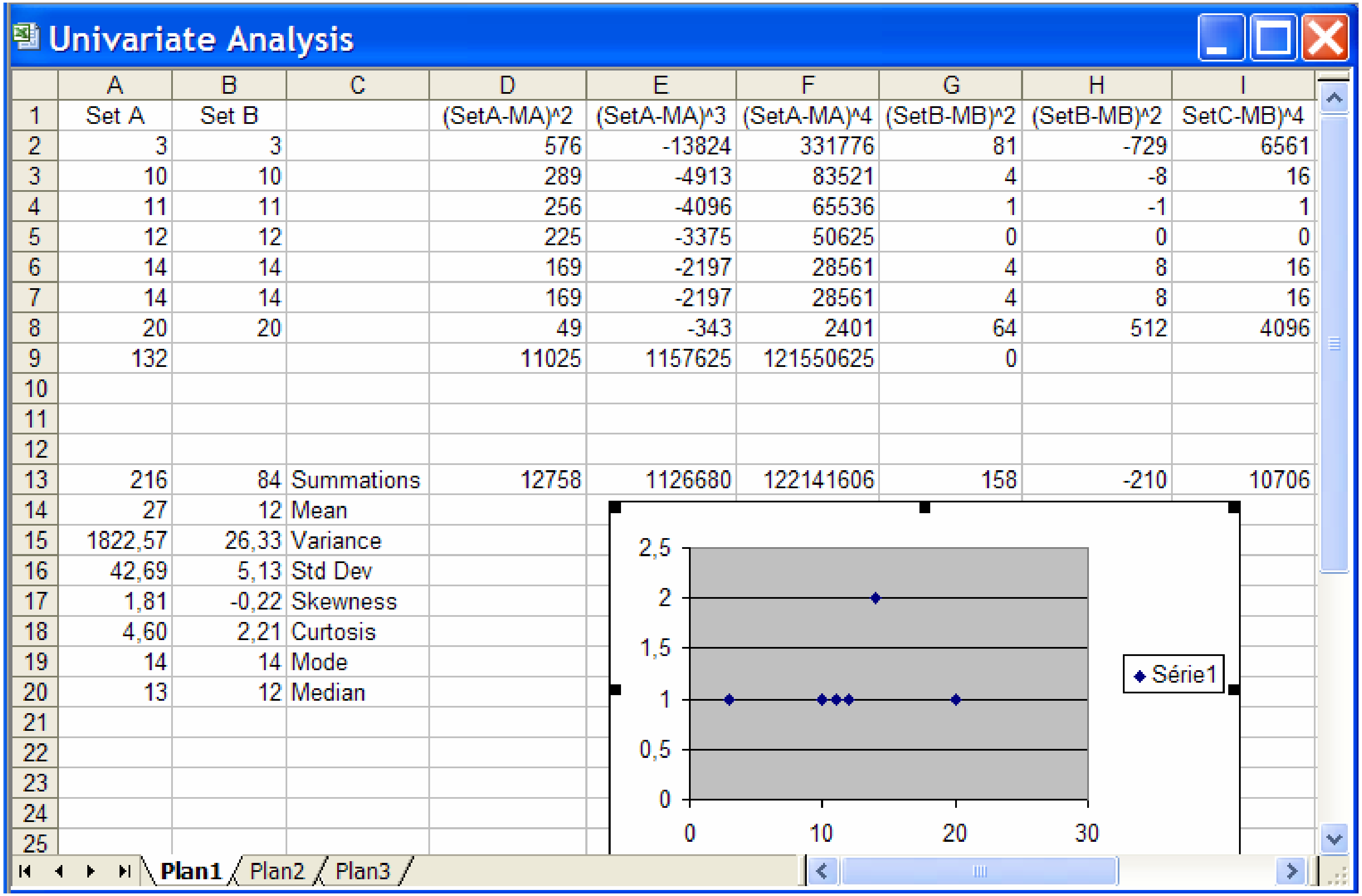
For some data analyses, for the spatial continuity of the data for example, the outliers must be removed.

Sample Sets	
Set A	Set B
3	3
10	10
11	-1111
12	12
1400	14
14	14
20	20
132	

Outliers

EDA– Univariate Description

- **Summary Statistics – Meas. of Location, Spread and Shape - Use of SpreadSheets**

