

Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali  
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY  
I YEAR

Course:

**EXPERIMENTAL DESIGN AND  
CHEMOMETRICS IN FOOD**  
(5 credits – 38 hours)

Teacher: Marcello Mascini  
([mmascini@unite.it](mailto:mmascini@unite.it))

The Teacher is available to answer questions at the end of the lesson, or on request by mail

# The course is split in 4 units

## UNIT 1: statistical regression

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the  $\chi^2$  method, Validation of the model

## UNIT 3: Data Matrices and sensor arrays

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

## UNIT 2: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

## UNIT 4: Elements of Pattern recognition

Cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

# **UNIT 1: statistical regression**

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

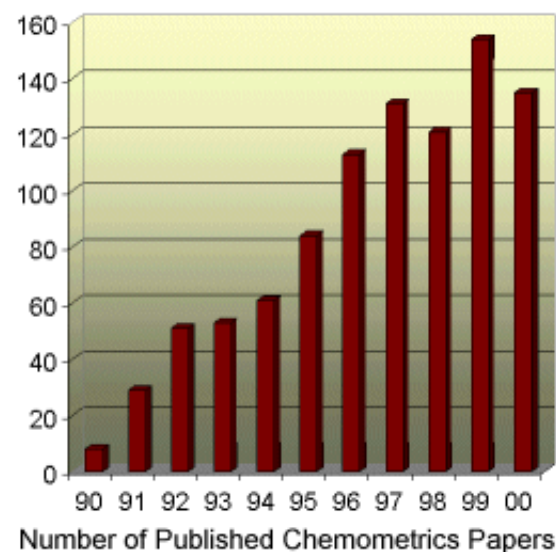
The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the  $\chi^2$  method, Validation of the model

# CHEMOMETRICS

- The science of extracting information from chemical systems by data-driven means.
- It is a highly interfacial discipline, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, biochemistry, medicine, biology and chemical engineering.
- The goal is using data from multidimensional signals for examples spectrometers or chromatograms



## Dedicated journals

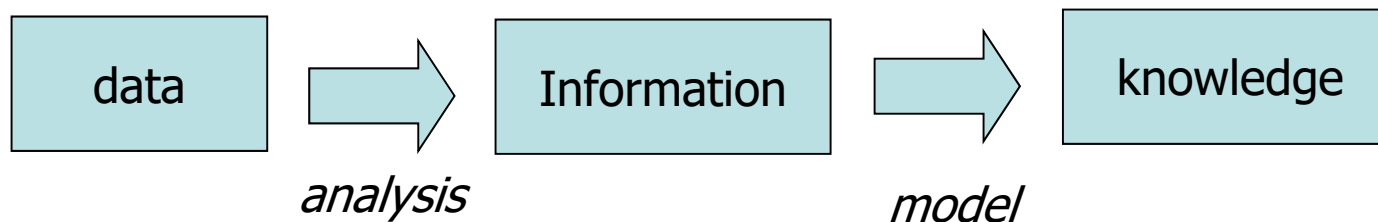
- Chemometrics and Intelligent Laboratory systems
- Journal of Chemometrics

## Articles are published also in:

- Analytical Chemistry
- Analytica Chimica Acta
- Trends in Analytical Chemistry
- J. computer aided molecular design
- .....

# DATA

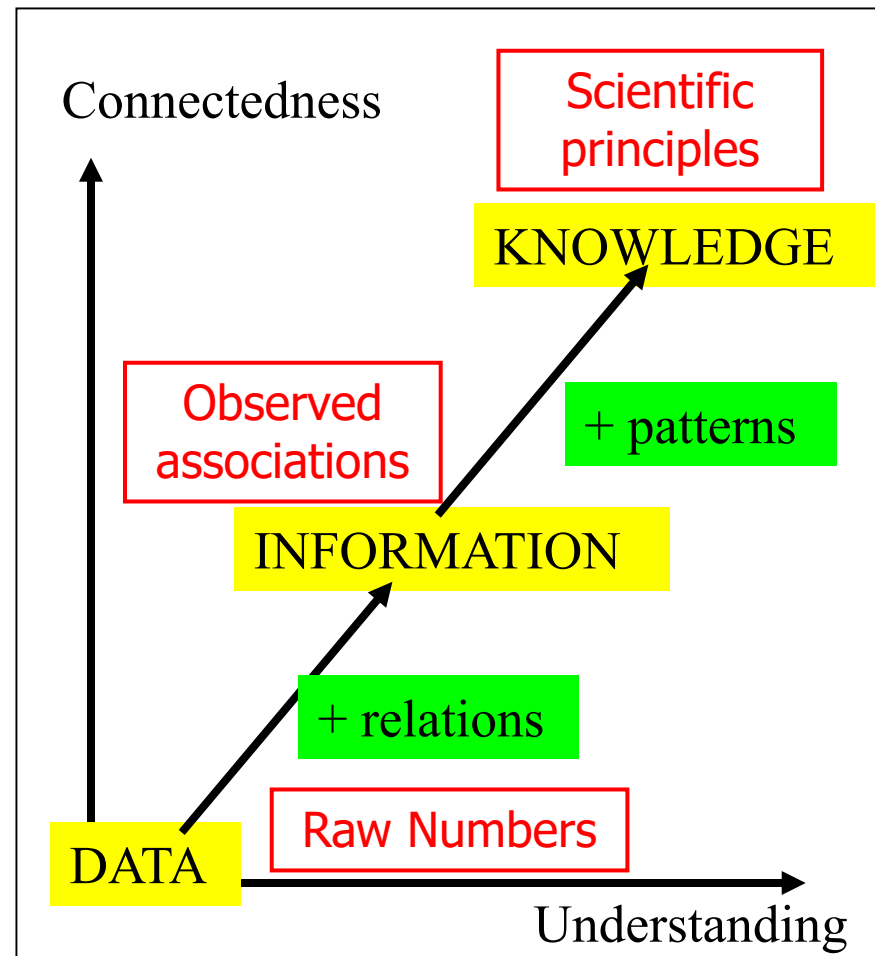
- Data are individual pieces of information.
- For Example human being data:
  - Height, weight, chemical blood analysis DNA composition, hair color...
- Data can be qualitative or quantitative
- Data must be analysed to have information and to increase knowledge
  - Example: a chemical blood analysis has to be supported by a human being model



# Data ⇒ Information ⇒ Knowledge

The aim of data-mining can be illustrated graphically as follows:

- Data
  - unrelated *facts*
- Information
  - facts plus *relations*
- Knowledge
  - information plus *patterns*



# Univariate analysis

❖ Describing the distribution of a single variable, including its central tendency (including the mean, median, and mode) and dispersion (including the range and quantiles of the data-set, and measures of spread such as the variance and standard deviation). Characteristics of a variable's distribution may also be depicted in graphical or tabular format, including histograms and stem-and-leaf display.

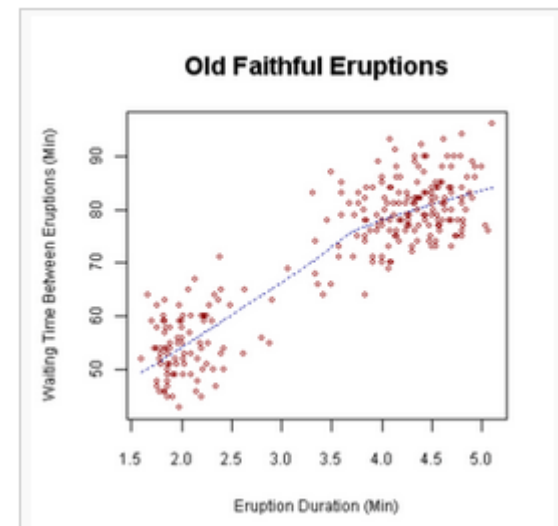
Age range	Number of cases	Percent
under 18	10	5
18–29	50	25
29–45	40	20
45–65	40	20
over 65	60	30
Valid cases: 200		
Missing cases: 0		

❖ Any measurement can be judged by the following meta-measurement criteria values: level of measurement (which includes magnitude), dimensions (units), and uncertainty:

- Electrical resistance is 100K $\Omega$
- The apple weight is è 80g
- The K<sup>+</sup> concentration in water is 1.02 mg/l

# Bivariate analysis

- ❖ It involves the analysis of two variables (often denoted as  $X$ ,  $Y$ ), for the purpose of determining the empirical relationship between them. In order to see if the variables are related to one another, it is common to measure how those two variables simultaneously change together (covariance).
- ❖ The major differentiating point between univariate and bivariate analysis, in addition to the latter's looking at more than one variable, is that the purpose of a bivariate analysis goes beyond simply descriptive: it is the analysis of the relationship between the two variables. Bivariate analysis is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously)





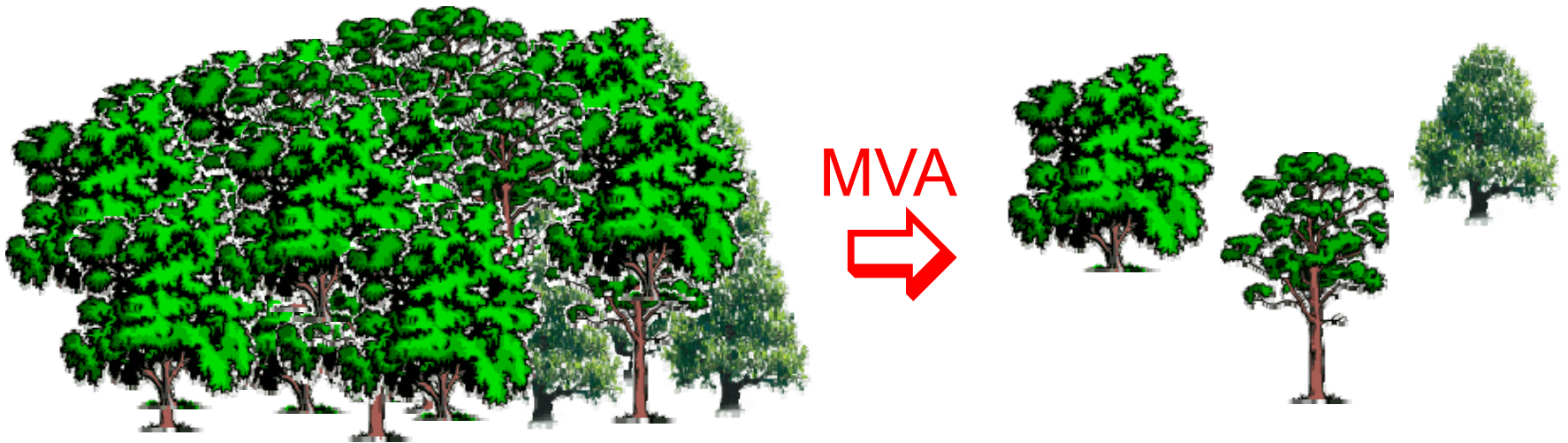
# Why Multivariate?

- Typically more than one measurement is taken on a given experimental unit
- Need to consider all the measurements together so that one can understand how they are related
- Need to consider all the measurements together so that one can extract essential structure

# What is MVA?

Multivariate analysis (MVA) is defined as the simultaneous analysis of more than five variables. Some people use the term “megavariate” analysis to denote cases where there are more than a hundred variables.

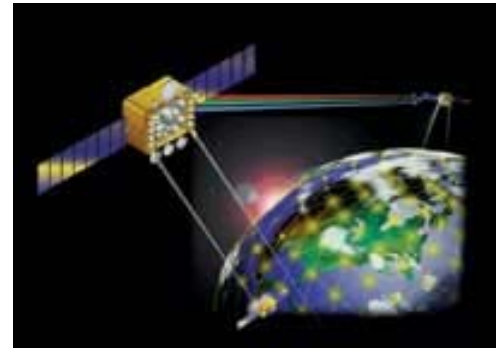
MVA uses ALL available data to capture the most information possible. The basic principle is to boil down hundreds of variables down to a *mere handful*.



# Process Integration Challenge: Make sense of masses of data

Many organisations today are faced with the same challenge: TOO MUCH DATA. These include:

- Business - *customer transactions*
- Communications - *website use*
- Government - *intelligence*
- Science - *astronomical data*
- Pharmaceuticals - *molecular configurations*
- Industry - *process data*



It is the last item that is of interest to us as chemical engineers...

# Graphical representation of MVA

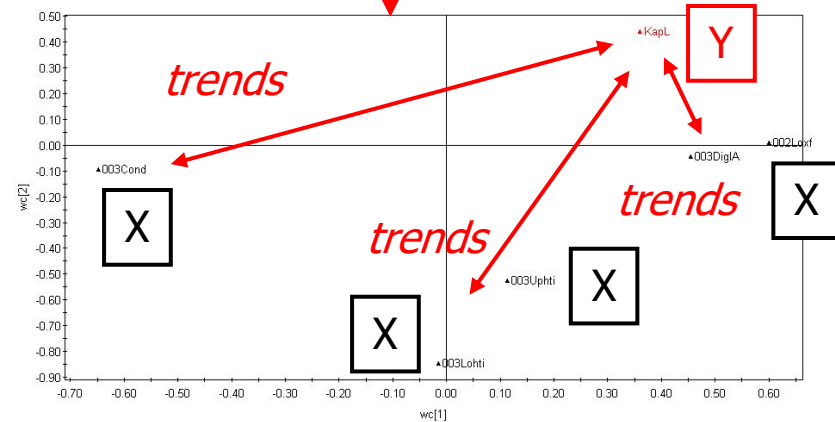
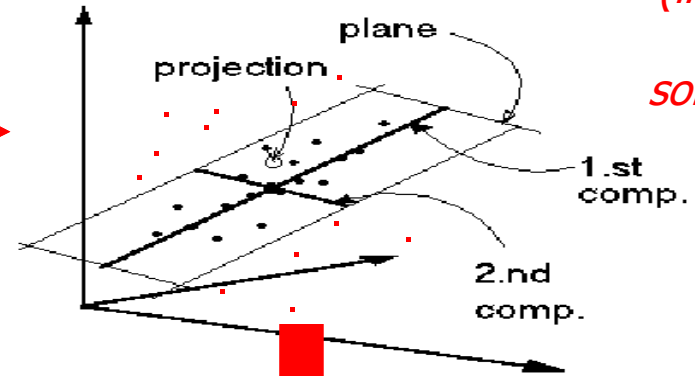
Tmt	X1	X4	X5	Rep	Y avec	Y sans
1	-1	-1	-1	1	2.51	2.74
1	-1	-1	-1	2	2.36	2.89
1	-1	-1	-1	3	2.45	2.56
2	-1	0	1	1	2.63	3.23
2	-1	0	1	2	2.55	2.47
2	-1	0	1	3	2.65	2.31
3						2.67
3						2.45
3						2.98
4						3.22
4						2.57
4	0	-1	1	3	2.97	2.63
5	0	0	0	1	2.89	3.16
5	0	0	0	2	2.56	3.32
5	0	0	0	3	2.52	3.26
6	0	1	-1	1	2.44	3.1
6	0	1	-1	2	2.22	2.97
6	0	1	-1	3	2.27	2.92

Raw Data:  
*impossible to interpret*



## Statistical Model

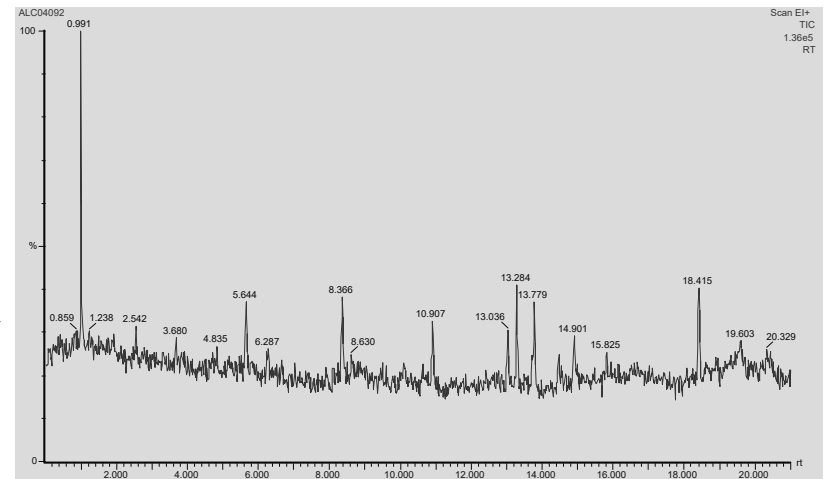
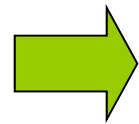
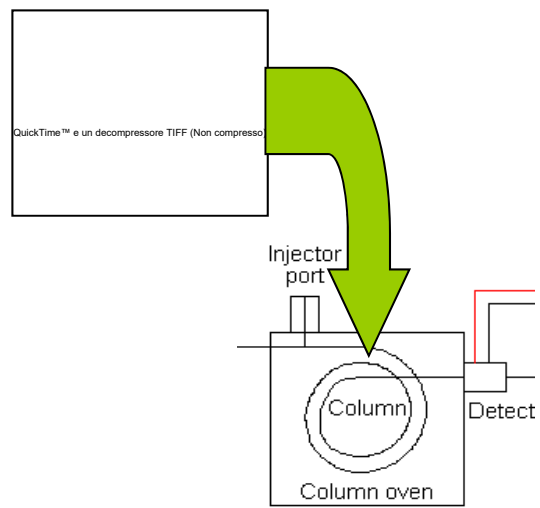
*(internal to software)*



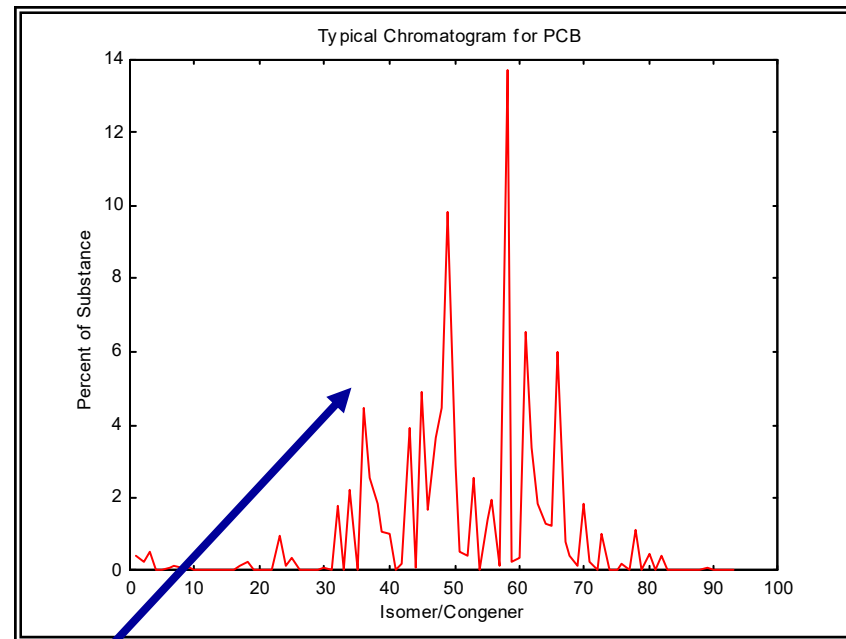
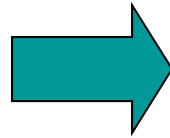
## 2-D Visual Outputs

# Multidimensional Instruments

- **Gas chromatography**



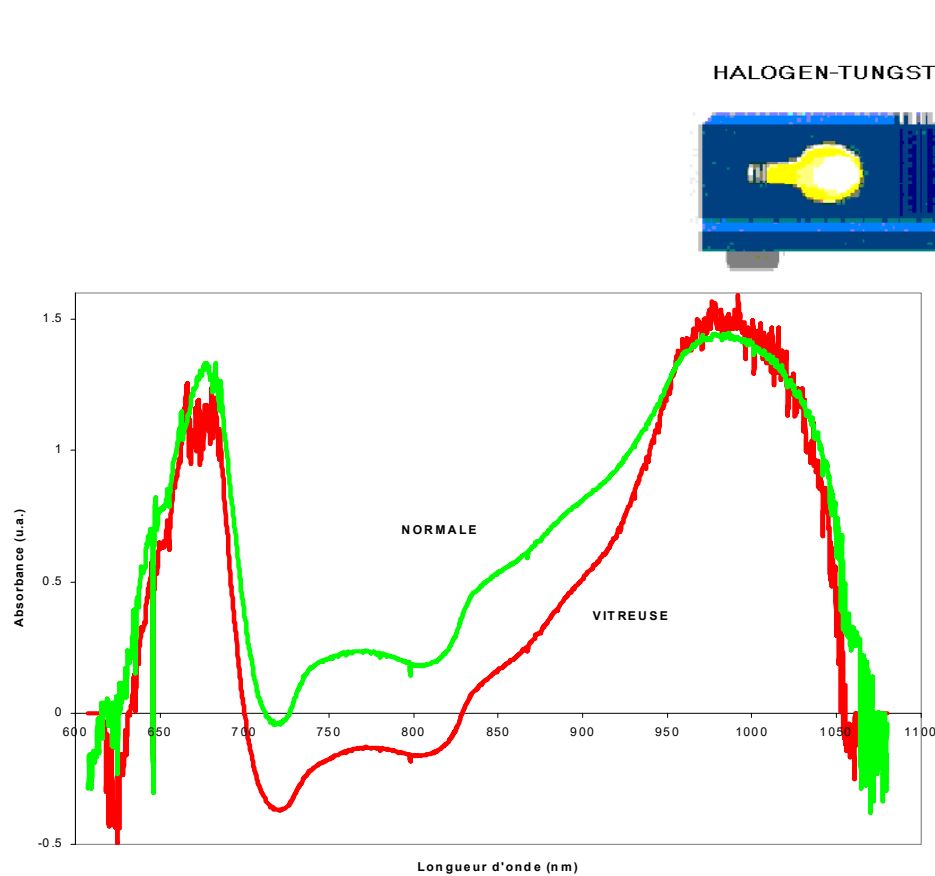
# In Chromatography



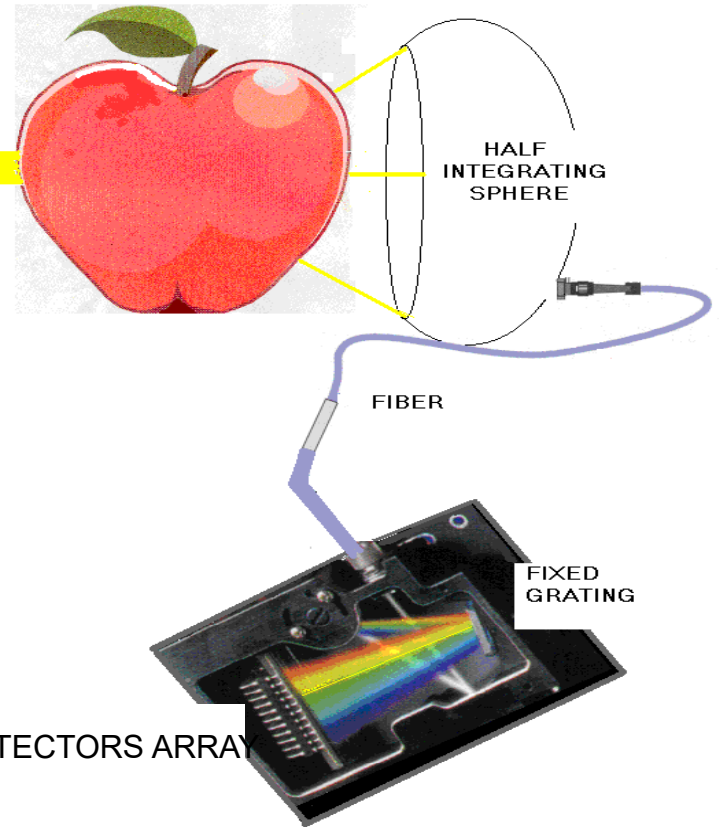
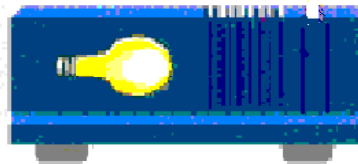
one observation