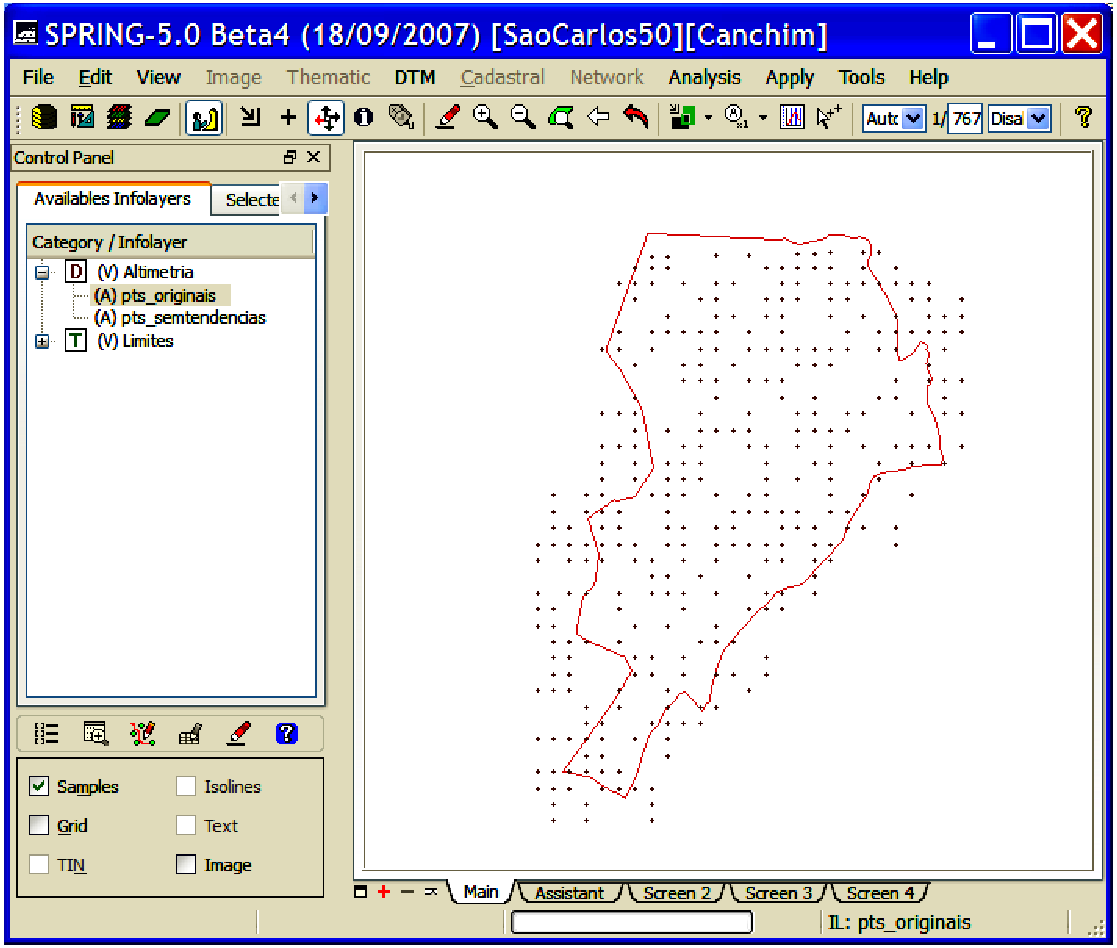


Univariate Analysis

	A	B	C	D	E	F	G	H	I
1	Set A	Set B		(SetA-MA)^2	(SetA-MA)^3	(SetA-MA)^4	(SetB-MB)^2	(SetB-MB)^2	SetC-MB)^4
2	3	3		576	-13824	331776	81	-729	6561
3	10	10		289	-4913	83521	4	-8	16
4	11	11		256	-4096	65536	1	-1	1
5	12	12		225	-3375	50625	0	0	0
6	14	14		169	-2197	28561	4	8	16
7	14	14		169	-2197	28561	4	8	16
8	20	20		49	-343	2401	64	512	4096
9	132			11025	1157625	121550625	0		
10									
11									
12									
13	216	84	Summations	12758	1126680	122141606	158	-210	10706
14	27	12	Mean						
15	1822,57	26,33	Variance						
16	42,69	5,13	Std Dev						
17	1,81	-0,22	Skewness						
18	4,60	2,21	Curtosis						
19	14	14	Mode						
20	13	12	Median						
21									
22									
23									
24									
25									

EDA – Univariate Description
An Example with Elevation Data

A set of 329 elevations sampled inside São Carlos farm region



Category / Infolayer

- D (V) Altimetria
  - (A) pts_originais
  - (A) pts_semtendencias
- T (V) Limites