# Multidimensional Instruments

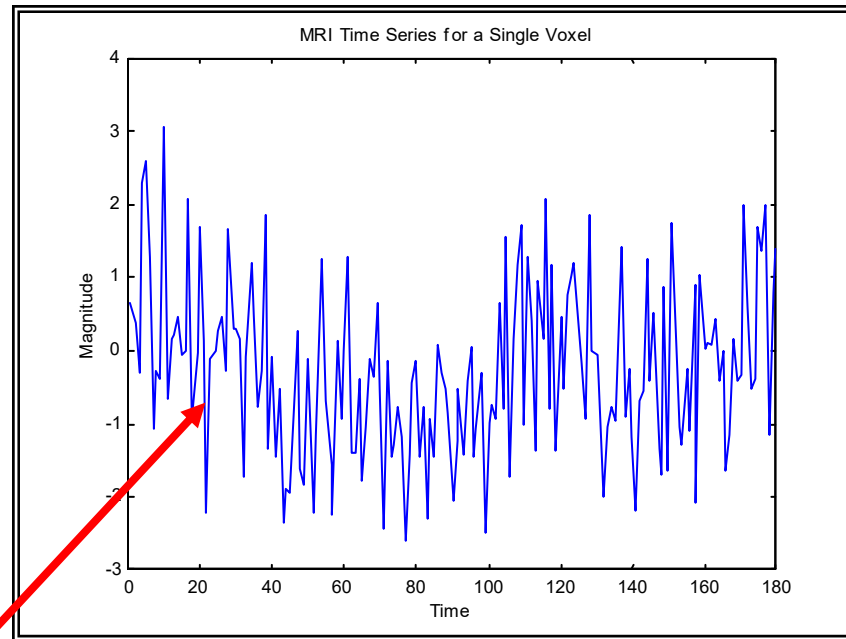
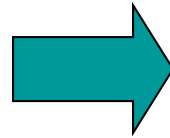
- **Spectroscopy**
  - Vis/NIR of an apple



HALOGEN-TUNGSTEN 250W



# In Neuroimaging

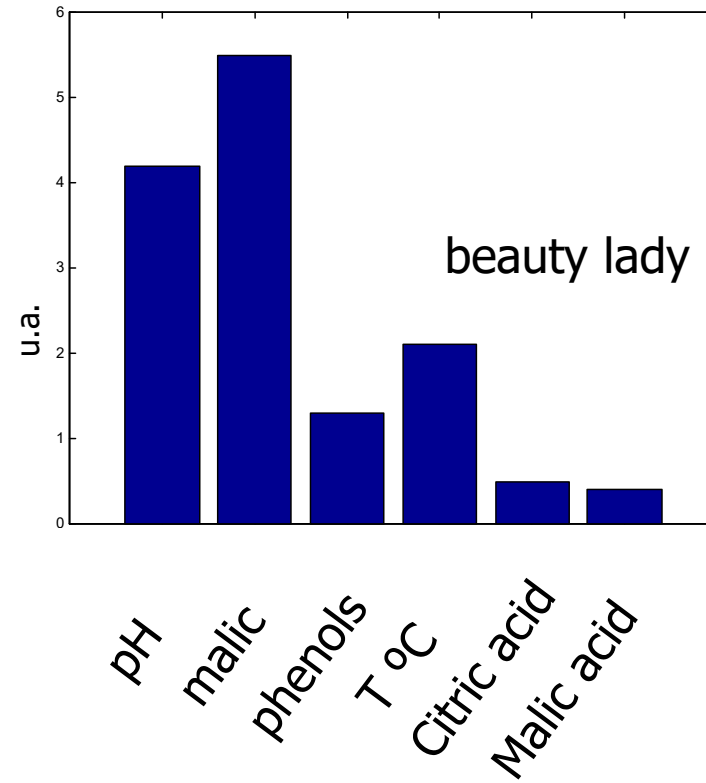
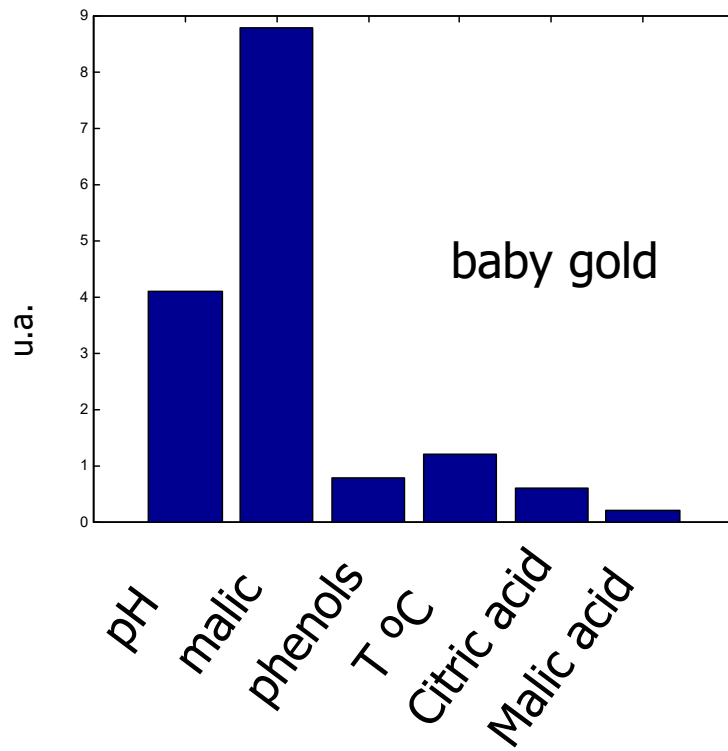


one observation



# Multidimensional Instruments

- **Sensors Array**



# Illustrative Data Set: Food Consumption in European Countries

To illustrate these concepts, we take an easy-to-understand example involving food.

Data on food preferences in 16 different European countries are considered, involving the consumption patterns for 18 different food groups.



Look at the table on the following page. Can you tell anything from the raw numbers? Of course not. No one could.

# Data Table: Food Consumption in European Countries

Table 1.1: The relative consumption of 20 food products across 16 European countries. Each entry shows the percentage of households that normally use each food item.

Primary ID	ONAM	Gr Coffe	Inst Coffe	Tea	Sweetner	Biscuits	Pa Soup	Ti Soup	In Potat	Fro Fish	Fro Veg	Apples	Oranges	Ti Fruit	Jam	Garlic	Butter	Margarine	Olive Oil	Yoghurt	Crisp Bread
1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
2	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
3	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
4	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
5	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
6	Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
7	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
8	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
9	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
10	Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
11	Sweden	97	13	93	31	43	43	39	54	45	56	78	53	75	9	68	32	48	2	93	
12	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
13	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	83	94	28	2	62
14	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17	64	
15	Spain	70	40	40	62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13	
16	Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

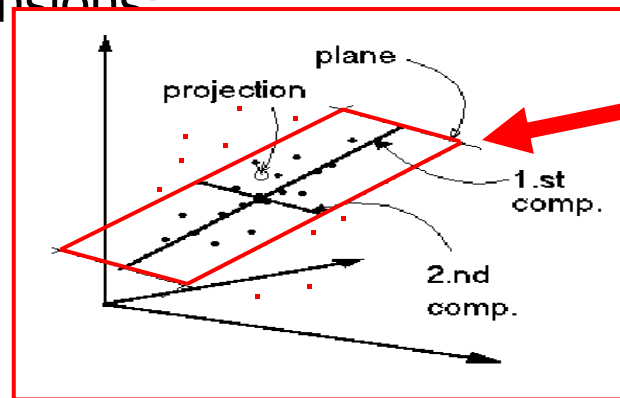
Note that MVA can handle up to 10-20% missing data

Courtesy of Umetrics corp.

# Score Plot

The MVA software generates two main types of plots to represent the data: *Score* plots and *Loadings* plots.

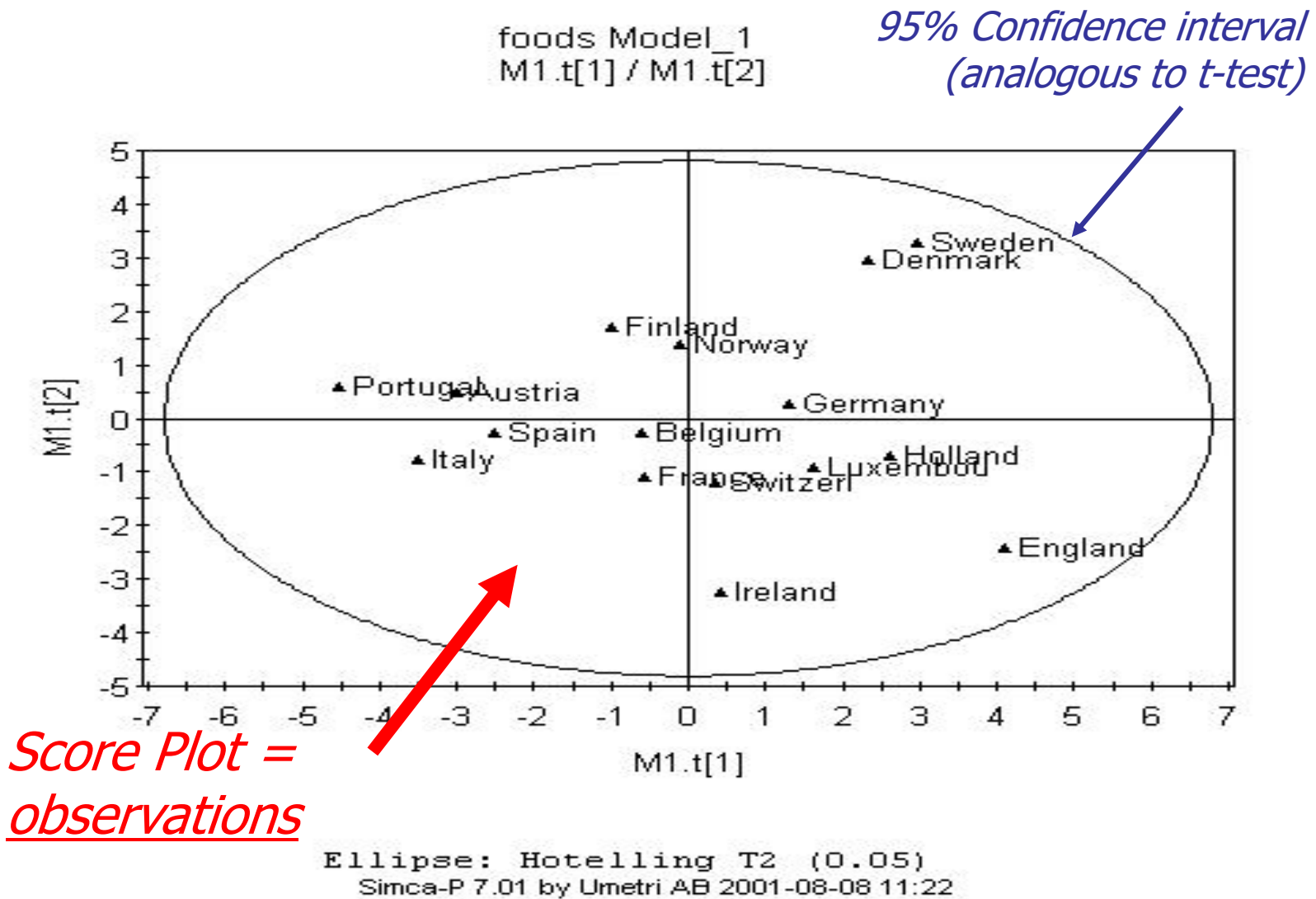
The first of these, the Score plot, shows all the original data points (observations) in a new set of coordinates or *components*. Each score is the value of that data point on one of the *new* component dimensions:



The *Score Plot* is the projection of the original data points onto a plane defined by two new *components*.

A score plot shows how the observations are arranged in the new component space. The score plot for the food data is shown on the next page. Note how similar countries cluster together...

# Score Plot for Food Example



# Loadings Plot

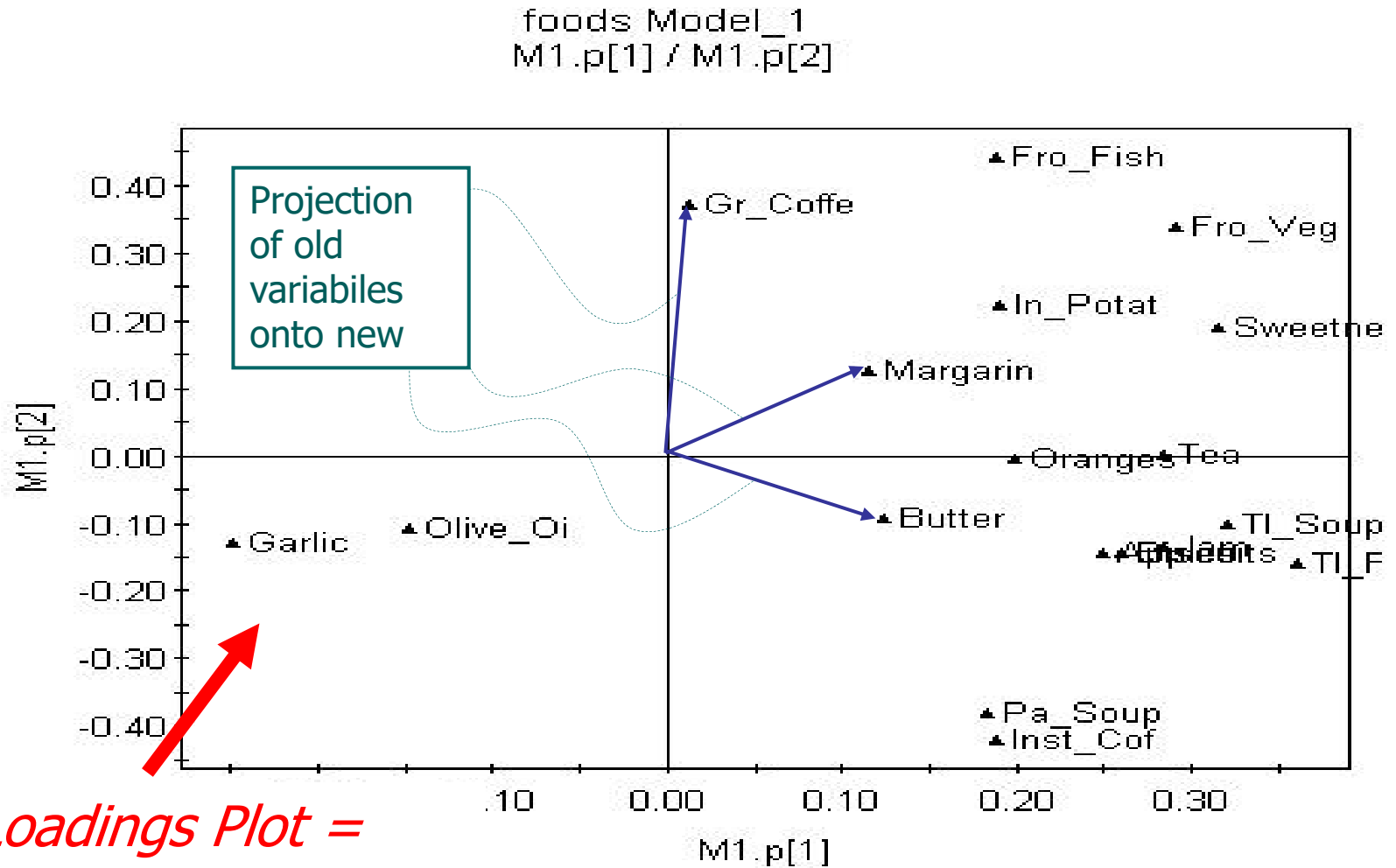
The second type of data plot generated by the MVA software is the *Loadings* plots. This is the equivalent to the score plot, only from the point of view of the original *variables*.

Each component has a set of *loadings* or weights, which express the projection of each original variable onto each new component.

Loadings show how strongly each variable is associated with each new component. The loadings plot for the food example is shown on the next page. The further from the origin, the more significant the correlation.

Note that the *quadrants* are the same on each type of plot. Sweden and Denmark are in the top-right corner; so are frozen fish and vegetables. Using both plots, variables and observations can be correlated with one another.

# Use of loadings (illustration)



*Loadings Plot =  
variables*

# To MVA, Data Overload is Good!

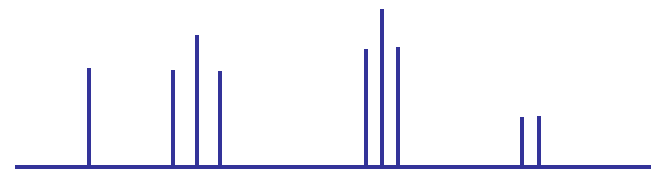
One great advantage of MVA is that the more data are available, the less noise matters (assuming that the noise is normally distributed). This is one of the reasons MVA is used to mine huge amounts of data.

This is analogous to NMR measurements in a laboratory. The more trials there are, the clearer the spectrum becomes:



Looks random

After  
1500  
trials



Not random at all  
(+ve and -ve noise  
cancels out)



# Multivariate Analysis: Benefits

What is the point of doing MVA?

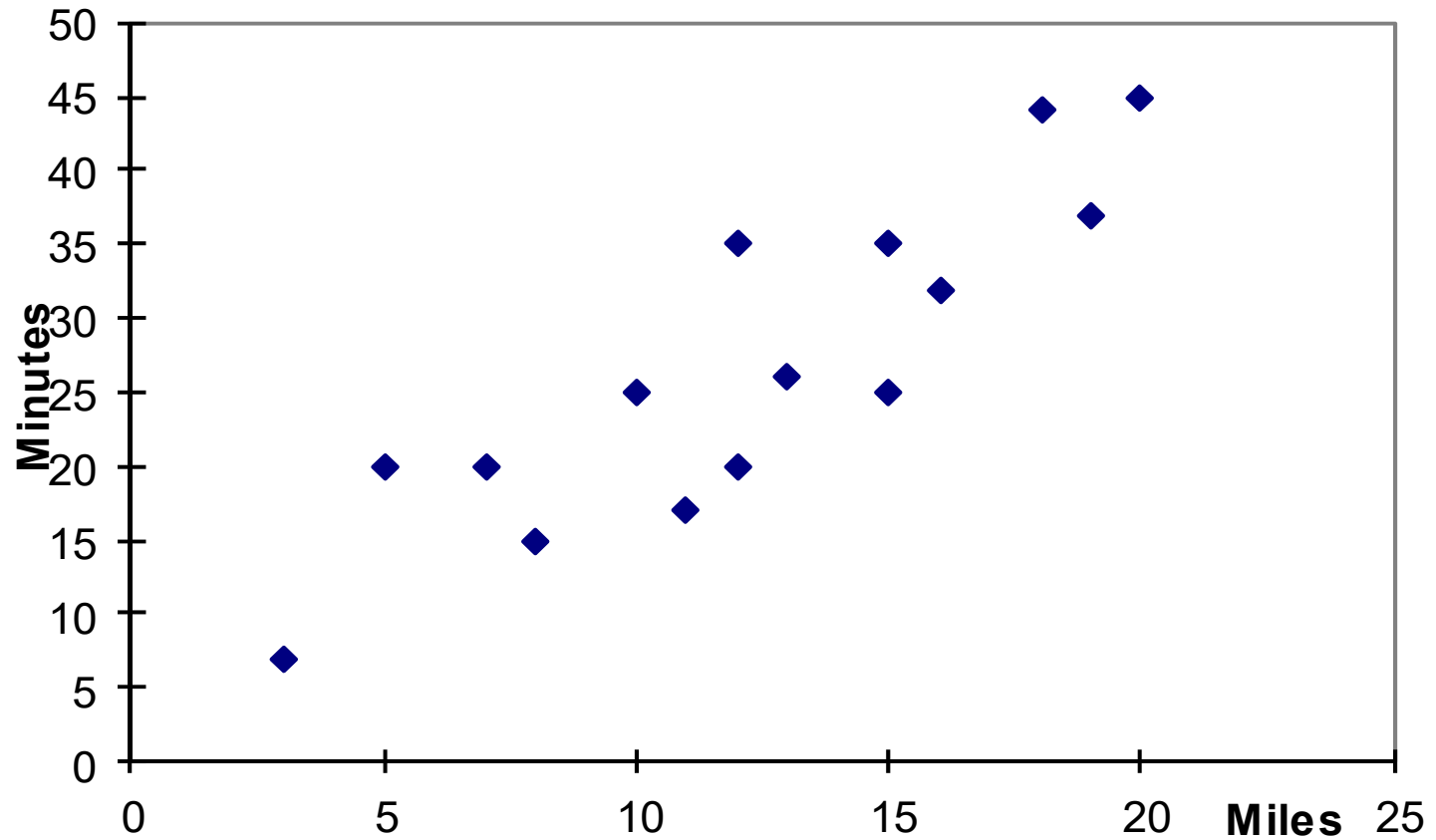
The first potential benefit is to explore the inter-relationships between different process variables. It is well known that simply creating a model can provide insight in the process itself (“Learn by modelling”).

Once a representative model has been created, the engineer can perform “what if?” exercises without affecting the real process. This is a low-cost way to investigate options.

Some important parameters, like final product quality, cannot be measured in real time. They can, however, be inferred from other variables that are measured on-line. When incorporated in the process control system, this inferential controller or “soft sensor” can greatly improve process performance.

# Statistics

## Scatterplot



# What is Statistics?

**Statistics:** The science of collecting, describing, and interpreting data.

Two areas of statistics:

**Descriptive Statistics:** collection, presentation, and description of sample data.

**Inferential Statistics:** making decisions and drawing conclusions about populations.

*Example:* A recent study examined the math and verbal SAT scores of high school seniors across the country. Which of the following statements are descriptive in nature and which are inferential.

- The mean math SAT score was 492.
- The mean verbal SAT score was 475.
- Students in the Northeast scored higher in math but lower in verbal.
- 80% of all students taking the exam were headed for college.
- 32% of the students scored above 610 on the verbal SAT.
- The math SAT scores are higher than they were 10 years ago.

# Introduction to Basic Terms

**Population:** A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

**Sample:** A subset of the population.

**Variable:** A characteristic about each individual element of a population or sample.

**Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

**Data (plural):** The set of values collected for the variable from each of the elements belonging to the sample.

**Experiment:** A planned activity whose results yield a set of data.

**Parameter:** A numerical value summarizing all the data of an entire population.

**Statistic:** A numerical value summarizing the sample data.

*Example:* A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the "age" of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *experiment* would be the method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

The *parameter* of interest is the "average" age of all faculty at the college.

The *statistic* is the "average" age for all faculty in the sample.

Two kinds of variables:

**Qualitative, or Attribute, or Categorical, Variable:** A variable that categorizes or describes an element of a population.

*Note:* Arithmetic operations, such as addition and averaging, are *not* meaningful for data resulting from a qualitative variable.

**Quantitative, or Numerical, Variable:** A variable that quantifies an element of a population.

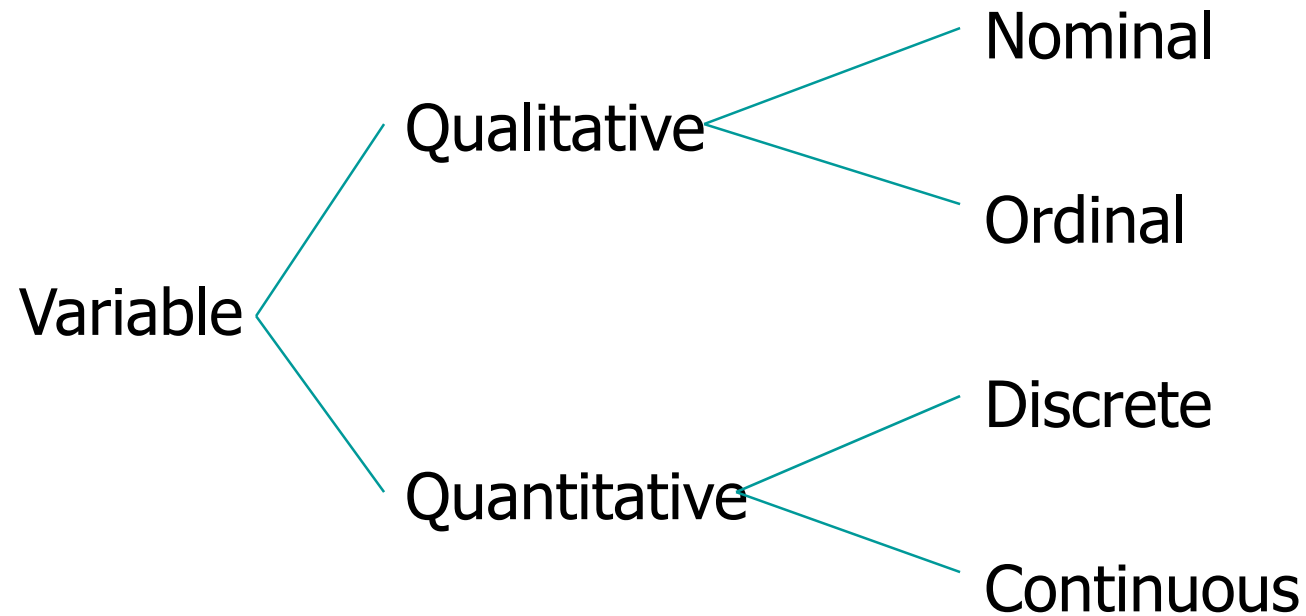
*Note:* Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.



*Example:* Identify each of the following examples as attribute (qualitative) or numerical (quantitative) variables.

1. The residence hall for each student in a statistics class.  
(Attribute)
2. The amount of gasoline pumped by the next 10 customers at the local Unimart. (Numerical)
3. The amount of radon in the basement of each of 25 homes in a new development. (Numerical)
4. The color of the baseball cap worn by each of 20 students.  
(Attribute)
5. The length of time to complete a mathematics homework assignment. (Numerical)
6. The state in which each truck is registered when stopped and inspected at a weigh station. (Attribute)

Qualitative and quantitative variables may be further subdivided:



**Nominal Variable:** A qualitative variable that categorizes (or describes, or names) an element of a population.

**Ordinal Variable:** A qualitative variable that incorporates an ordered position, or ranking.

**Discrete Variable:** A quantitative variable that can assume a countable number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

**Continuous Variable:** A quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

*Note:*

1. In many cases, a discrete and continuous variable may be distinguished by determining whether the variables are related to a count or a measurement.
2. Discrete variables are usually associated with counting. If the variable cannot be further subdivided, it is a clue that you are probably dealing with a discrete variable.
3. Continuous variables are usually associated with measurements. The values of discrete variables are only limited by your ability to measure them.

# Measure and Variability

- No matter what the response variable: there will always be **variability** in the data.
- One of the primary objectives of statistics: measuring and characterizing variability.
- Controlling (or reducing) variability in a manufacturing process: statistical process control.

*Example:* A supplier fills cans of soda marked 12 ounces. How much soda does each can really contain?

- It is very *unlikely* any one can contains exactly 12 ounces of soda.
- There is variability in any process.
- Some cans contain a little more than 12 ounces, and some cans contain a little less.
- On the average, there are 12 ounces in each can.
- The supplier hopes there is little variability in the process, that most cans contain *close* to 12 ounces of soda.

# Data Collection

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.

**Biased Sampling Method:** A sampling method that produces data which systematically differs from the sampled population. An **unbiased sampling method** is one that is not biased.

Sampling methods that often result in biased samples:

1. **Convenience sample:** sample selected from elements of a population that are easily accessible.
2. **Volunteer sample:** sample collected from those elements of the population which chose to contribute the needed information on their own initiative.



## Process of data collection:

1. Define the objectives of the survey or experiment.

*Example:* Estimate the average life of an electronic component.

2. Define the variable and population of interest.

*Example:* Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

Methods used to collect data:

**Experiment:** The investigator controls or modifies the environment and observes the effect on the variable under study.

**Survey:** Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

**Census:** A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

**Sampling Frame:** A list of the elements belonging to the population from which the sample will be drawn.

*Note:* It is important that the sampling frame be representative of the population.

**Sample Design:** The process of selecting sample elements from the sampling frame.

*Note:* There are many different types of sample designs. Usually they all fit into two categories: judgment samples and probability samples.

**Judgment Samples:** Samples that are selected on the basis of being “typical.”

Items are selected that are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector’s judgment.

**Probability Samples:** Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

**Random Samples:** A sample selected in such a way that every element in the population has a equal probability of being chosen. Equivalently, all samples of size  $n$  have an equal chance of being selected. Random samples are obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.

*Note:*

1. Inherent in the concept of randomness: the next result (or occurrence) is not predictable.
2. Proper procedure for selecting a random sample: use a random number generator or a table of random numbers.

*Example:* An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

**Systematic Sample:** A sample in which every  $k$ th item of the sampling frame is selected, starting from the first element which is randomly selected from the first  $k$  elements.

*Note:* The systematic technique is easy to execute. However, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature. In these situations the results may not approximate a simple random sample.

**Stratified Random Sample:** A sample obtained by stratifying the sampling frame and then selecting a fixed number of items from each of the strata by means of a simple random sampling technique.

**Proportional Sample (or Quota Sample):** A sample obtained by stratifying the sampling frame and then selecting a number of items in proportion to the size of the strata (or by quota) from each strata by means of a simple random sampling technique.

**Cluster Sample:** A sample obtained by stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.



# Numerical Presentation

A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data. Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70

B: 40, 50, 60

The mean of both two data sets is 50. But, the distance of the observations from the mean in data set A is larger than in the data set B. Thus, the mean of data set B is a better representation of the data set than is the case for set A.

# Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are mean, median, mode, geometric mean etc.

Mean: Summing up all the observation and dividing by number of observations.  
Mean of 20, 30, 40 is  $(20+30+40)/3 = 30$ .

Notation : Let  $x_1, x_2, \dots, x_n$  are  $n$  observations of a variable  $x$ . Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

# Methods of Center Measurement

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case,  $(5 + 6) / 2 = 5.5$ .

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

# Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is  $(20+30+40+990)/4 = 270$ . The median of these four observations is  $(30+40)/2 = 35$ . Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

# Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range, variance, standard deviation, interquartile range, coefficient of variation etc.*

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is  $(100-2)=98$ . It's a crude measure of variability.

# Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the  $n$  observations  $x_1, x_2, \dots, x_n$  is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

Variance of 5, 7, 3? Mean is  $(5+7+3)/3 = 5$  and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

# Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the  $((n+1)/4)q^{\text{th}}$  observation of the data, where  $q$  is the desired quartile and  $n$  is the number of observations of data.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25<sup>th</sup> and 50<sup>th</sup> percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

# Methods of Variability Measurement

In the following example  $Q1 = ((15+1)/4)1 = 4^{\text{th}}$  observation of the data. The 4<sup>th</sup> observation is 11. So  $Q1$  of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

$Q1$

$Q2$

$Q3$

The first quartile is  $Q1=11$ . The second quartile is  $Q2=40$  (This is also the Median.) The third quartile is  $Q3=61$ .

Inter-quartile Range: Difference between  $Q3$  and  $Q1$ . Inter-quartile range of the previous example is  $61 - 40 = 21$ . The middle half of the ordered data lie between 40 and 61.



# Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25<sup>th</sup> percentile is the Q1, 50<sup>th</sup> percentile is the Median (Q2) and the 75<sup>th</sup> percentile of the data is Q3.

In notations, percentiles of a data is the  $((n+1)/100)p$  th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$

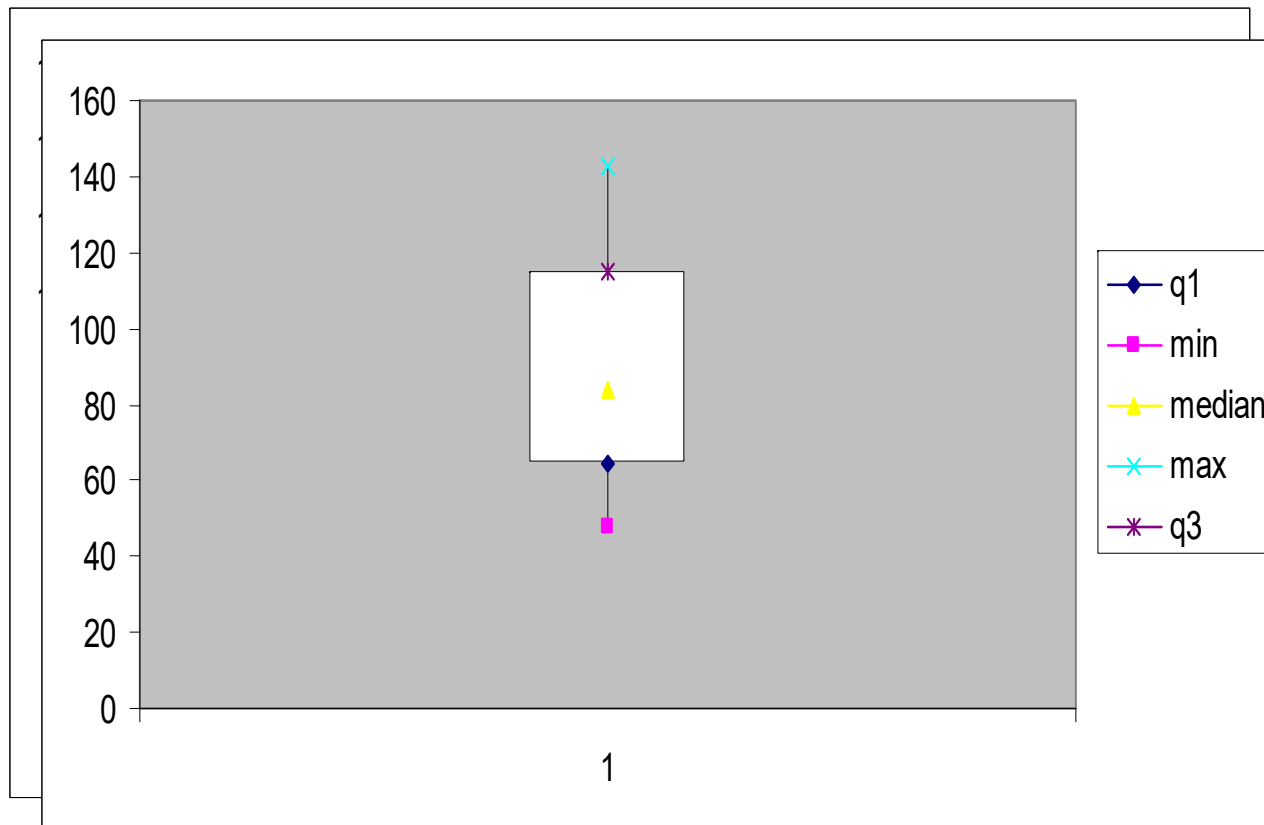
# Five Number Summary

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), The median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

# Boxplot

Distribution of Age in Month



# Choosing a Summary

The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.

In real life we can't always expect symmetry of the data. It's a common practice to include number of observations ( $n$ ), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like  $Q_1$ ,  $Q_3$ , Coefficient of variation if it is considered to be important for describing data.

# Shape of Data

- Shape of data is measured by
  - Skewness
  - Kurtosis

# Skewness

- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let  $x_1, x_2, \dots, x_n$  be  $n$  observations. Then,

$$\text{Skewness} = \frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

## Kurtosis

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

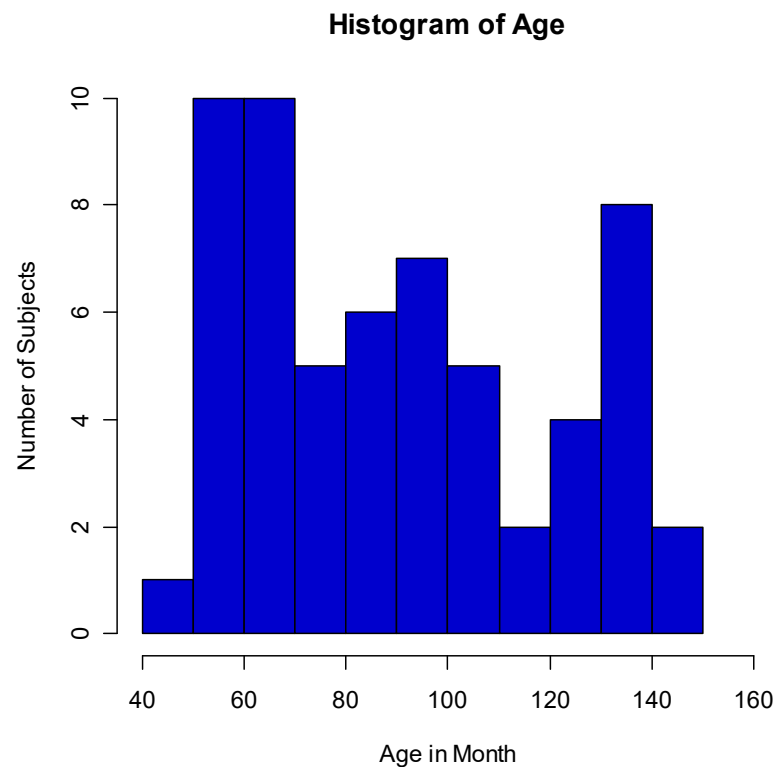
Let  $x_1, x_2, \dots, x_n$  be  $n$  observations. Then,

$$\text{Kurtosis} = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left( \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

# Summary of the Variable 'Age' in the given data set

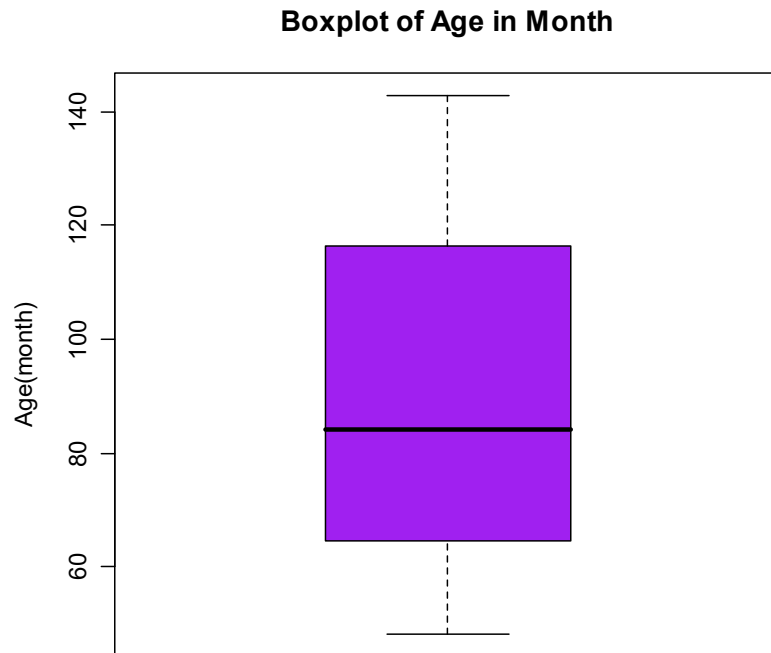
Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

---





# Summary of the Variable 'Age' in the given data set



# Microsoft Excel

A Spreadsheet Application. It features calculation, graphing tools, pivot tables and a macro programming language called VBA (Visual Basic for Applications).

There are many versions of MS-Excel. Excel XP, Excel 2003, Excel 2007 are capable of performing a number of statistical analyses.

Starting MS Excel: Double click on the Microsoft Excel icon on the desktop or Click on Start --> Programs --> Microsoft Excel.

Worksheet: Consists of a multiple grid of cells with numbered rows down the page and alphabetically-tilted columns across the page. Each cell is referenced by its coordinates. For example, A3 is used to refer to the cell in column A and row 3. B10:B20 is used to refer to the range of cells in column B and rows 10 through 20.

# Microsoft Excel

Opening a document: File → Open (From a existing workbook). Change the directory area or drive to look for file in other locations.

Creating a new workbook: File→New→Blank Document

Saving a File: File→Save

Selecting more than one cell: Click on a cell e.g. A1), then hold the Shift key and click on another (e.g. D4) to select cells between and A1 and D4 or Click on a cell and drag the mouse across the desired range.

Creating Formulas: 1. Click the cell that you want to enter the formula, 2. Type = (an equal sign), 3. Click the Function Button, 4. Select the formula you want and step through the on-screen instructions.

# Microsoft Excel

Entering Date and Time: Dates are stored as MM/DD/YYYY. No need to enter in that format. For example, Excel will recognize jan 9 or jan-9 as 1/9/2007 and jan 9, 1999 as 1/9/1999. To enter today's date, press Ctrl and ; together. Use a or p to indicate am or pm. For example, 8:30 p is interpreted as 8:30 pm. To enter current time, press Ctrl and : together.

Copy and Paste all cells in a Sheet: Ctrl+A for selecting, Ctrl +C for copying and Ctrl+V for Pasting.

Sorting: Data → Sort → Sort By ...

Descriptive Statistics and other Statistical methods: Tools → Data Analysis → Statistical method. If Data Analysis is not available then click on Tools → Add-Ins and then select Analysis ToolPack and Analysis toolPack-Vba

# Microsoft Excel

Statistical and Mathematical Function: Start with '=' sign and then select function from function wizard  $f_x$ .

Inserting a Chart: Click on Chart Wizard (or Insert→Chart), select chart, give, Input data range, Update the Chart options, and Select output range/ Worksheet.

Importing Data in Excel: File →open →FileType →Click on File→ Choose Option ( Delimited/Fixed Width) →Choose Options (Tab/ Semicolon/ Comma/ Space/ Other) → Finish.

Limitations: Excel uses algorithms that are vulnerable to rounding and truncation errors and may produce inaccurate results in extreme cases.