☑ Samples □ Isolines

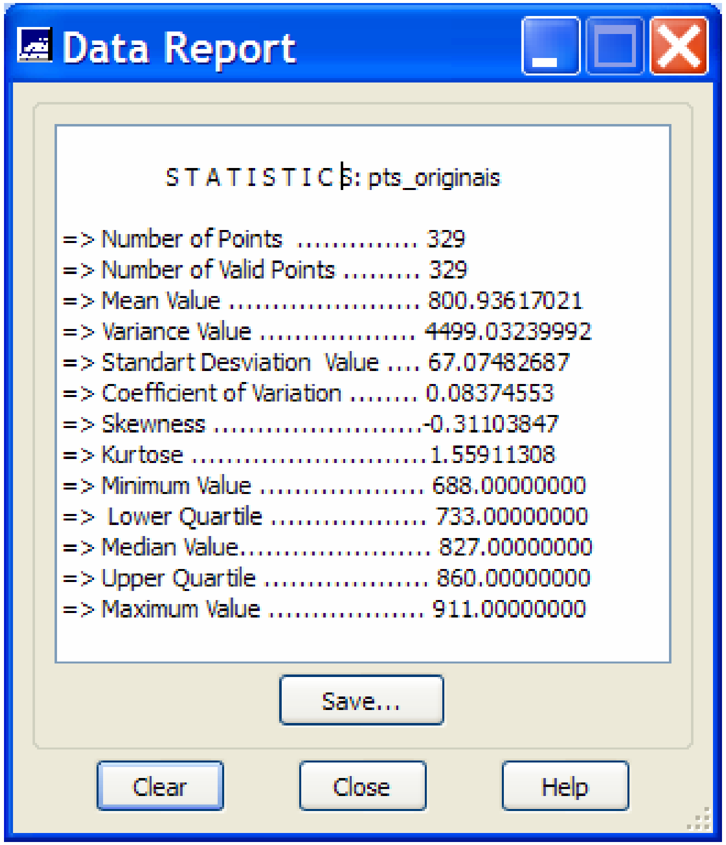
□ Grid □ Text

□ TIN □ Image

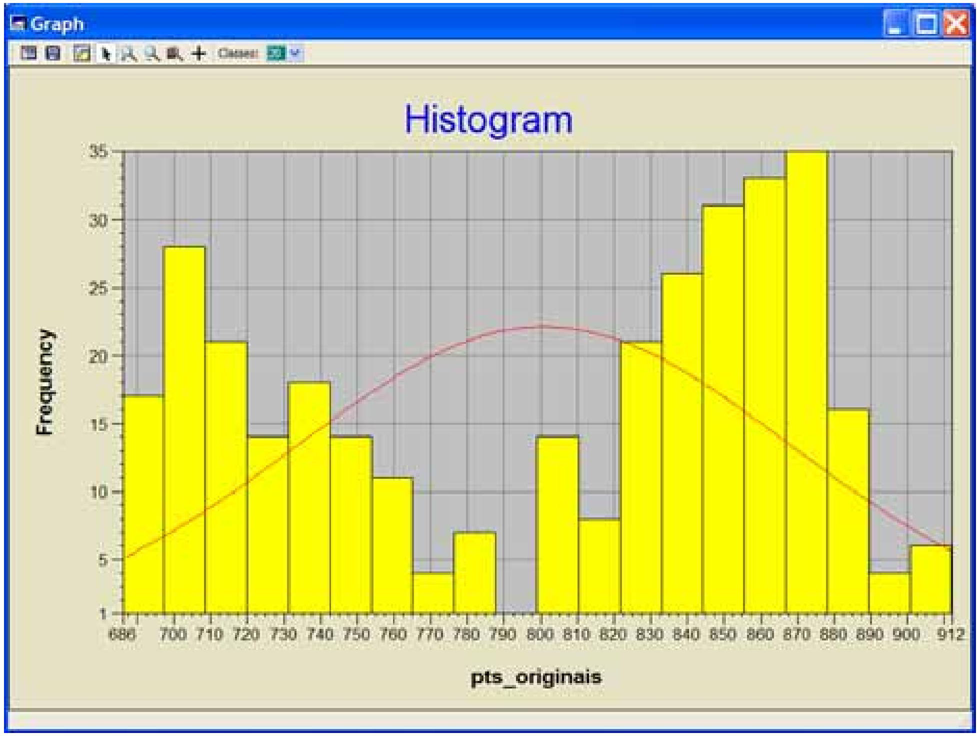
IL: pts_originais

EDA – Univariate Description
An Example with Elevation data

Summary Reports and Histograms

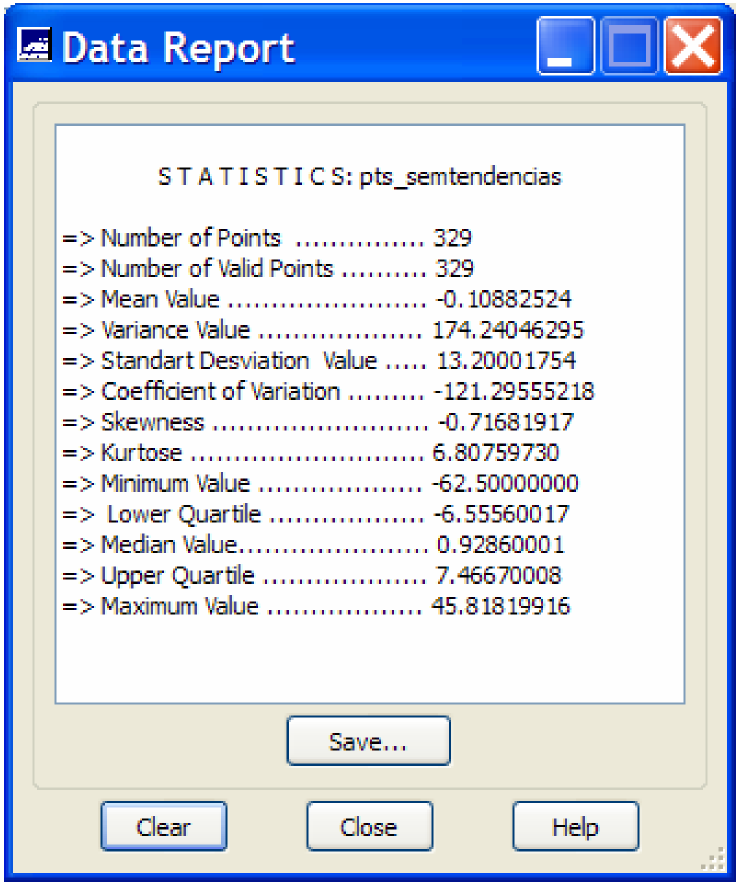


```
STATISTICS: pts_originais

=> Number of Points ............. 329
=> Number of Valid Points ......... 329
=> Mean Value ...................... 800.93617021
=> Variance Value ................. 4499.03239992
=> Standart Desviation  Value .... 67.07482687
=> Coefficient of Variation ........ 0.08374553
=> Skewness .......................-0.31103847
=> Kurtose ..........................1.55911308
=> Minimum Value ................... 688.00000000
=>  Lower Quartile ................. 733.00000000
=> Median Value..................... 827.00000000
=> Upper Quartile .................. 860.00000000
=> Maximum Value ................... 911.00000000
```