# Statistical Inference

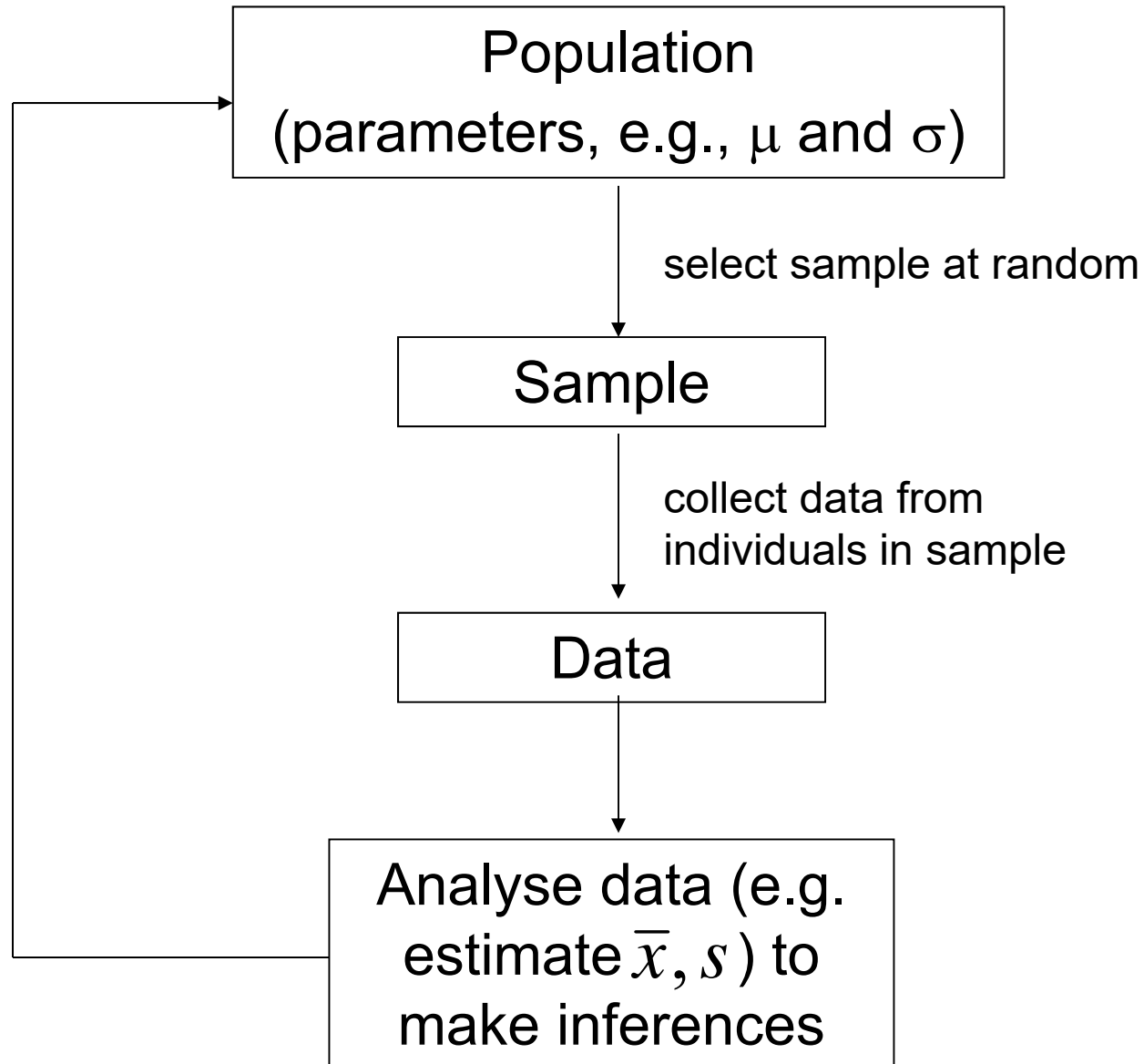
- **Statistical Inference** – the process of drawing conclusions about a population based on information in a sample
- Unlikely to see this published...

“In our study of a new antihypertensive drug we found an effective 10% reduction in blood pressure for those on the new therapy. However, the effects seen are only specific to the subjects in our study. We cannot say this drug will work for hypertensive people in general”.

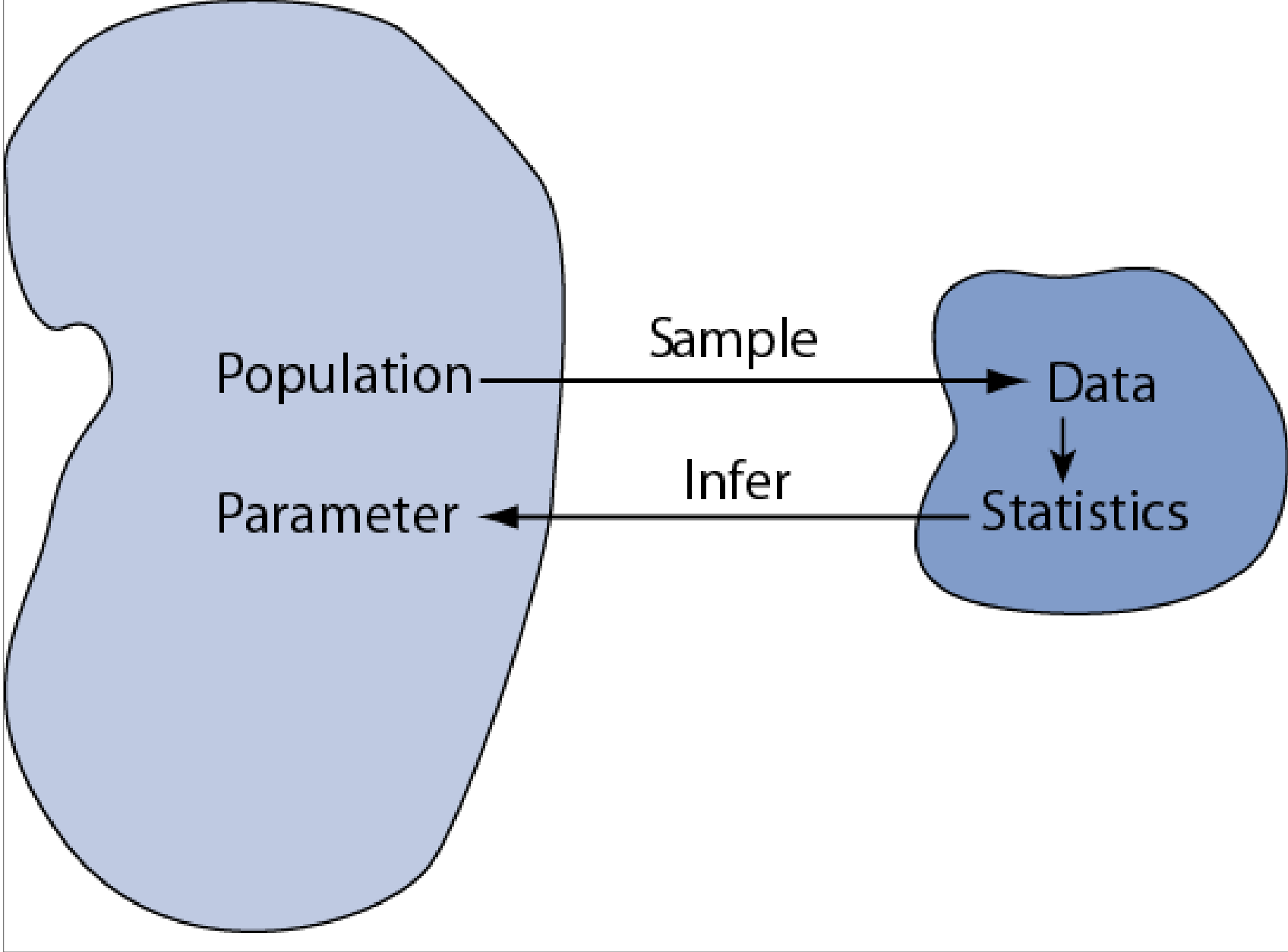
## Describing a population

- Characteristics of a population, e.g. the **population mean  $\mu$**  and the **population standard deviation  $\sigma$**  are never known exactly
- Sample characteristics, e.g.  $\bar{x}$  and  $s$  are **estimates** of population characteristics  $\mu$  and  $\sigma$
- A sample characteristic, e.g.  $\bar{x}$ , is called a **statistic** and a population characteristic, e.g.  $\mu$  is called a **parameter**

# Statistical Inference







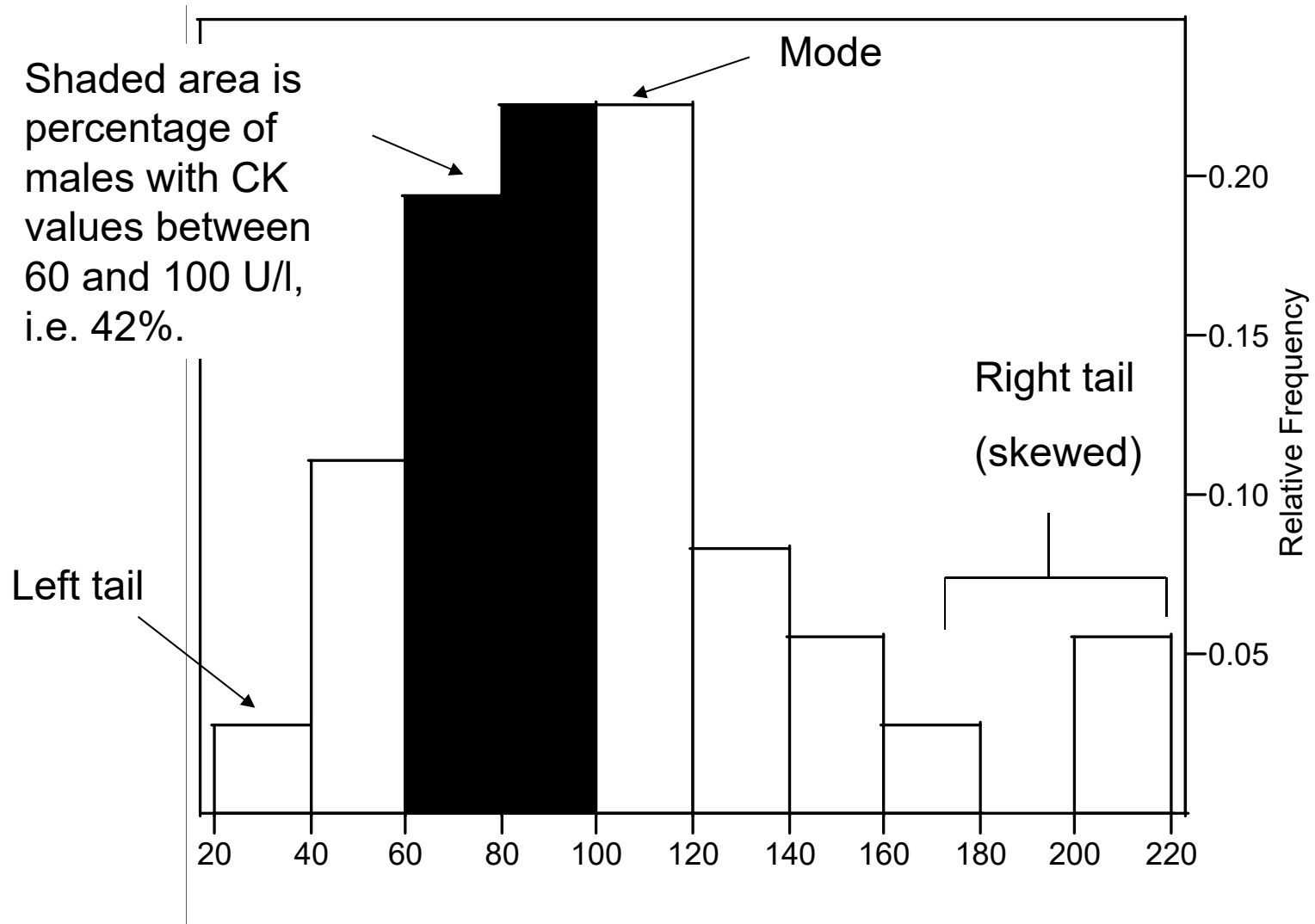
# Distributions

- As sample size increases, histogram class widths can be narrowed such that the histogram eventually becomes a smooth curve
- The population histogram of a random variable is referred to as the **distribution** of the random variable, i.e. it shows how the population is distributed across the number line

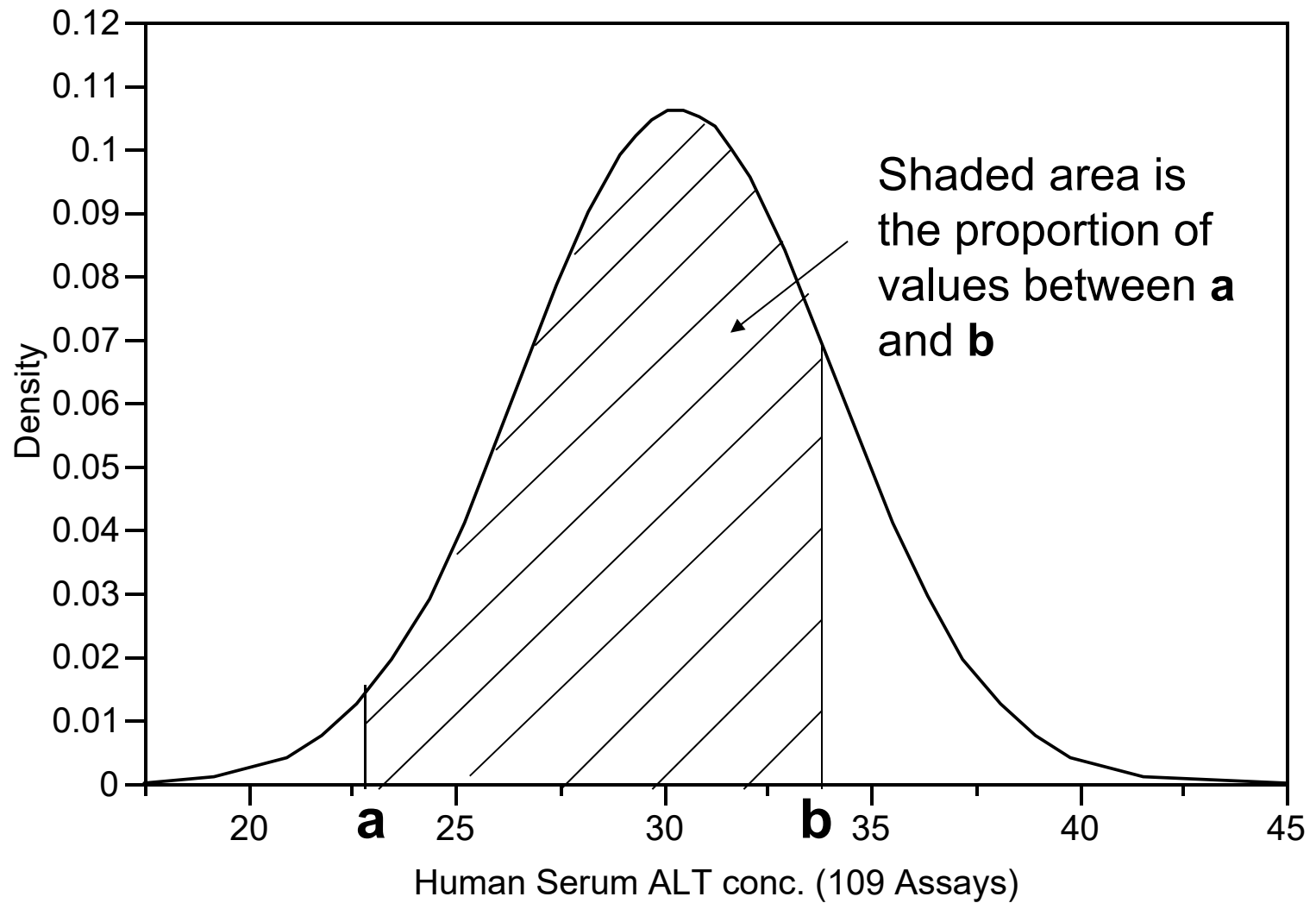
# Density curve

- A smooth curve representing a relative frequency distribution is called a **density** curve
- The area under the density curve between any two points **a** and **b** is the proportion of values between **a** and **b**.

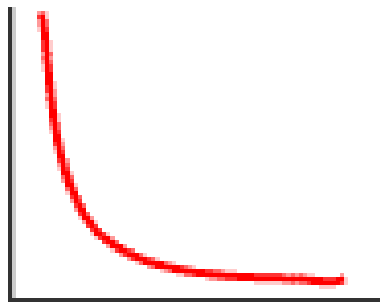
# Sample Relative Frequency Distribution



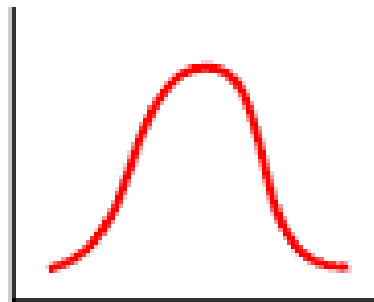
# Population Relative Frequency Distribution (Density)



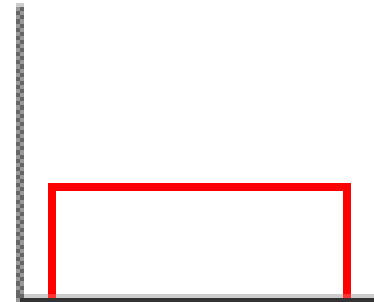
# Distribution Shapes



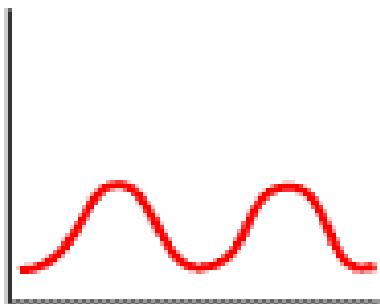
**J-shaped**



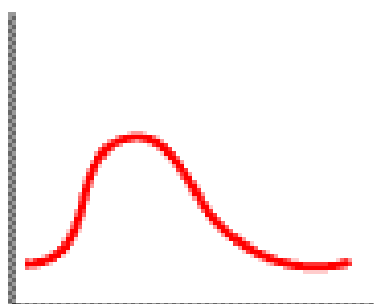
**Normal**



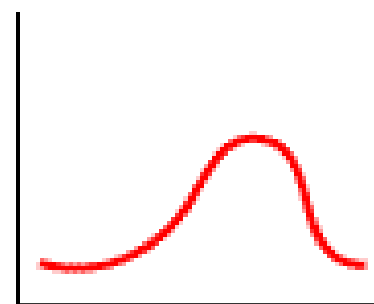
**Rectangular**



**Bimodal**



**Positive (right) skew**



**Negative (left) skew**

# The Normal Distribution

- The **Normal distribution** is considered to be the most important distribution in statistics
- It occurs in “nature” from processes consisting of a very large number of elements acting in an **additive** manner
- However, it would be very difficult to use this argument to assume normality of your data
  - Later, we will see exactly why the Normal is so important in statistics

# Normal Distribution

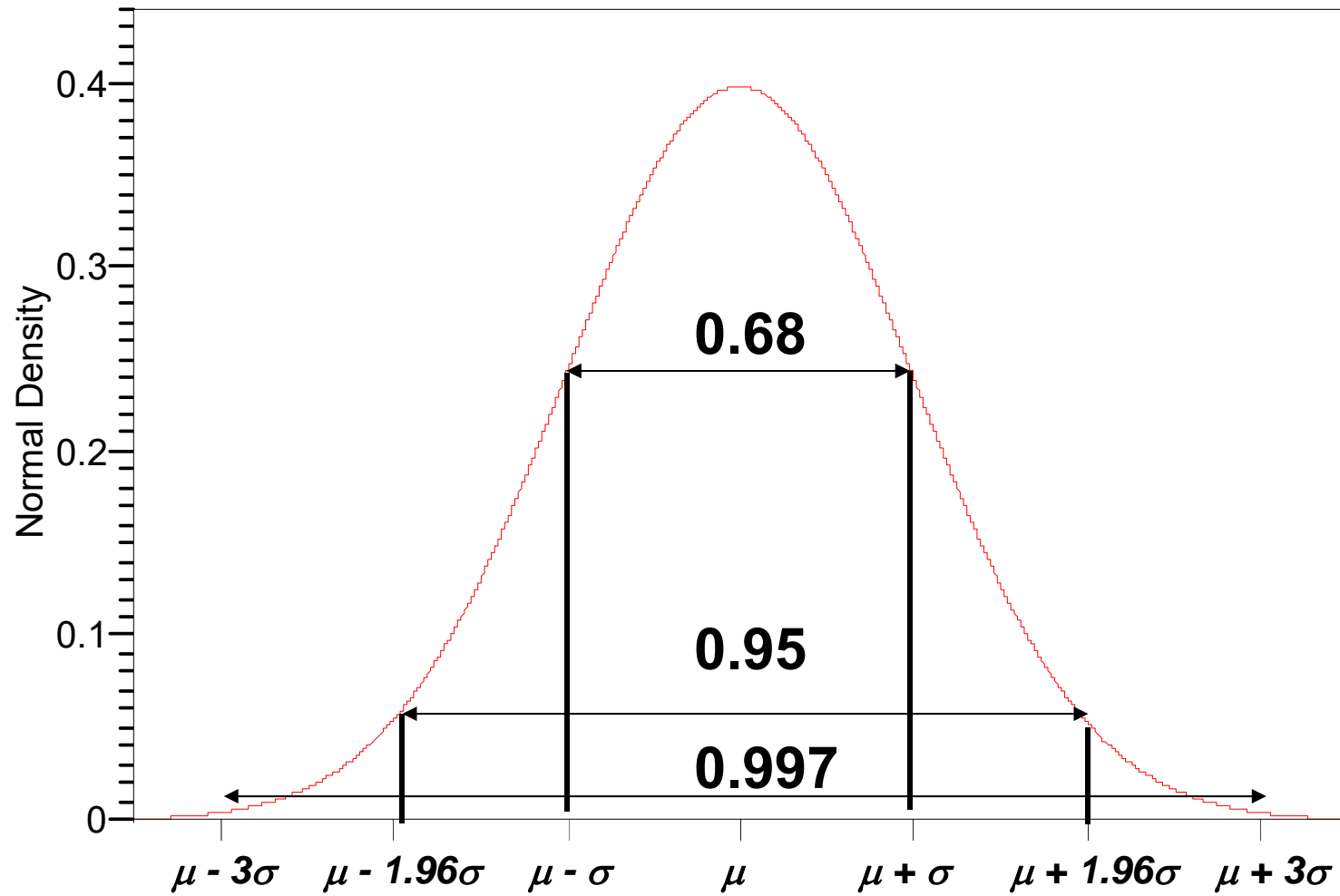
- Closely related is the **log-normal** distribution, based on factors acting **multiplicatively**. This distribution is right-skewed.
  - Note: The logarithm of the data is thus normal.
- The log-transformation of data is very common, mostly to eliminate skew in data



# Properties of the Normal Distribution

- The Normal distribution has a symmetric bell-shaped density curve
- Characterised by two parameters, i.e. the mean  $\mu$ , and standard deviation  $\sigma$ 
  - 68% of data lie within  $1\sigma$  of the mean  $\mu$
  - 95% of data lie within  $2\sigma$  of the mean  $\mu$
  - 99.7% of data lie within  $3\sigma$  of the mean  $\mu$

# Normal curve




# Probability Density Functions...

- Unlike a discrete random variable which we studied in Chapter 7, a ***continuous random variable*** is one that can assume an **uncountable** number of values.
- → We cannot list the possible values because there is an infinite number of them.
- → Because there is an infinite number of values, the probability of each individual value is virtually 0.