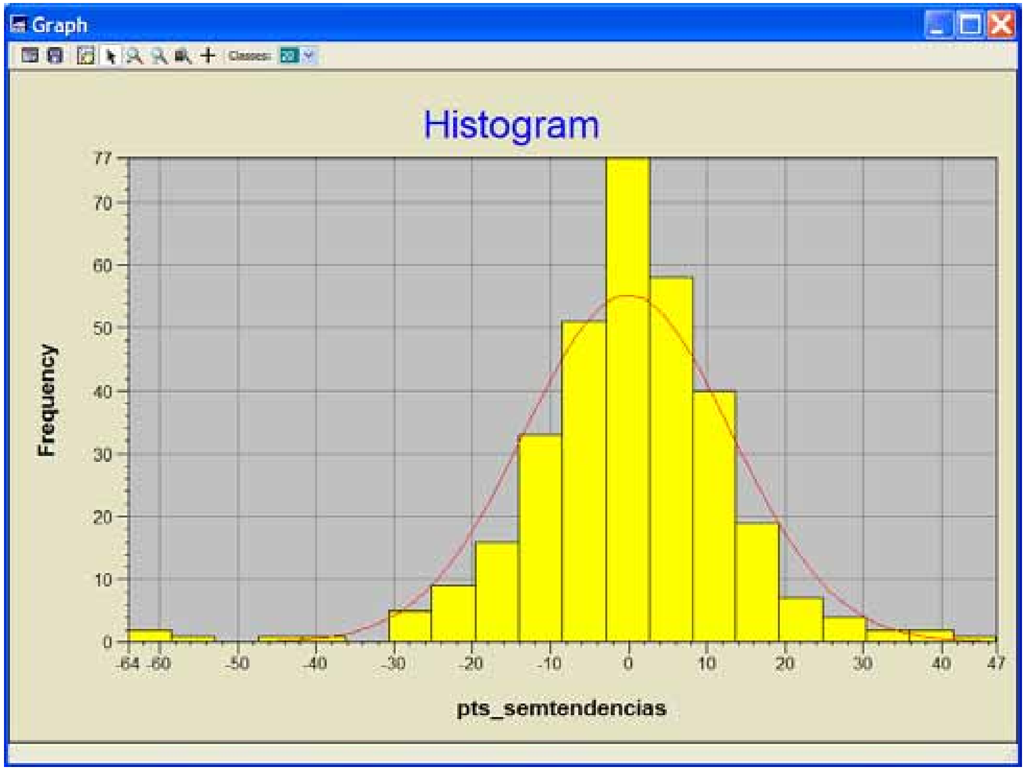


EDA – Univariate Description
An Example with Elevation data

Summary Reports and Histograms

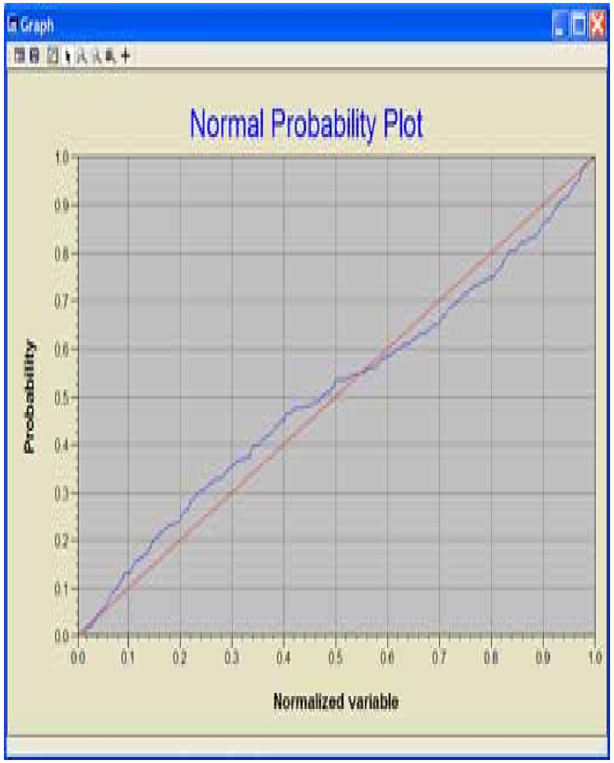


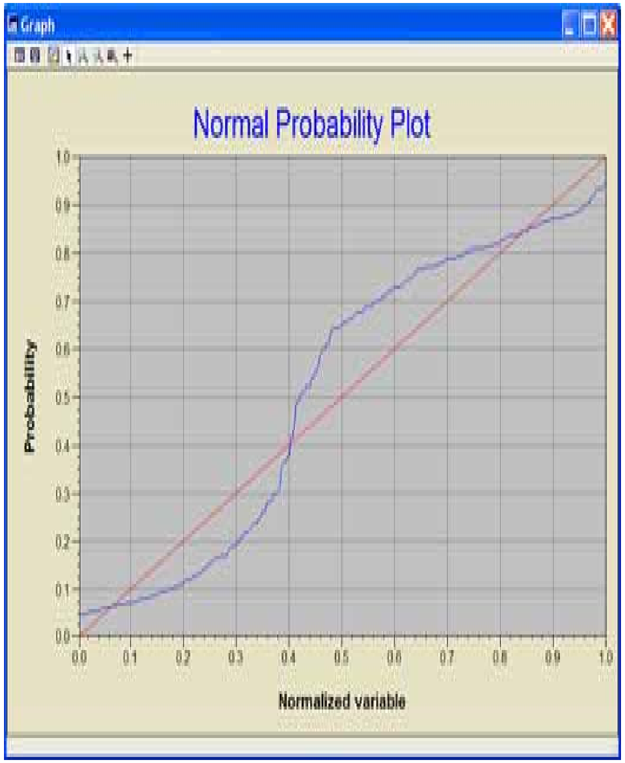
```
S T A T I S T I C S: pts_semtendencias

=> Number of Points ............... 329
=> Number of Valid Points .......... 329
=> Mean Value ....................... -0.10882524
=> Variance Value .................. 174.24046295
=> Standart Desviation  Value ..... 13.20001754
=> Coefficient of Variation ......... -121.29555218
=> Skewness ......................... -0.71681917
=> Kurtose .......................... 6.80759730
=> Minimum Value .................. -62.50000000
=>  Lower Quartile ................. -6.55560017