# Point Probabilities are Zero

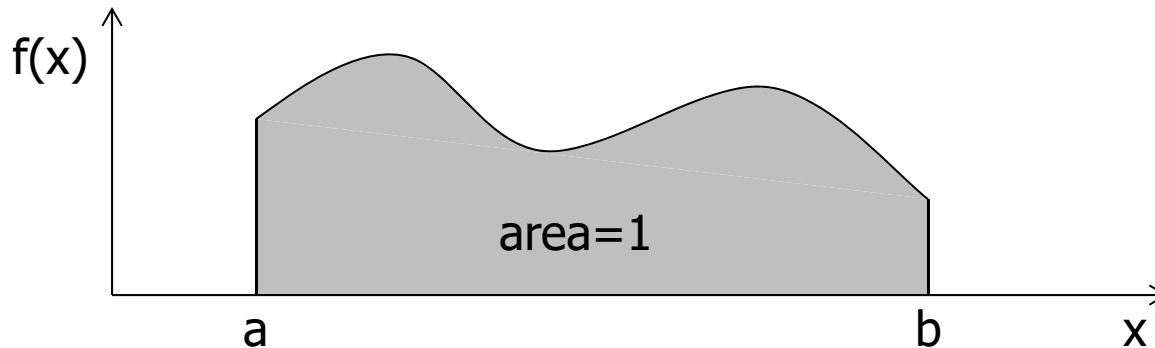
→ Because there is an infinite number of values, the probability of each individual value is virtually 0.

Thus, we can determine the probability of a *range of values* only.

- E.g. with a **discrete** random variable like tossing a die, it is meaningful to talk about  $P(X=5)$ , say.
  - In a **continuous** setting (e.g. with time as a random variable), the probability the random variable of interest, say task length, takes exactly 5 minutes is infinitesimally small, hence  $P(X=5) = 0$ .
    - ***It is meaningful to talk about  $P(X \leq 5)$ .***
- 

# Probability Density Function...

- A function  $f(x)$  is called a ***probability density function*** (over the range  $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ ) if it meets the following requirements:
  - 1)  $f(x) \geq 0$  for all  $\mathbf{x}$  between  $\mathbf{a}$  and  $\mathbf{b}$ , and
  - 2) The total area under the curve between  $\mathbf{a}$  and  $\mathbf{b}$  is 1.0

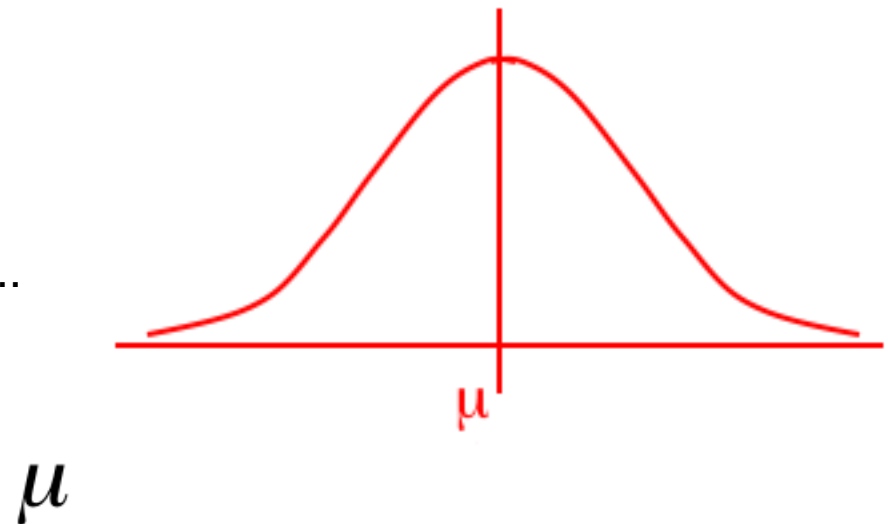


# The Normal Distribution...

- The **normal distribution** is the most important of all probability distributions. The probability density function of a **normal random variable** is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

- It looks like this:
- Bell shaped,
- Symmetrical around the mean ...



# The Normal Distribution...

The normal distribution is fully defined by two parameters:  
its **standard deviation** and **mean**

- **Important things to note:**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

The normal distribution is bell shaped and  
symmetrical about the **mean**

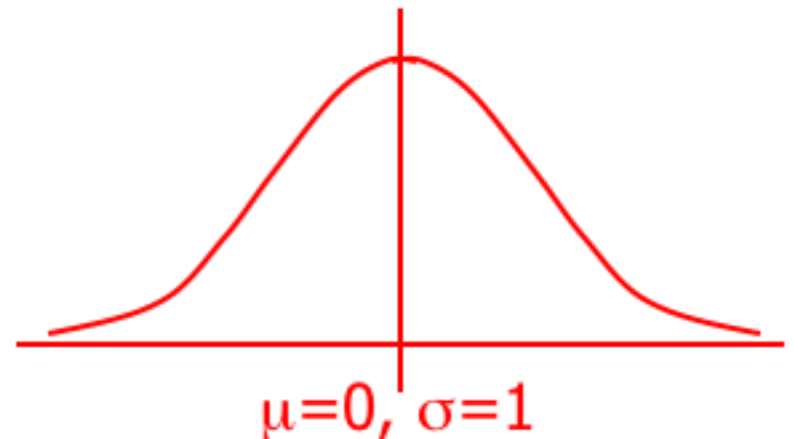
Unlike the range of the uniform distribution ( $a \leq x \leq b$ )  
Normal distributions ***range from minus infinity to plus infinity***

# Standard Normal Distribution...

- A normal distribution whose **mean is zero** and **standard deviation is one** is called the ***standard normal distribution***.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} \quad -\infty < x < \infty$$

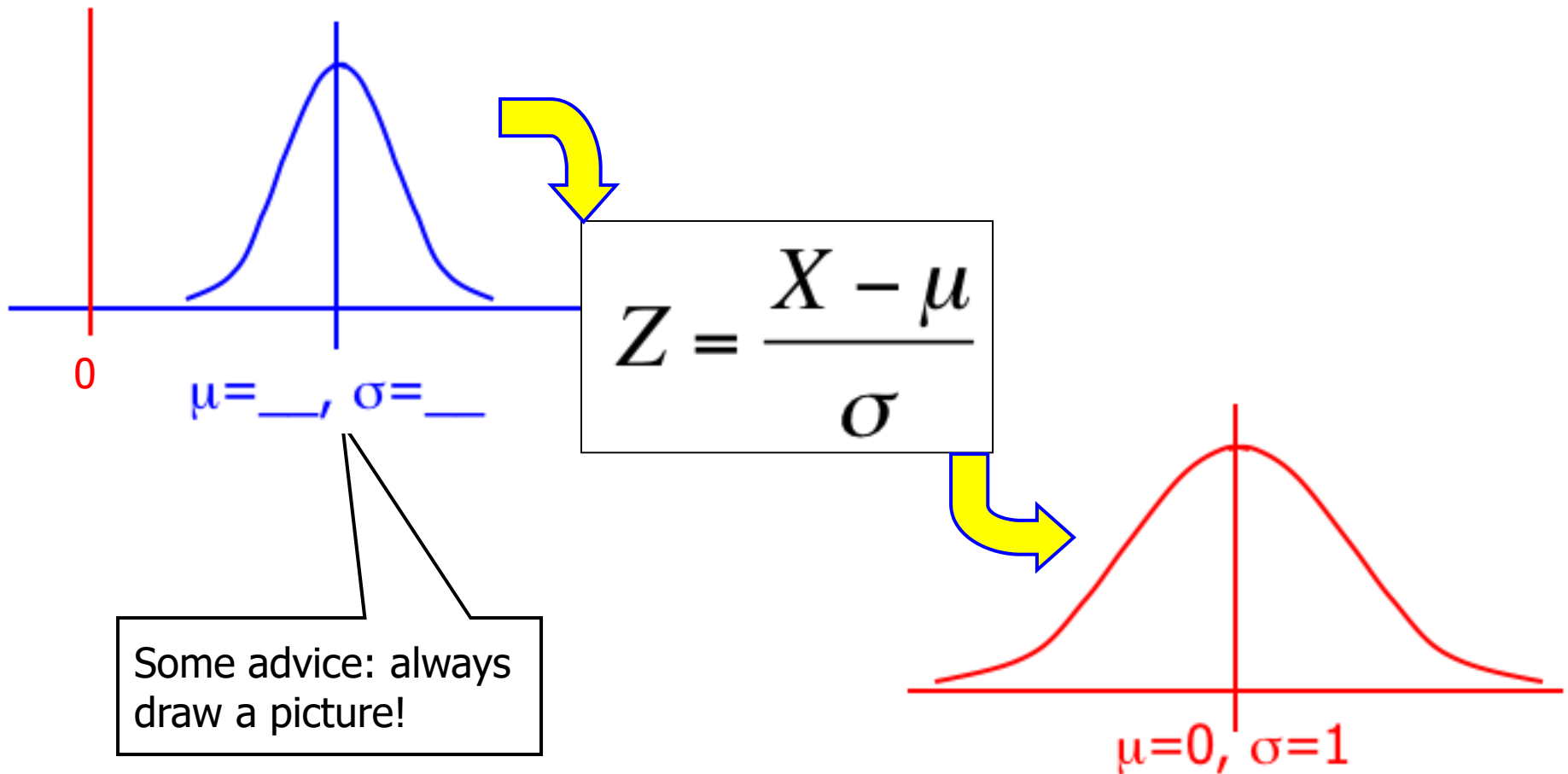
- As we shall see shortly, any normal distribution can be ***converted*** to a standard normal distribution with simple algebra. This makes calculations much easier.





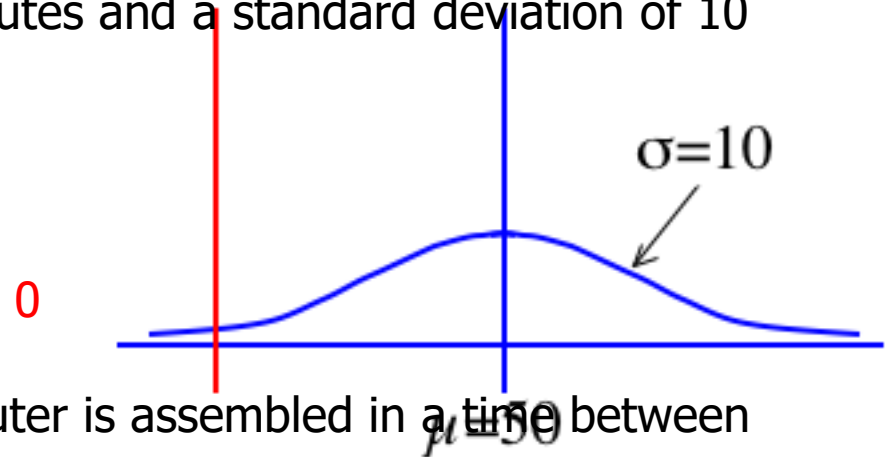
# Calculating Normal Probabilities...

- We can use the following function to convert any normal random variable to a **standard** normal random variable...



# Calculating Normal Probabilities...

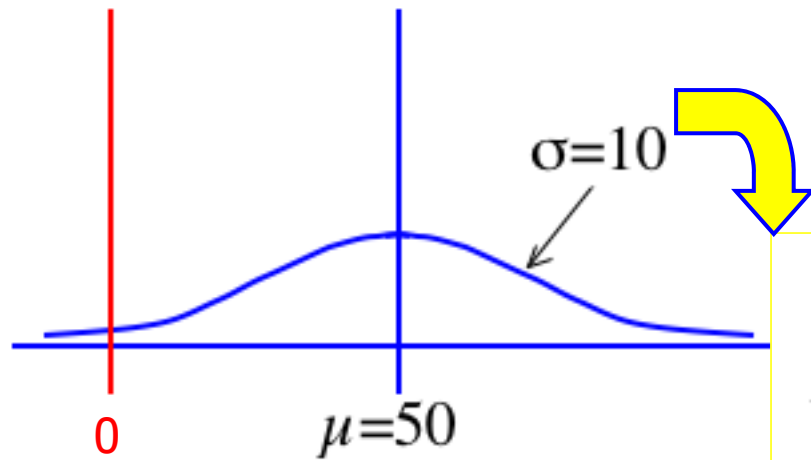
- Example: The time required to build a computer is ***normally distributed*** with a mean of 50 minutes and a standard deviation of 10 minutes:



- What is the probability that a computer is assembled in a time between 45 and 60 minutes?
- Algebraically speaking, what is  **$P(45 < X < 60)$**  ?

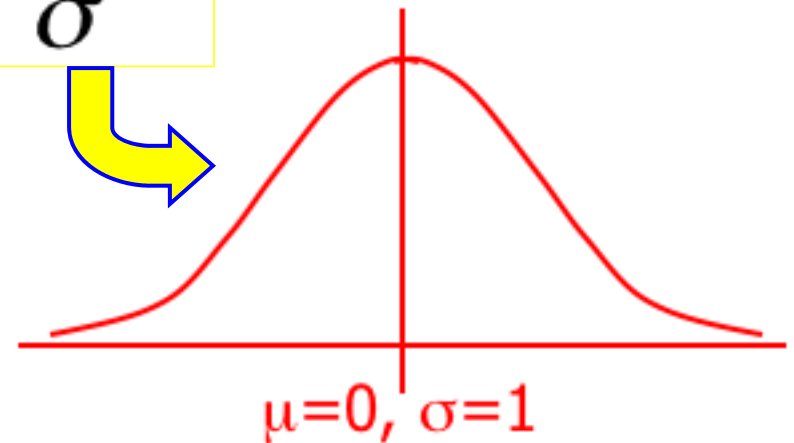
# Calculating Normal Probabilities...

...mean of 50 minutes and a standard deviation of 10 minutes...



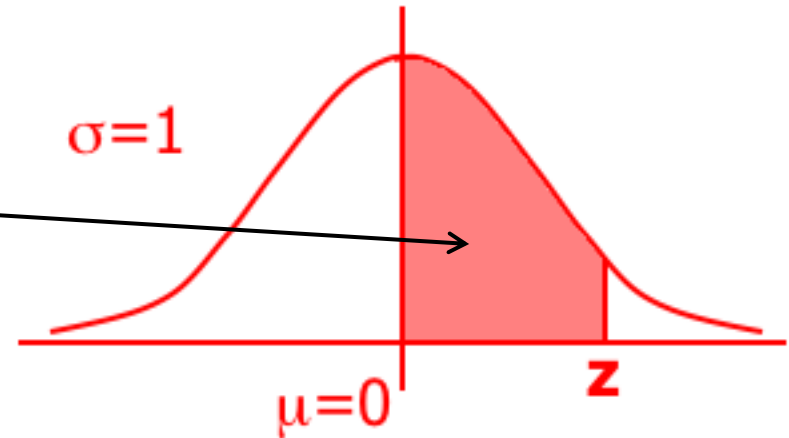
$$Z = \frac{X - \mu}{\sigma}$$

$$P(45 < X < 60) =$$
$$P\left(\frac{45 - 50}{10} < \frac{X - \mu}{\sigma} < \frac{60 - 50}{10}\right) =$$
$$P(-.5 < Z < 1)$$



# Calculating Normal Probabilities...

- We can use [Table 3](#) in
- Appendix B to look-up
- probabilities  **$P(0 < Z < z)$**



- We can break up  **$P(-.5 < Z < 1)$**  into:
  - **$P(-.5 < Z < 0)$**  +  **$P(0 < Z < 1)$**
- The distribution is *symmetric* around zero, so we have:
  - $P(-.5 < Z < 0) =$   **$P(0 < Z < .5)$**
  - Hence:  **$P(-.5 < Z < 1) = P(0 < Z < .5) + P(0 < Z < 1)$**

# Calculating Normal Probabilities...

- How to use [Table](#)...

This table gives probabilities  $P(0 < Z < z)$

First column = integer + first decimal

Top row = second decimal place

$$P(0 < Z < 0.5)$$

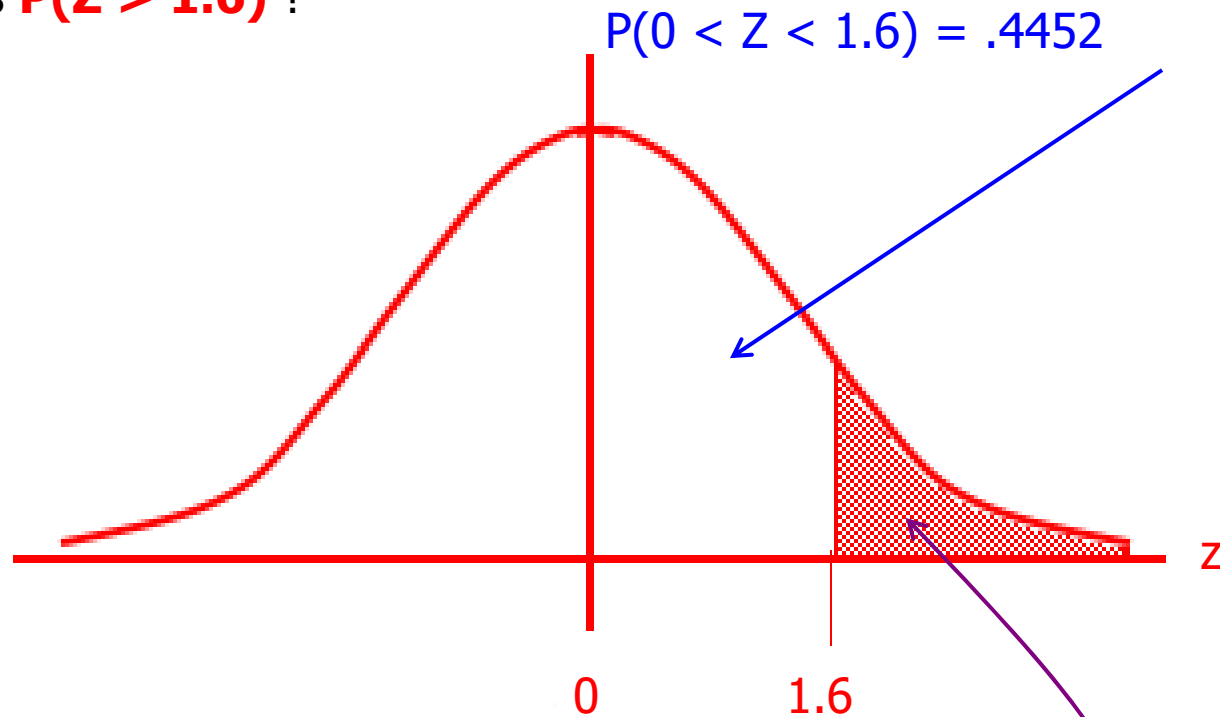
$$P(0 < Z < 1)$$

$$P(-.5 < Z < 1) = .1915 + .3414 = .5328$$

z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	.3840	.3860	.3880	.3899

# Using the Normal Table

- What is  $P(Z > 1.6)$  ?

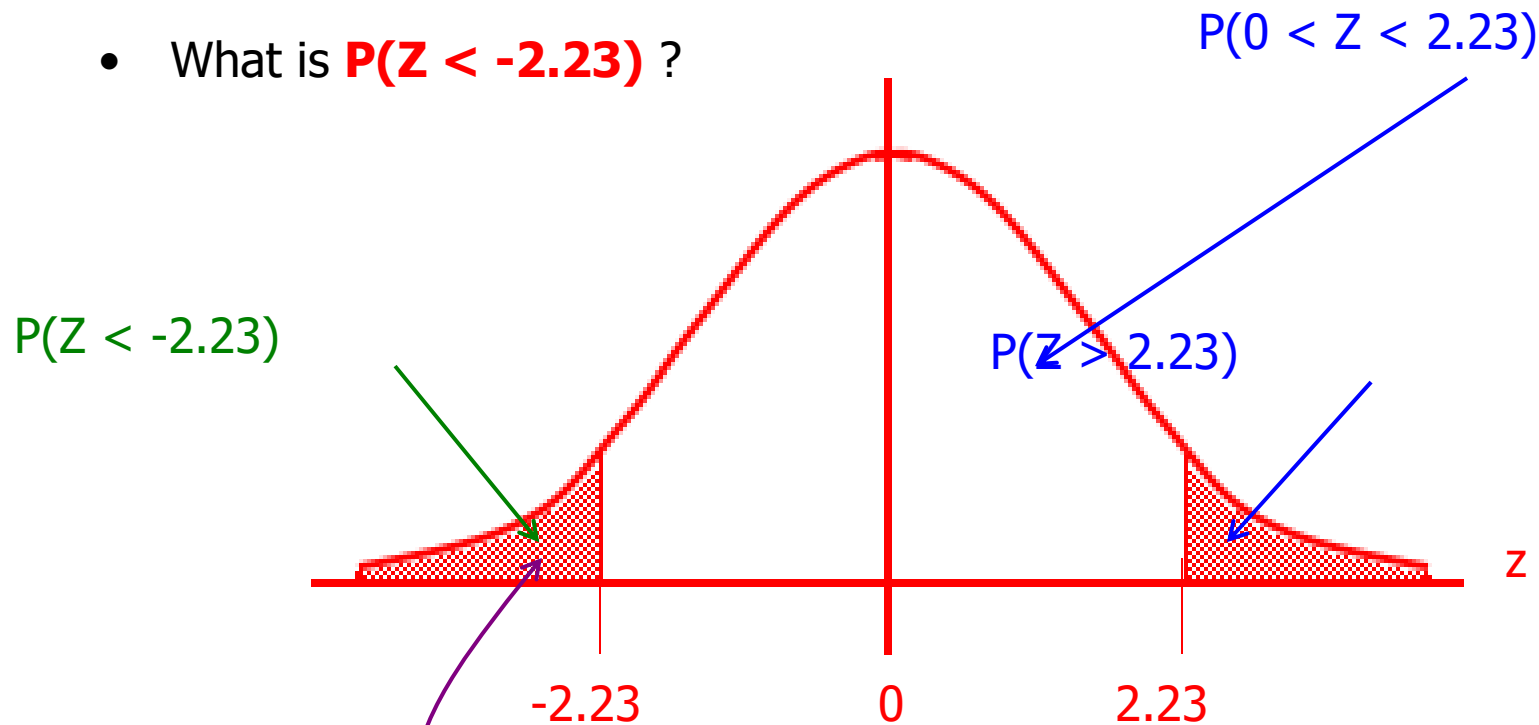


$$P(0 < Z < 1.6) = .4452$$

$$\begin{aligned} P(Z > 1.6) &= .5 - P(0 < Z < 1.6) \\ &= .5 - .4452 \\ &= \mathbf{.0548} \end{aligned}$$

## Using the Normal Table ([Table 3](#))...

- What is  $P(Z < -2.23)$  ?



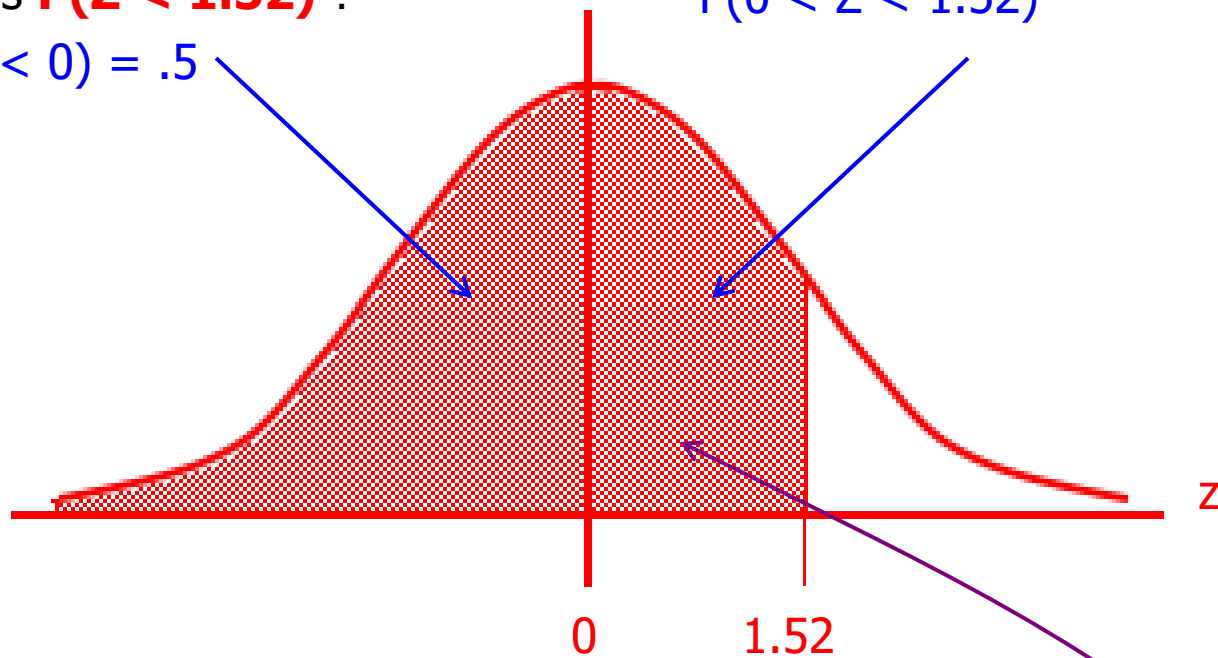
$$\begin{aligned} P(Z < -2.23) &= P(Z > 2.23) \\ &= .5 - P(0 < Z < 2.23) \\ &= \mathbf{.0129} \end{aligned}$$

## Using the Normal Table ([Table 3](#))...

- What is  **$P(Z < 1.52)$**  ?

$$P(Z < 0) = .5$$

$$P(0 < Z < 1.52)$$

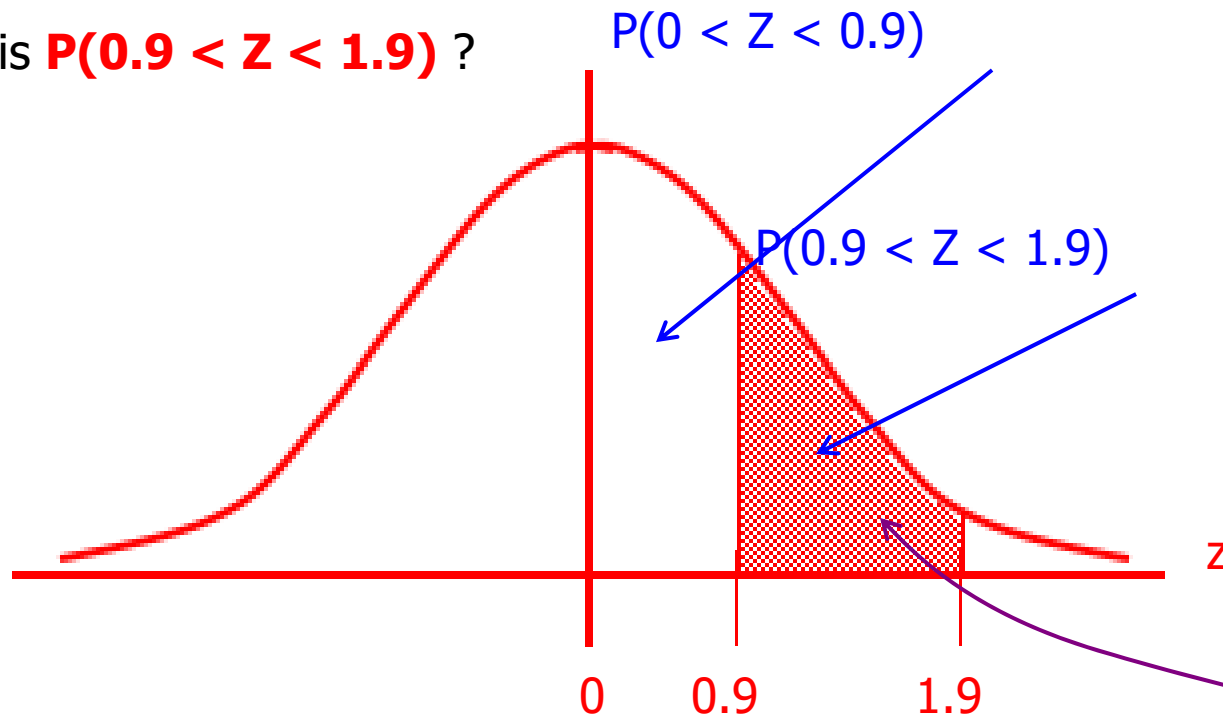


$$\begin{aligned} P(Z < 1.52) &= .5 + P(0 < Z < 1.52) \\ &= .5 + .4357 \\ &= \mathbf{.9357} \end{aligned}$$



## Using the Normal Table ([Table 3](#))...

- What is  **$P(0.9 < Z < 1.9)$**  ?



$$\begin{aligned} P(0.9 < Z < 1.9) &= P(0 < Z < 1.9) - P(0 < Z < 0.9) \\ &= .4713 - .3159 \\ &= \mathbf{.1554} \end{aligned}$$

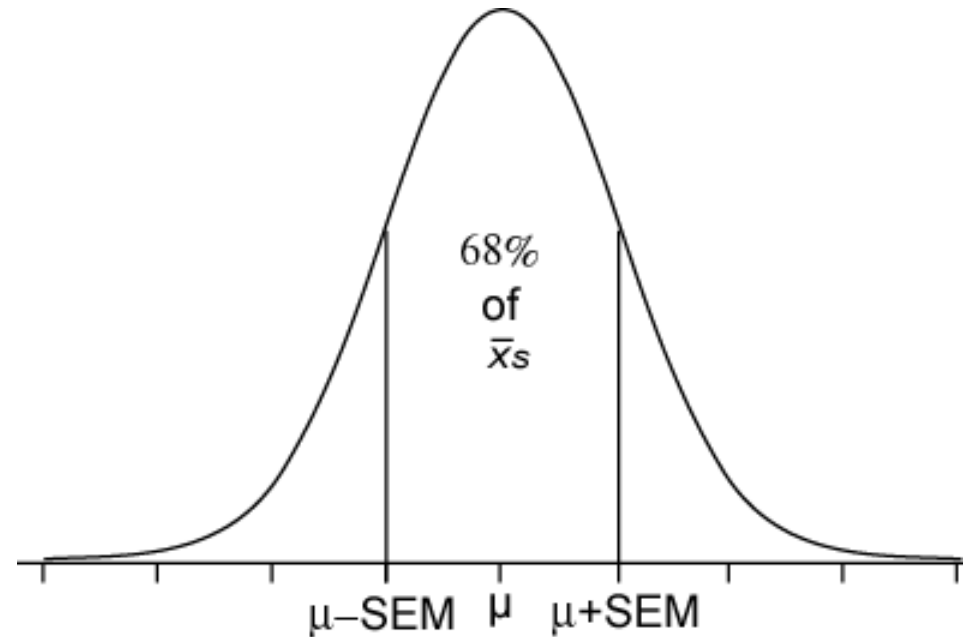
# Sampling Distributions of a Mean

The **sampling distributions of a mean (SDM)** describes the behavior of a sampling mean

*SE=standard error*

$$\bar{x} \sim N(\mu, SE_{\bar{x}})$$

$$\text{where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



# Standard Normal distribution

- If  $X$  is a Normally distributed random variable with mean =  $\mu$  and standard deviation =  $\sigma$ , then  $X$  can be converted to a Standard Normal random variable  $Z$  using:

$$Z = \frac{X - \mu}{\sigma}$$

# Standard Normal distribution (contd.)

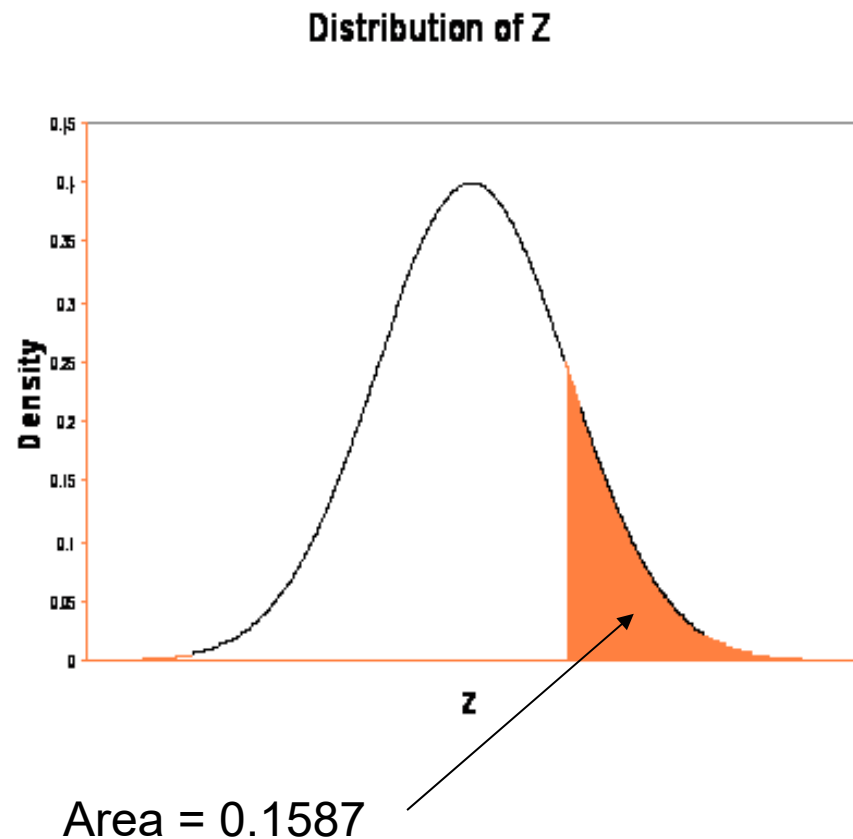
- $Z$  has mean = 0 and standard deviation = 1
- Using this transformation, we can calculate areas under **any** normal distribution

# Example

- Assume the distribution of blood pressure is Normally distributed with  $\mu = 80$  mm and  $\sigma = 10$  mm
- What percentage of people have blood pressure greater than 90?
- Z score transformation:  
 $Z = (90 - 80) / 10 = 1$

## Example (contd.)

- The percentage greater than 90 is equivalent to the area under the Standard Normal curve greater than  $Z = 1$ .
- From tables of the Standard Normal distribution, the area to the right of  $Z=1$  is 0.1587 (or 15.87%)



# Central Limit Theorem (CLT)

- Suppose you take any random sample from a population with mean  $\mu$  and variance  $\sigma^2$
- Then, for large sample sizes, the CLT states that the distribution of sample means is the Normal Distribution, with mean  $\mu$  and variance  $\sigma^2/n$  (i.e. standard deviation is  $\sigma/\sqrt{n}$  )
- If the original data is Normal then the sample means are Normal, irrespective of sample size

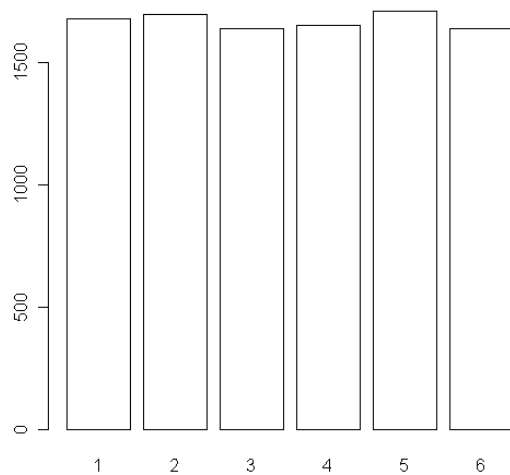
## What is it really saying?

- (1) It gives a relationship between the sample mean and population mean
  - This gives us a framework to extrapolate our sample results to the population (*statistical inference*);
- (2) It doesn't matter what the distribution of the original data is, the sample mean will always be Normally distributed when  $n$  is large.
  - This why the Normal is so central to statistics



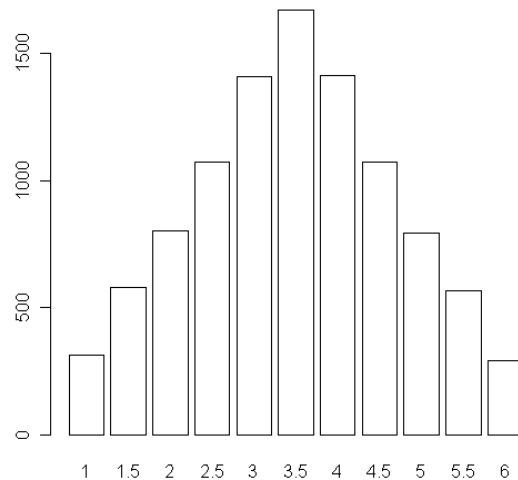
# Example: Toss 1, 2 or 10 dice (10,000 times)

Toss 1 dice  
Histogram of data



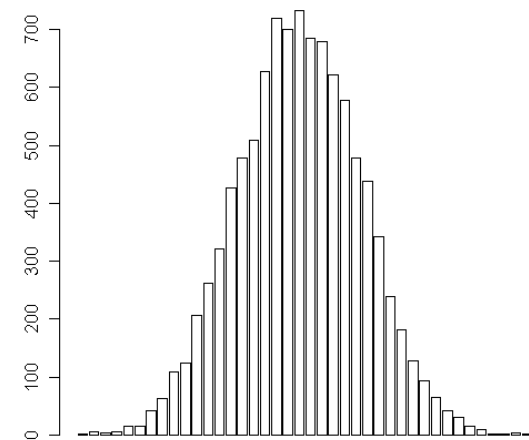
**Distribution of data is far from Normal**

Toss 2 dice  
Histogram of averages



**Distribution of averages approach Normal as sample size (no. of dice) increases**

Toss 10 dice  
Histogram of averages



## CLT cont'd

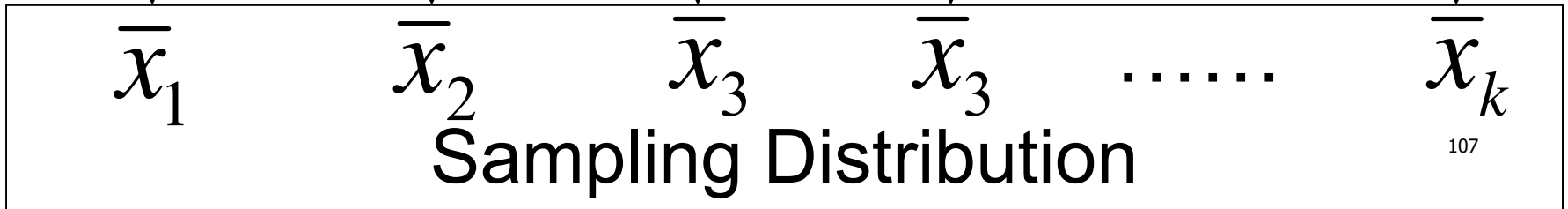
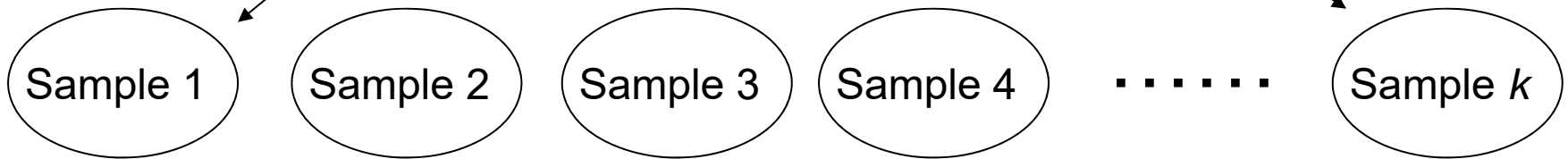
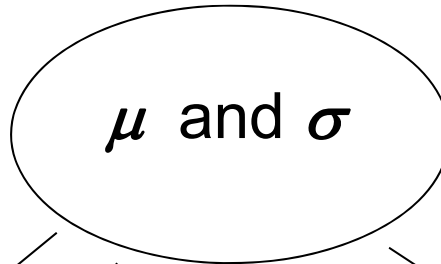
(3) It describes the distribution of the sample mean

- The values of  $\bar{x}$  obtained from repeatedly taking samples of size  $n$  describe a separate population
- The distribution of any statistic is often called the **sampling distribution**

# Sampling distribution of

 $\bar{X}$ 

Population



# CLT continued

(4) The mean of the sampling distribution of  $\bar{X}$  is equal to the population mean, i.e.