=> Median Value..................... 0.92860001
=> Upper Quartile .................. 7.46670008
=> Maximum Value ................. 45.81819916
```



EDA – Univariate Description
An Example with Elevation data

NORMAL PROBABILITY PLOTS





Exploratory Data Analysis

Bivariate Description

- The ScatterPlot
- Correlation between two variables
- The covariance value
- The correlation coefficient
- Linear Regression
- Conditional Expectation

EDA – Bivariate Description

Objectives

Representation of relations/co-relations between two variables

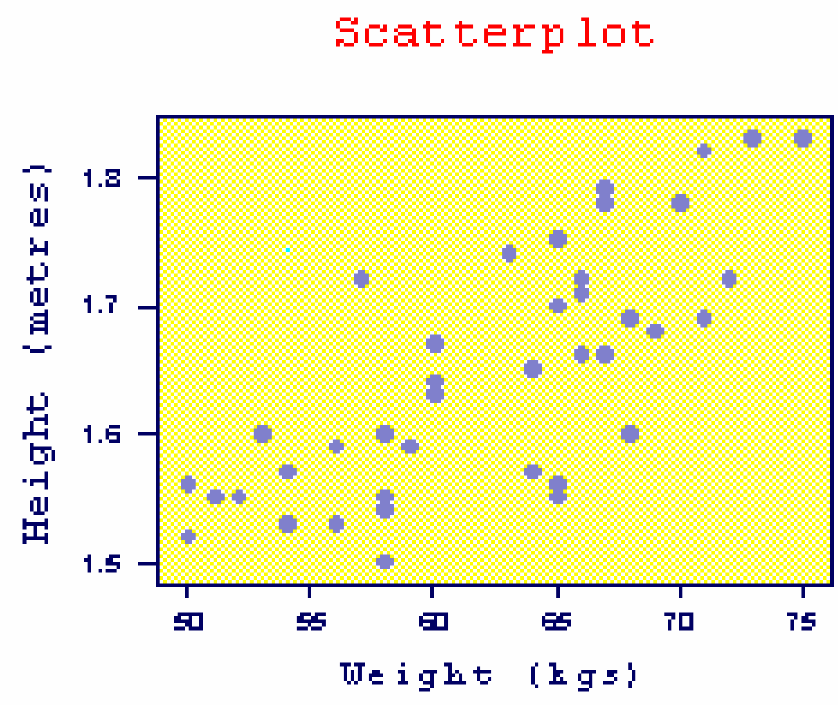
Comparing the distributions of the two variables

Use of graphics and statistic measurements that summarize the main features of the bivariate relations.

EDA – Bivariate Description

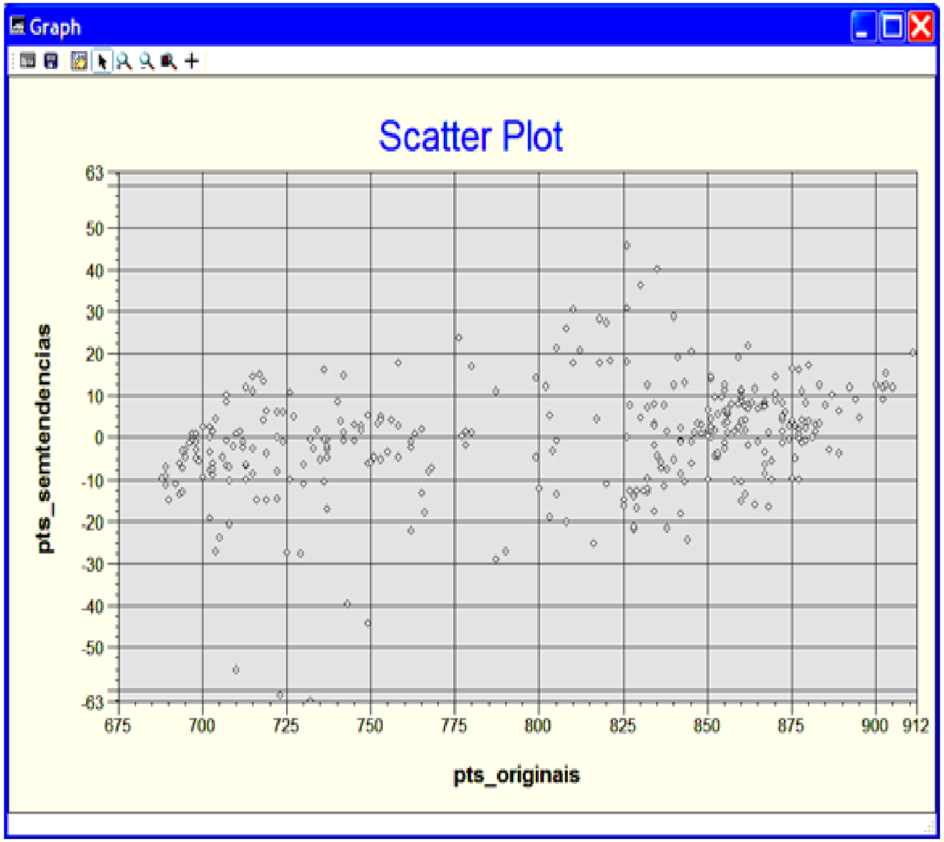
The Scatterplot

A scatterplot is a useful summary of a set of bivariate data (two variables X and Y), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model. Each unit contributes one point to the scatterplot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables.