(5) Standard deviation of the sampling distribution of  $\bar{X}$  is the population standard deviation  $\div$  square root of sample size, i.e.

$$\mu_{\bar{X}} = \mu$$

$$\bar{X}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

# Estimates

- Since  $s$  is an estimate of  $\sigma$ , an estimate of  $\frac{\sigma}{\sqrt{n}}$  is  $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$
- This is known as the **standard error of the mean**
- Be careful not to confuse the standard deviation and the standard error !
  - Standard deviation describes the variability of the data
  - Standard error is the measure of the precision of  $\bar{x}$  as a measure of  $\mu$

# Confidence Interval

- A **confidence interval** for a population characteristic is an interval of plausible values for the characteristic. It is constructed so that, with a chosen degree of confidence (the **confidence level**), the value of the characteristic will be captured inside the interval
- E.g. we claim with 95% confidence that the population mean lies between 15.6 and 17.2

# Methods for Statistical Inference

Confidence Intervals

Hypothesis Tests

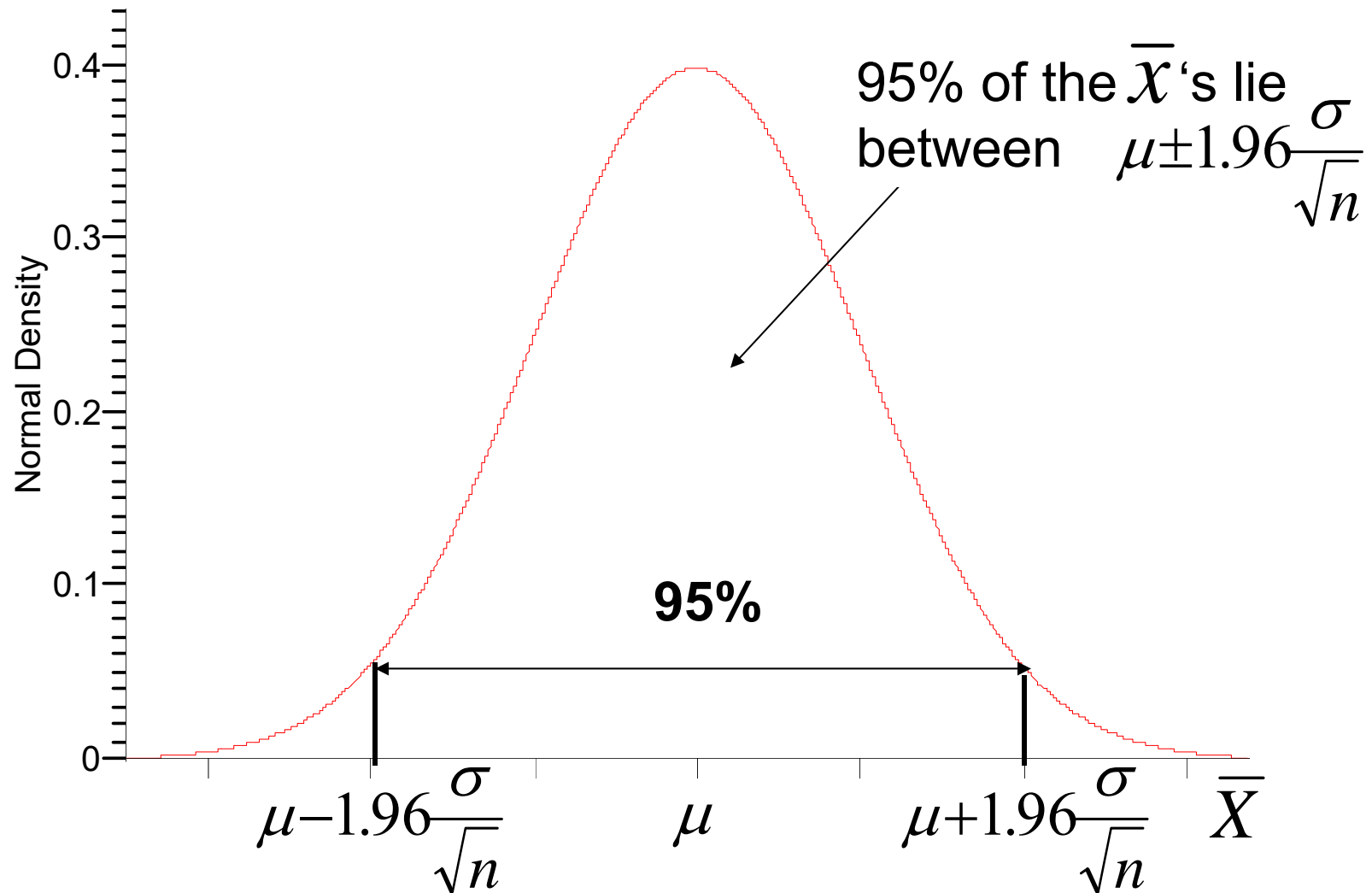
## Confidence Interval for $\mu$ when $\sigma$ is known

- A 95% confidence interval for  $\mu$  if  $\sigma$  is known is given by:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$



# Sampling distribution of $\bar{X}$



# Rationale for Confidence Interval

- From the sampling distribution of  $\bar{X}$  conclude that  $\mu$  and  $\bar{x}$  are within 1.96 standard errors ( $\frac{\sigma}{\sqrt{n}}$ ) of each other 95% of the time
- Otherwise stated, 95% of the intervals contain  $\mu$
- So, the interval  $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$  can be taken as an interval that typically would include  $\mu$

## Example

- A random sample of 80 tablets had an average potency of 15mg. Assume  $\sigma$  is known to be 4mg.
- $\bar{x} = 15, \sigma = 4, n = 80$
- A 95% confidence interval for  $\mu$  is

$$15 \pm 1.96 \times \frac{4}{\sqrt{80}}$$
$$= (14.12, 15.88)$$

# Confidence Interval for $\mu$ when $\sigma$ is unknown

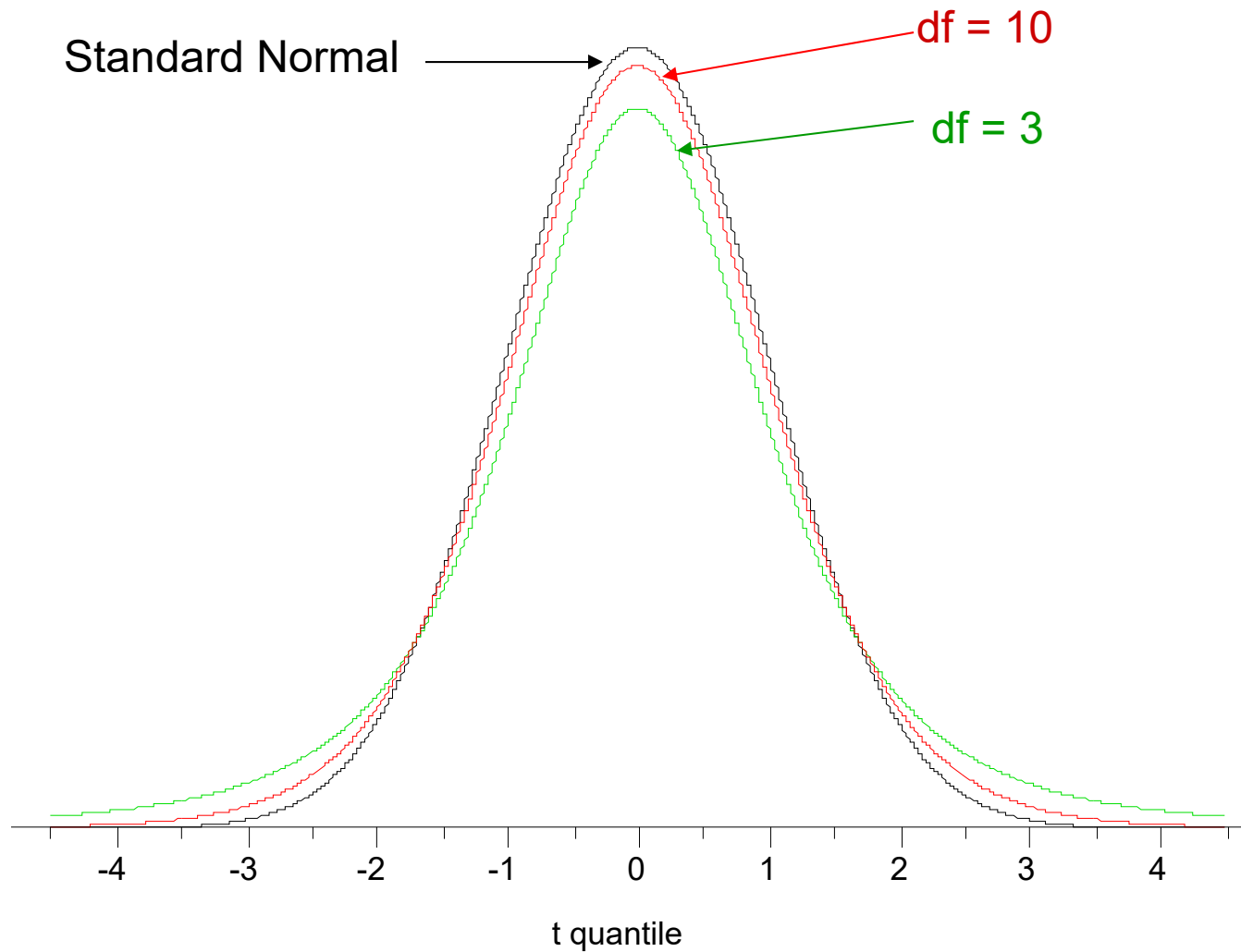
- Nearly always  $\sigma$  is unknown and is estimated using sample standard deviation  $s$
- The value 1.96 in the confidence interval is replaced by a new quantity, i.e.,  $t_{0.025}$
- The 95% confidence interval when  $\sigma$  is unknown is:

$$\bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}}$$

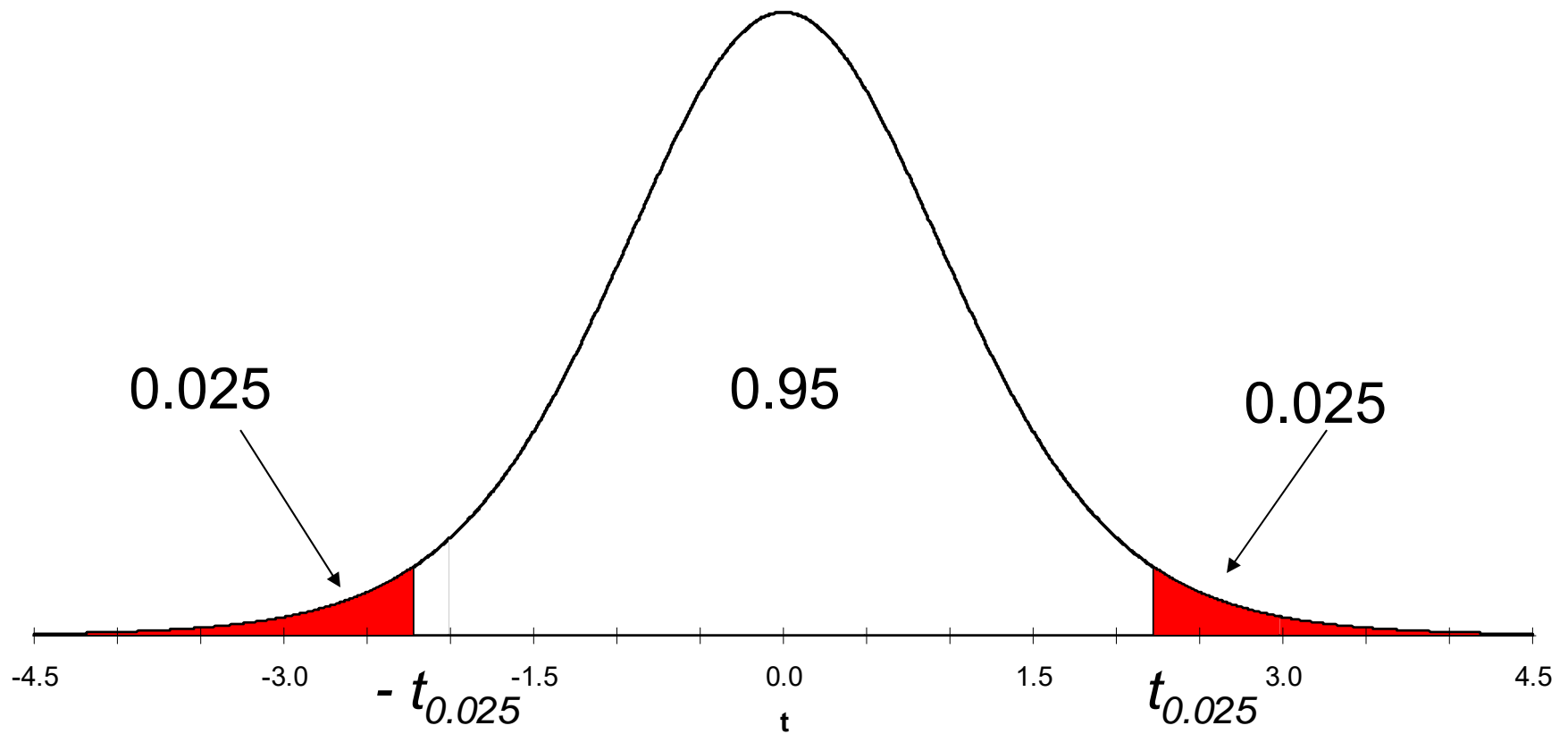
# Student's $t$ Distribution

- Closely related to the standard normal distribution  $Z$ 
  - Symmetric and bell-shaped
  - Has mean = 0 but has a larger standard deviation
- Exact shape depends on a parameter called **degrees of freedom** (df) which is related to sample size
  - In this context  $df = n-1$

# Student's $t$ distribution for 3, 10 df and standard Normal distribution



# Definition of $t_{0.025}$ values



## Example

- 26 measurements of the potency of a single batch of tablets in mg per tablet are as follows

498.38	489.31	505.50	495.24	490.17	483.2
488.47	497.71	503.41	482.25	488.14	
492.22	483.96	473.93	463.40	493.65	
499.48	496.05	494.54	508.58	488.42	
463.68	492.46	489.45	491.57	489.33	



## Example (contd.)

- $\bar{x} = 490.096$ , and  $s = 10.783$  mg per tablet
- $t_{0.025}$  with  $df = 25$  is 2.06

$$\begin{aligned}\bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} &= 490.096 \pm 2.06 \times \frac{10.783}{\sqrt{26}} \\ &= 490.096 \pm 4.356\end{aligned}$$

- So, the batch potency lies between 485.74 and 494.45 mg per tablet

# General Form of Confidence Interval

Estimate  $\pm$ (critical value from distribution).(standard error)

# Hypothesis testing

- Used to investigate the validity of a claim about the value of a population characteristic
- For example, the mean potency of a batch of tablets is 500mg per tablet, i.e.,  
 $\mu_0 = 500\text{mg}$

# Procedure

- Specify Null and Alternative hypotheses
- Specify test statistic
- Define what constitutes an exceptional outcome
- Calculate test statistic and determine whether or not to reject the Null Hypothesis

# Step 1

- Specify the hypothesis to be tested and the alternative that will be decided upon if this is rejected
  - The hypothesis to be tested is referred to as the **Null Hypothesis** (labelled  $H_0$ )
  - The alternative hypothesis is labelled  $H_1$
- For the earlier example this gives:

$$H_0 : \mu = 500\text{mg}$$

$$H_a : \mu \neq 500\text{mg}$$

## Step 1 (continued)

- The Null Hypothesis is assumed to be true unless the data clearly demonstrate otherwise

## Step 2

- Specify a test statistic which will be used to measure departure from

$$H_0 : \mu = \mu_0$$

where  $\mu_0$  is the value specified under the Null Hypothesis, e.g.  $\mu_0 = 500$  in the earlier example.

- For hypothesis tests on sample means the test statistic is:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

## Step 2 (contd.)

- The test statistic  $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$

is a 'signal to noise ratio', i.e. it measures how far is from  $\bar{x}$  in terms of standard error units  $\mu_0$

- The  $t$  distribution with  $df = n-1$  describes the distribution of the test statistics **if** the Null Hypothesis is true
- In the earlier example, the test statistic  $t$  has a  $t$  distribution with  $df = 25$



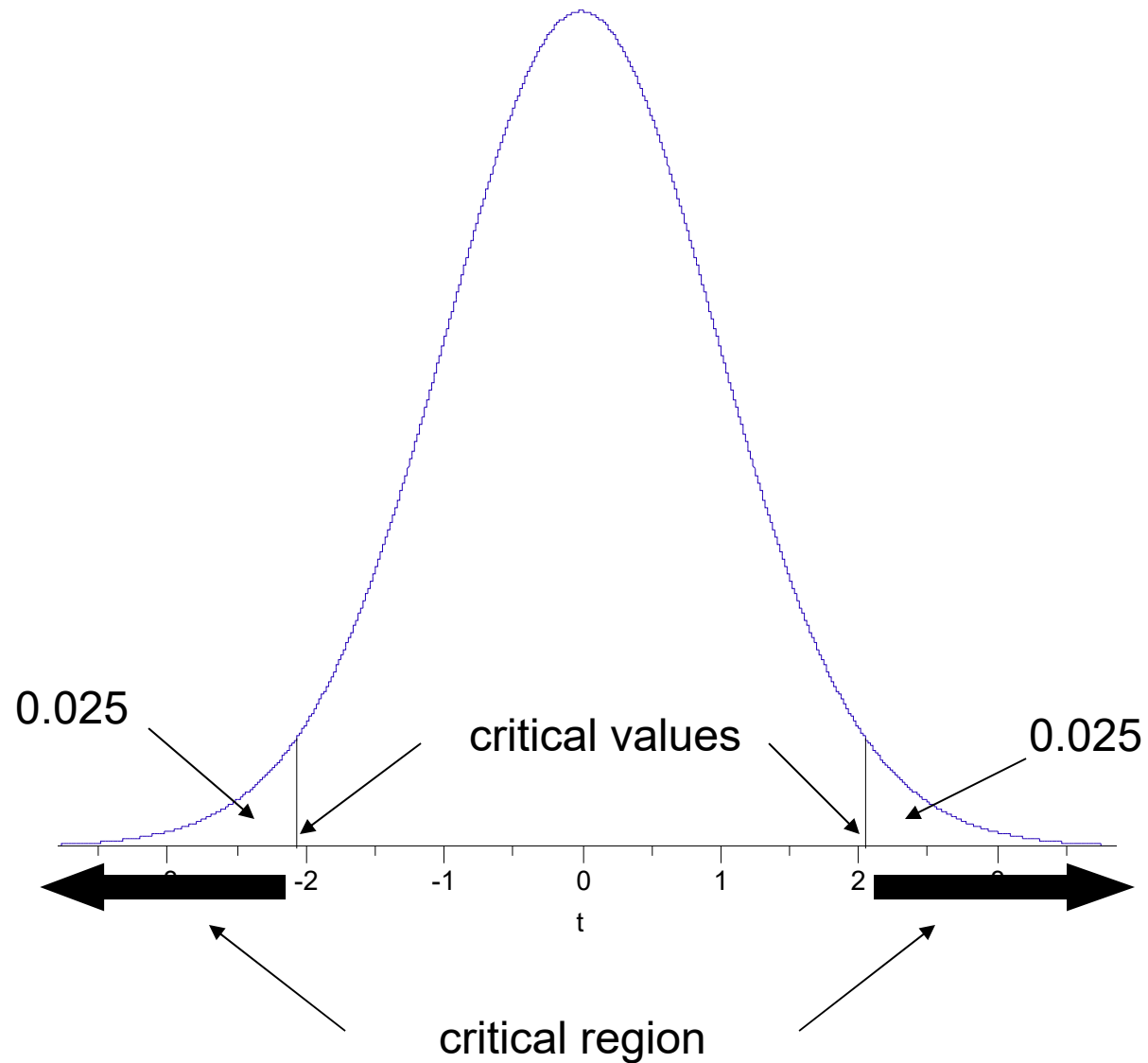
## Step 3

- Define what will be an exceptional outcome
  - a value of the test statistic is exceptional if it has only a small chance of occurring when the null hypothesis is true
- The probability chosen to define an exceptional outcome is called the **significance level** of the test and is labelled  $\alpha$ 
  - Conventionally,  $\alpha$  is chosen to be = 0.05

## Step 3 (contd.)

- $\alpha = 0.05$  gives cut-off values on the sampling distribution of  $t$  called **critical values**
  - values of the test statistic  $t$  lying beyond the critical values lead to rejection of the null hypothesis
- For the earlier example the critical value for a  $t$  distribution with  $df = 25$  is 2.06

# $t$ distribution with $df=25$ showing critical region



## Step 4

- Calculate the test statistic and see if it lies in the critical region
- For the example

$$t = \frac{490.096 - 500}{10.783 / \sqrt{26}}$$
$$= -4.683$$

- $t = -4.683$  is  $< -2.06$  so the hypothesis that the batch potency is 500 mg/tablet is rejected

## P value

The **P value** associated with a hypothesis test is the probability of getting sample values **as extreme or more extreme** than those actually observed, assuming null hypothesis to be true

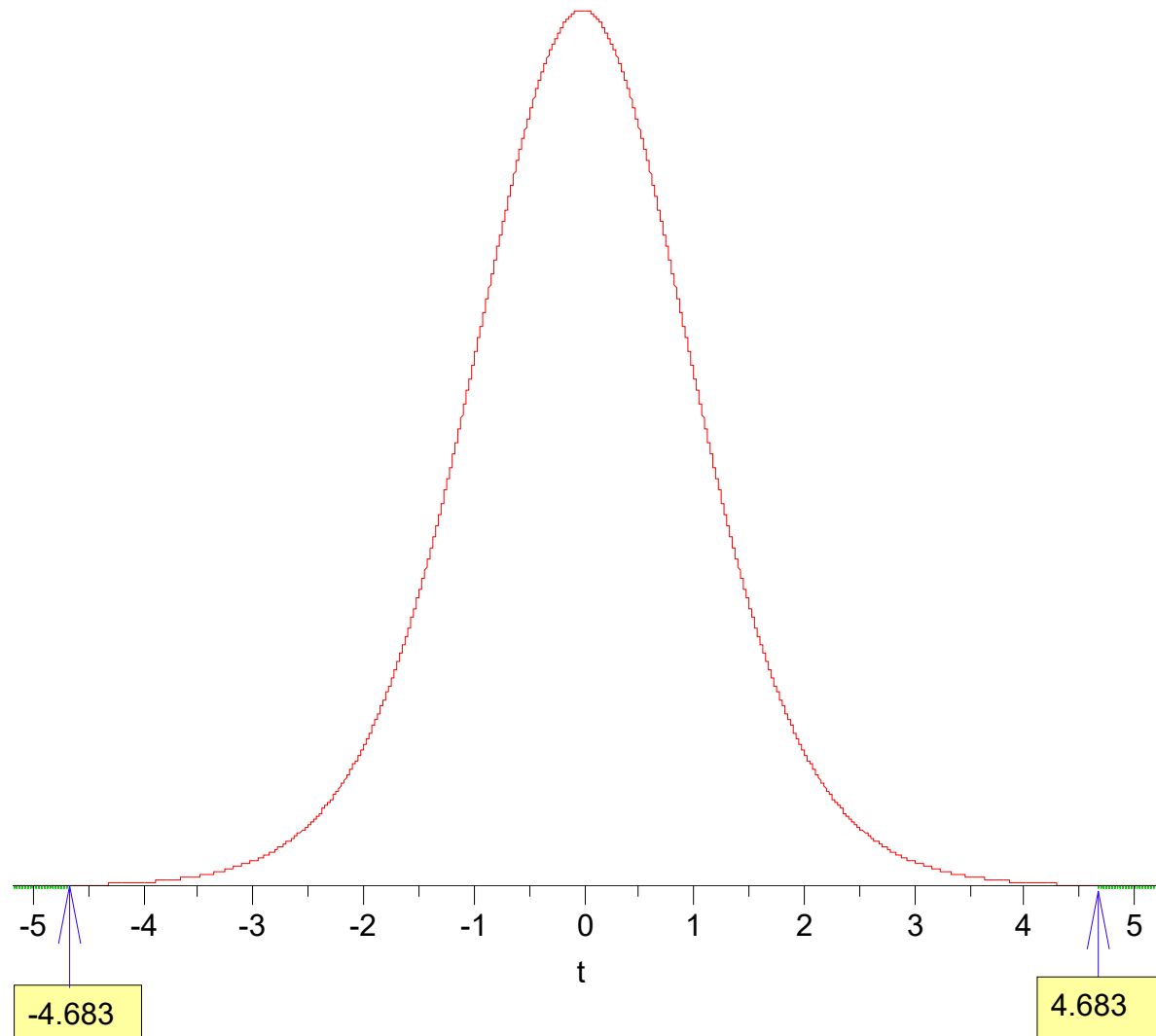
## Example (contd)

- P value = probability of observing a more extreme value of  $t$
- The observed  $t$  value was -4.683, so the P value is the probability of getting a value more extreme than  $\pm 4.683$
- This P value is calculated as the area under the  $t$  distribution below -4.683 plus the area above 4.683, i.e., 0.00008474 !