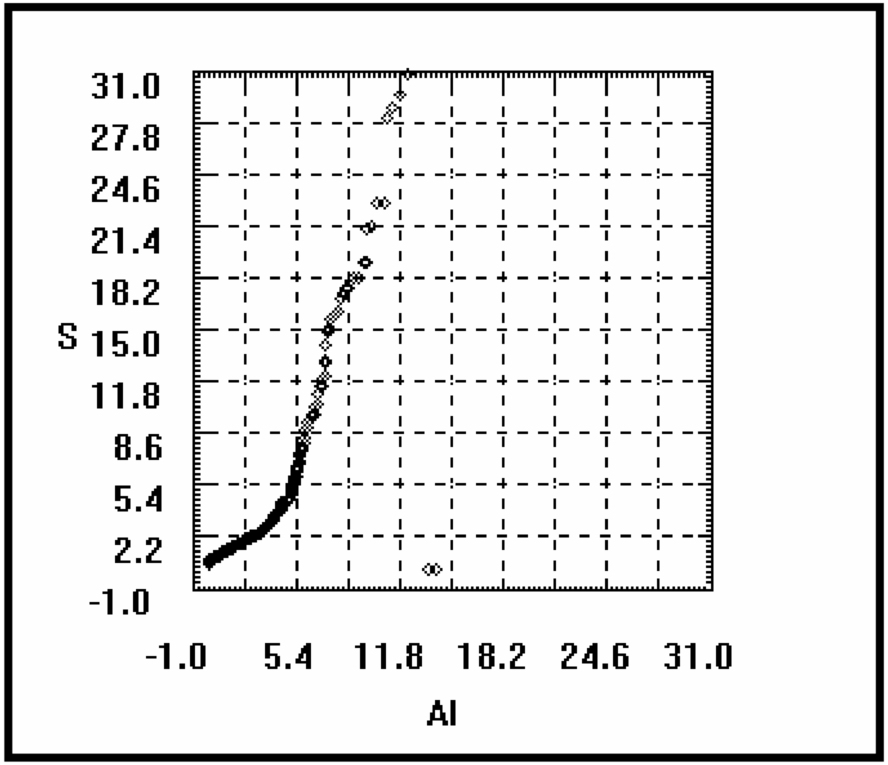


EDA – ScatterPlot Examples

SCATTERPLOTS



Elevation Data



Sulphur x Aluminium Data

EDA – Bivariate Description

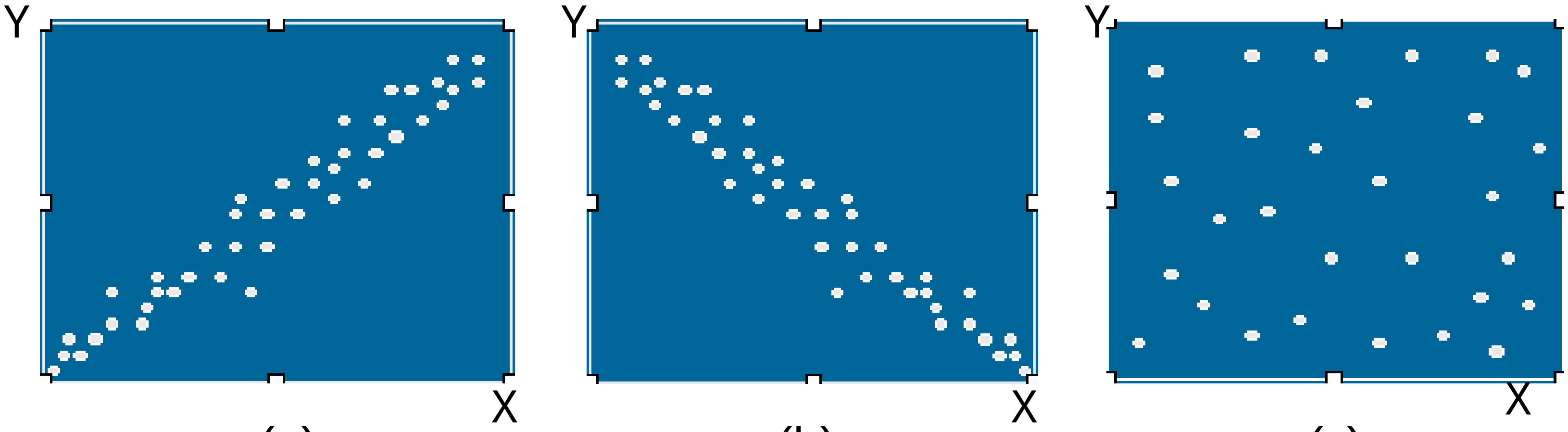
Correlation Between two Variables – Considering variables Z_i and Z_j

In a scatterplot we can observe, generally, one of the three patterns:

(a) *positively correlated*: larger values of Z_i tend to be associated to larger values of Z_j

(b) *negatively correlated*: larger values of Z_i tend to be associated to larger values of Z_j

(c) *uncorrelated*: the two variables are not related



(a) (b) (c)

EDA – Bivariate Description

Covariance Value – Considering variables Z_i and Z_j

The **Covariance** between two variables is used itself as a summary statistic of a scatterplot. It measures the joint variation of the two variables around their means. It is computed as:

$$C_{ij} = \sigma_{ij} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(z_i(\alpha) - m_i\right)\cdot\left(z_j(\alpha) - m_j\right)$$

The covariance depends on the unit of the variables

It is formula can be extended for more than two variables

EDA – Bivariate Description

Correlation Coefficient – Considering variables Z_i and Z_j

The **Correlation Coefficient** is most commonly used to summarize the relationship between two variables. It can be calculated from:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{1}{n} \frac{\sum_{\alpha=1}^{n} \left(z_i(\alpha) - m_i\right) \cdot \left(z_j(\alpha) - m_j\right)}{\sigma_i \sigma_j}$$

The correlation value is units free – covariance standardized by the standard deviations

The correlation value varies between -1 to 1, $\rho_{ij} \in [-1,1]$

-1 – means that the variables are negatively correlated

0 – means that the variables are totally uncorrelated

1 – means that the variables are positively correlated

EDA – Bivariate Description

Linear Regression – Considering variables V_x and V_Y

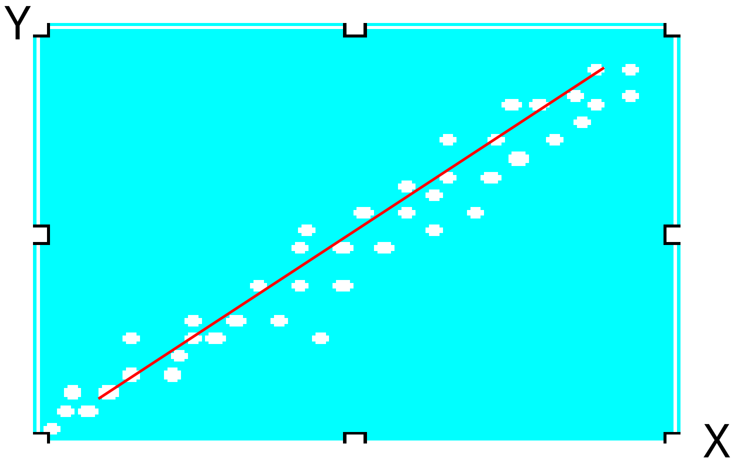
Assuming that the dependence of one variable on to the other is linear, the **linear regression** help us to predict one variable if the other is known. In this case the linear equation is given by the straight line representation:

$$y = ax + b$$

where the slope, *a*, and the constant, *b*, are given by:

$$a = \rho \frac{\sigma_y}{\sigma_x}$$

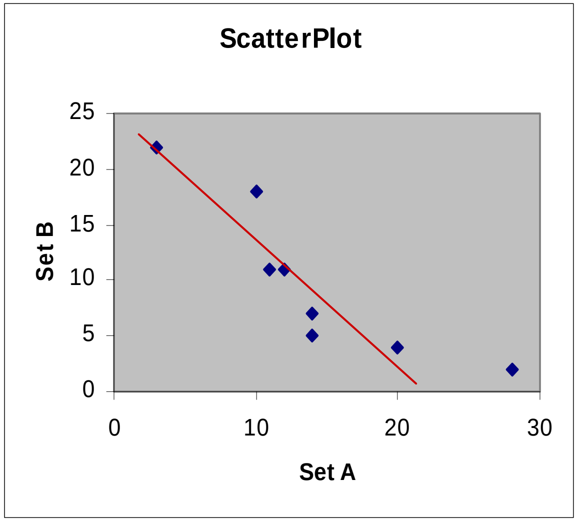
$$b = m_x - a \cdot m_y$$



EDA – Bivariate Description

Simple Numerical Example

Sample Sets	
Set A	Set B
3	22
10	18
11	11
12	11
14	7
14	5
20	4
28	2



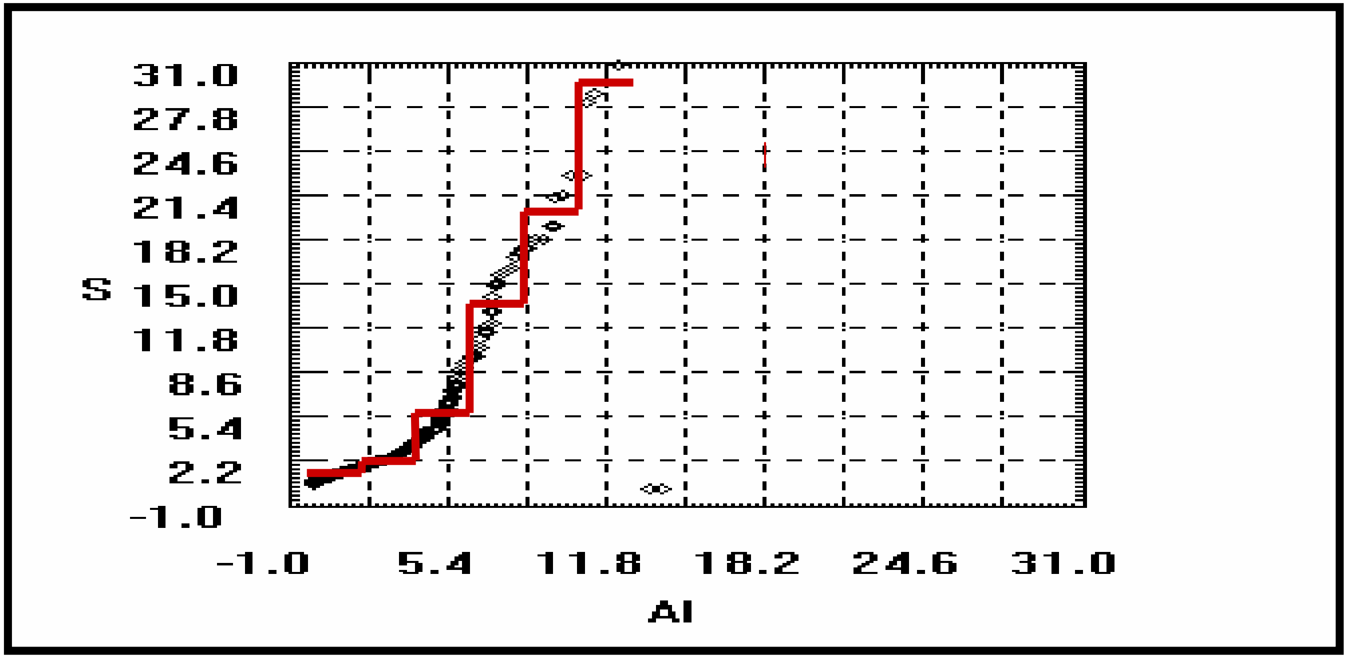
	Set A	Set B
Mean	14	10
Std Var	7,39	7,01
Covariance	-39,63	
Skewness	-0,875	
Linear Regression	a =-0,83 (-41,19^{o})	b=22,3

$n_A = n_B = 8$

EDA – Bivariate Description

Conditional Expectation – Considering variables V_i and V_j

- Alternative when linear regression is not a good fit (curve with bends on it)
- Consists on calculating the mean value of y for different ranges of x (conditional expectation values)
- Huge number of very narrow classes give smooth curves
- Problems can occur for too narrow classes (no points in some ranges of x)



EDA – Summary and Conclusions

Summary and Conclusions

- EDA/ESDA provides a set of robust tools for exploring spatial data, which do not require a knowledge of advanced statistics for their use.

- GIS are currently only poorly equipped with many of these tools, despite containing the basic functionality to allow them to be implemented.

Geostatistics – EDA

END of Presentation

Questions about EDA

1. Explain, using your own words, the difference between the EDA and ESDA terms.
2. Given the data sets A and B (in the right side) perform Univariate and Bivariate Exploratory Data Analyses on them. Compare your results with the results given by statistical (or not) function that are available in a spreadsheet (excel for example).
3. Write a report with the EDA results and with considerations related to the analysis you have done. Send the report to the professor e-mail (cfelgueiras@isegi.unl.pt) before 12/Oct/2007.
4. Download, from the geostatistics course in ISEGI online, the Text input data for the first lab and the Lab1 – Exercise – Creating a DataBase in SPRING. Run the lab1 by yourself. Contact the geostatistics professor by e-mail or at his office (room 21, ISEGI building) if you have problems on running the lab.

Sample Sets		
Id.	Set A	Set B
1	33,3	101
2	10,18	34,3
3	11	29,7
4	22	65
5	78,4	234,3
6	13	35,7
7	2220	22
8	28	57,57
9	66,6	330
10	12.34	40
11	37,2	115,2