## Example (contd)

- Less than 1 in 10,000 chance of observing a value of  $t$  more extreme than -4.683 if the Null Hypothesis is true
- Evidence in favour of the alternative hypothesis is very strong

# P value (contd.)





# Two-tail and One-tail tests

- The test described in the previous example is a **two-tail** test
  - The null hypothesis is rejected if either an unusually large or unusually small value of the test statistic is obtained, i.e. the rejection region is divided between the two tails

# One-tail tests

- Reject the null hypothesis only if the observed value of the test statistic is
  - Too large
  - Too small
- In both cases the critical region is entirely in one tail so the tests are **one-tail** tests

# Statistical versus Practical Significance

- When we reject a null hypothesis it is usual to say the result is **statistically significant** at the chosen level of significance
- But should also always consider the **practical significance** of the **magnitude** of the difference between the estimate (of the population characteristic) and what the null hypothesis states that to be

# Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample)
- The technique is introduced by considering a one-sample z test
- The procedure is broken into four steps
- *Each* element of the procedure must be understood

# Hypothesis Testing Steps

- A. Null and alternative hypotheses
- B. Test statistic
- C. P-value and interpretation
- D. Significance level (optional)

# Null and Alternative Hypotheses

- Convert the research question to null and alternative hypotheses
- The **null hypothesis ( $H_0$ )** is a claim of “no difference in the population”
- The **alternative hypothesis ( $H_a$ )** claims “ $H_0$  is false”
- Collect data and seek evidence against  $H_0$  as a way of bolstering  $H_a$  (deduction)

# Illustrative Example: “Body Weight”

- **The problem:** In the 1970s, 20–29 year old men in the U.S. had a mean  $\mu$  body weight of 170 pounds. Standard deviation  $\sigma$  was 40 pounds. We test whether mean body weight in the population now differs.
- **Null hypothesis**  $H_0: \mu = 170$  (“no difference”)
- The **alternative hypothesis** can be either  $H_a: \mu > 170$  (**one-sided test**)  
or  
 $H_a: \mu \neq 170$  (**two-sided test**)

## §9.2 Test Statistic

This is an example of a one-sample test of a mean when  $\sigma$  is known. Use this statistic to test the problem:

$$Z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

where  $\mu_0 \equiv$  population mean assuming  $H_0$  is true

$$\text{and } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



## Illustrative Example: $z$ statistic

- For the illustrative example,  $\mu_0 = 170$
- We know  $\sigma = 40$
- Take an SRS of  $n = 64$ . Therefore

- If we found a sample mean of 173, then

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = 5$$

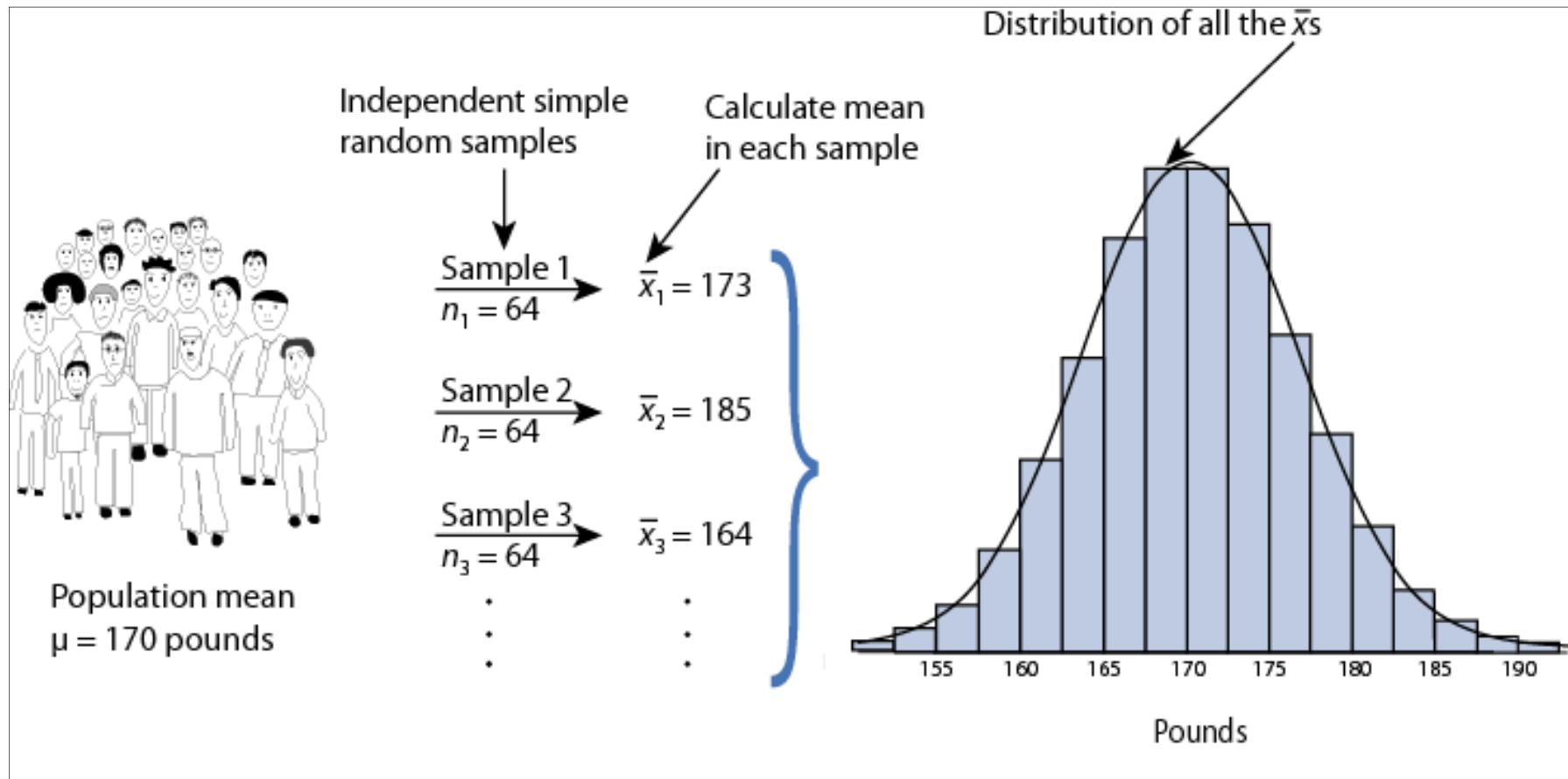
$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{173 - 170}{5} = 0.60$$

## Illustrative Example: z statistic

If we found a sample mean of 185, then

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{185 - 170}{5} = 3.00$$

# Reasoning Behind $\mu Z_{\text{stat}}$



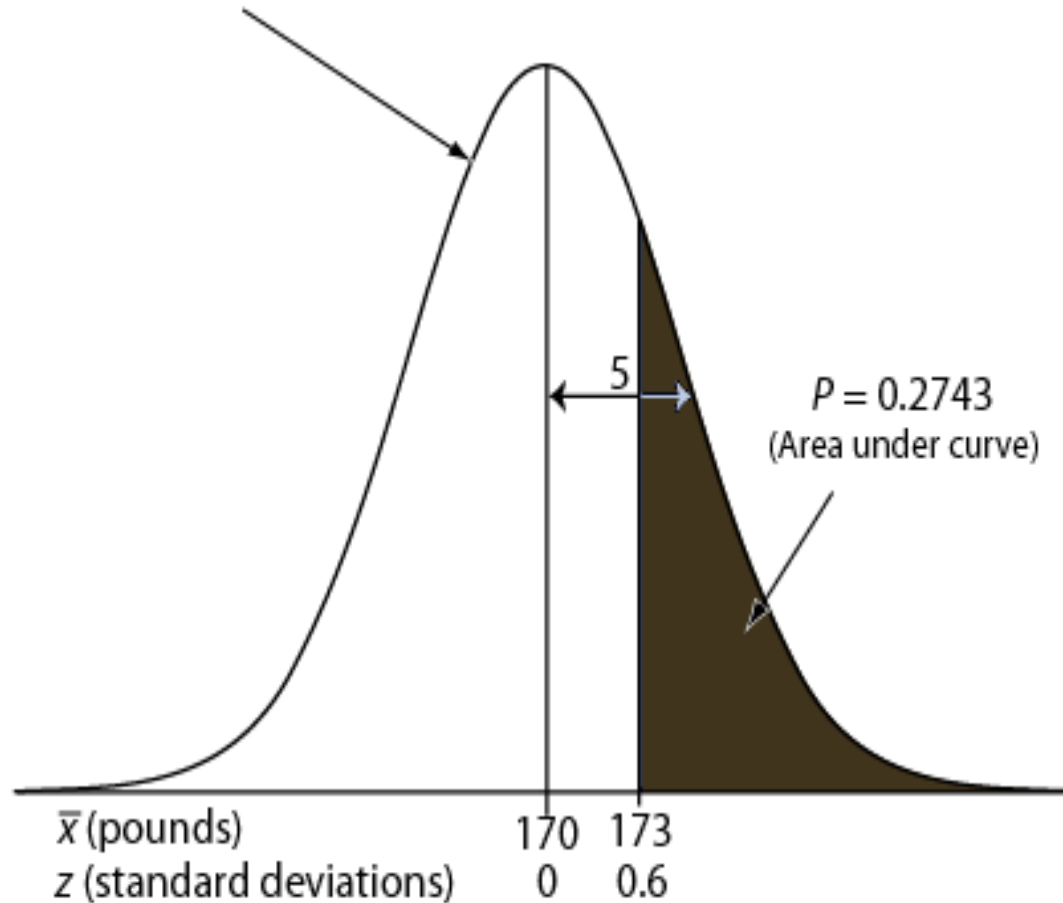
Sampling distribution of  $\bar{x}$  under  $H_0: \mu = 170$  for  $n = 64 \Rightarrow \bar{x} \sim N(170, 5)$

## $P$ -value

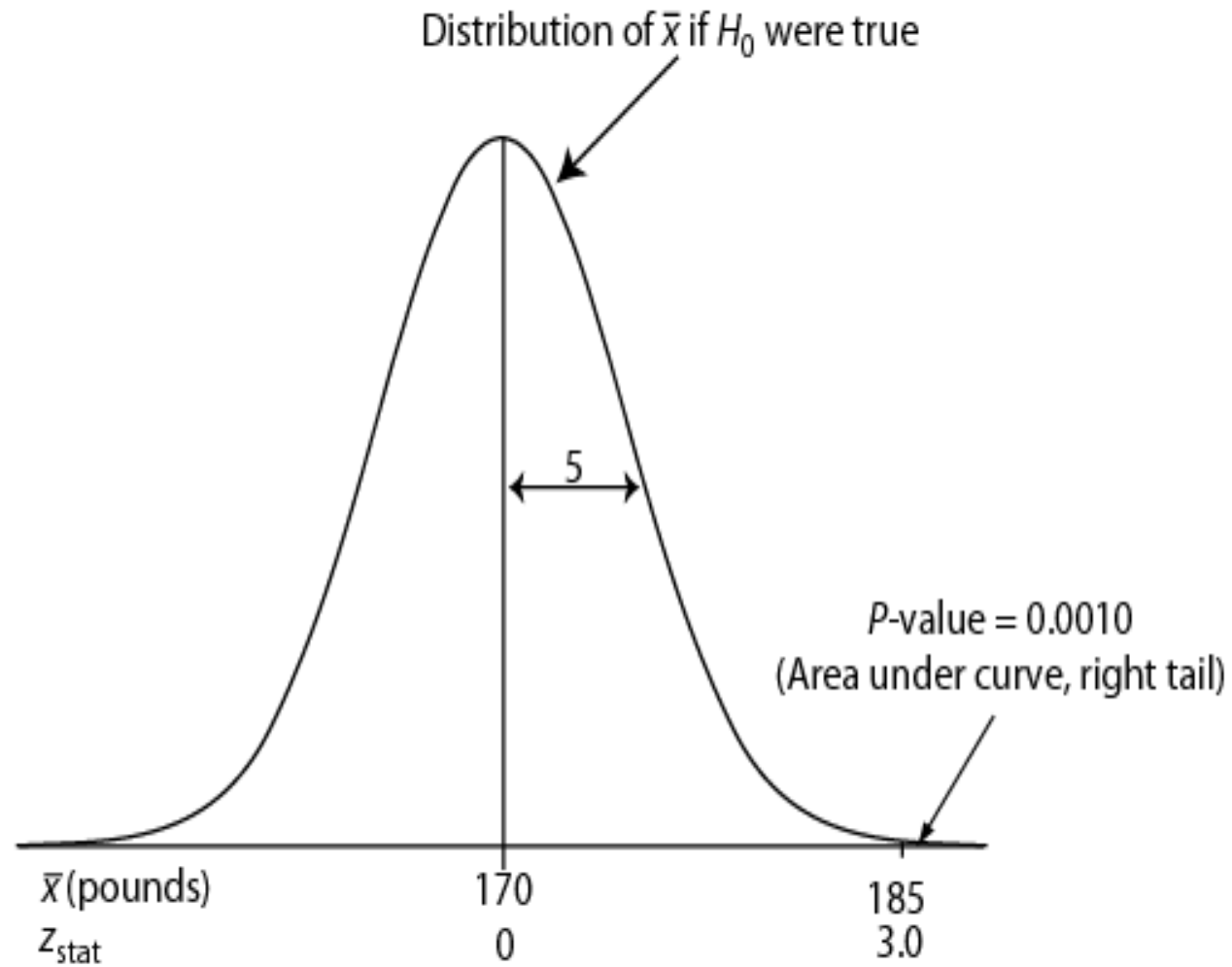
- The  $P$ -value answer the question: What is the probability of the observed test statistic or one more extreme **when  $H_0$  is true?**
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond the  $z_{\text{stat}}$ .
- Convert  $z$  statistics to  $P$ -value :
  - For  $H_a: \mu > \mu_0 \Rightarrow P = \Pr(Z > z_{\text{stat}}) =$  right-tail beyond  $z_{\text{stat}}$
  - For  $H_a: \mu < \mu_0 \Rightarrow P = \Pr(Z < z_{\text{stat}}) =$  left tail beyond  $z_{\text{stat}}$
  - For  $H_a: \mu \neq \mu_0 \Rightarrow P = 2 \times$  one-tailed  $P$ -value
- Use Table B or software to find these probabilities (next two slides).

# One-sided $P$ -value for $z_{\text{stat}}$ of 0.6

Distribution of  $\bar{x}$  and  $z_{\text{stat}}$  if  $H_0$  were true

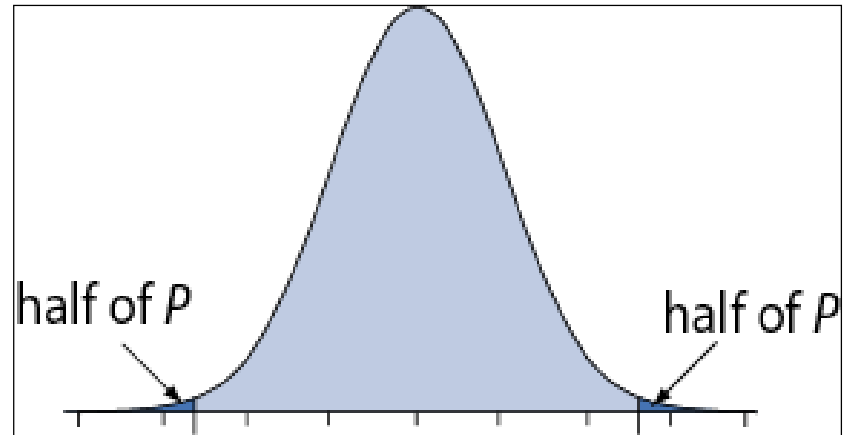


# One-sided $P$ -value for $z_{\text{stat}}$ of 3.0



# Two-Sided $P$ -Value

- One-sided  $H_a \Rightarrow$  AUC in tail beyond  $z_{\text{stat}}$
- Two-sided  $H_a \Rightarrow$  consider potential deviations in both directions  $\Rightarrow$  double the one-sided  $P$ -value



Examples: If one-sided  $P = 0.0010$ , then two-sided  $P = 2 \times 0.0010 = 0.0020$ .  
If one-sided  $P = 0.2743$ , then two-sided  $P = 2 \times 0.2743 = 0.5486$ .

# Interpretation

- $P$ -value answer the question: What is the probability of the observed test statistic ... **when  $H_0$  is true?**
- Thus, smaller and smaller  $P$ -values provide stronger and stronger evidence against  $H_0$
- Small  $P$ -value  $\Rightarrow$  strong evidence



# Interpretation

## Conventions\*

$P > 0.10 \Rightarrow$  non-significant evidence against  $H_0$

$0.05 < P \leq 0.10 \Rightarrow$  marginally significant evidence

$0.01 < P \leq 0.05 \Rightarrow$  significant evidence against  $H_0$

$P \leq 0.01 \Rightarrow$  highly significant evidence against  $H_0$

## Examples

$P = .27 \Rightarrow$  non-significant evidence against  $H_0$

$P = .01 \Rightarrow$  highly significant evidence against  $H_0$

**\* It is *unwise* to draw firm borders for “significance”**

# $\alpha$ -Level (Used in some situations)

- Let  $\alpha \equiv$  probability of erroneously rejecting  $H_0$
- Set a threshold (e.g., let  $\alpha = .10, .05, \text{ or } \textit{whatever}$ )
- Reject  $H_0$  when  $P \leq \alpha$
- Retain  $H_0$  when  $P > \alpha$
- Example: Set  $\alpha = .10$ . Find  $P = 0.27 \Rightarrow$  retain  $H_0$
- Example: Set  $\alpha = .01$ . Find  $P = .001 \Rightarrow$  reject  $H_0$

## (Summary) One-Sample z Test

A. Hypothesis statements

$H_0: \mu = \mu_0$  vs.

$H_a: \mu \neq \mu_0$  (two-sided) or

$H_a: \mu < \mu_0$  (left-sided) or

$H_a: \mu > \mu_0$  (right-sided)

B. Test statistic

C. P-value: convert  $z_{\text{stat}}$  to P value

D. Significance statement (usually not necessary)

$$Z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \text{ where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

## Two-Sample Inferences

- So far, we have dealt with inferences about  $\mu$  for a **single** population using a **single** sample.
- Many studies are undertaken with the objective of comparing the characteristics of two populations. In such cases we need two samples, one for each population
- The two samples will be **independent** or dependent (**paired**) according to how they are selected

# Example

- Animal studies to compare toxicities of two drugs

2 independent samples:

Select sample of rats for drug 1 and another sample of rats for drug 2

2 paired samples:

Select a number of pairs of litter mates and use one of each pair for drug 1 and drug 2

# Two Sample t-test

- Consider inferences on 2 independent samples
- We are interested in testing whether a difference exists in the population means,  $\mu_1$  and  $\mu_2$

Formulate hypotheses

$$H_0 : \mu_2 - \mu_1 = 0$$

$$H_a : \mu_2 - \mu_1 \neq 0$$

## Two Sample t-Test

- It is natural to consider the statistic  $\bar{x}_2 - \bar{x}_1$  and its sampling distribution
- The distribution is centred at  $\mu_2 - \mu_1$ , with standard error

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the two populations are normal, the sampling distribution is normal
- For large sample sizes ( $n_1$  and  $n_2 > 30$ ), the sampling distribution is approximately normal even if the two populations are not normal (CLT)

# Two Sample t-Test

- The two-sample t-statistic is defined as

$$t = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The two sample standard deviations are combined to give a pooled estimate of the population standard deviation  $\sigma$



## Two-sample Inference

- The t statistic has  $n_1+n_2-2$  degrees of freedom
- Calculate critical value & p value as per usual
- The 95% confidence interval for  $\mu_2-\mu_1$  is

$$(\bar{x}_2 - \bar{x}_1) \pm t_{0.025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

## Example

Population	n	mean	s
Drug 1	20	35.9	11.9
Drug 2	38	36.6	12.3

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{(19)(141.61) + (37)(151.29)}{56} \\ &= 148.01 \end{aligned}$$

## Example (contd)

$$t = \frac{(\bar{x}_2 - \bar{x}_1) - 0}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= -0.21$$

- Two-tailed test with 56 df and  $\alpha=0.05$  therefore we reject the null hypothesis if  $t > 2$  or  $t < -2$
- Fail to reject - there is insufficient evidence of a difference in mean between the two drug populations
- Confidence interval is -7.42 to 6.02

## Paired t-test

- Methods for independent samples are **not** appropriate for paired data.
- Two related observations (i.e. two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another.
- Calculation of the t-statistic, 95% confidence intervals for the mean difference and P-values are estimated as presented previously for one-sample testing.

# Example

- 14 cardiac patients were placed on a special diet to lose weight. Their weights (kg) were recorded before starting the diet and after one month on the diet
- Question: Do the data provide evidence that the diet is effective?

<b>Patient</b>	<b>Before</b>	<b>After</b>	<b>Difference</b>
1	62	59	3
2	62	60	2
3	65	63	2
4	88	78	10
5	76	75	1
6	57	58	-1
7	60	60	0
8	59	52	7
9	54	52	2
10	68	65	3
11	65	66	-1
12	63	59	4
13	60	58	2
14	56	55	1

## Example

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d > 0$$

$$\bar{x}_d = 2.5 \quad s_d = 2.98 \quad n = 14$$

$$t = \frac{\bar{x}_d - 0}{\frac{s_d}{\sqrt{n}}} = \frac{2.5}{\frac{2.98}{\sqrt{14}}} = 3.14$$

## Example (contd)

- Critical Region (1 tailed)  $t > 1.771$
- Reject  $H_0$  in favour of  $H_a$
- P value is the area to the right of 3.14  
 $= 1 - 0.9961 = 0.0039$
- 95% Confidence Interval for  $\mu_d = \mu_1 - \mu_2$   
 $2.5 \pm 2.17 (2.98/\sqrt{14})$   
 $= 2.5 \pm 1.72$   
 $= 0.78 \text{ to } 4.22$



## Example (cont)

- Suppose these data were (incorrectly) analysed as if the two samples were independent...  
→  $t=0.80$

## Example (contd)

- We calculate  $t=0.80$
- This is an upper tailed test with 26 df and  $\alpha=0.05$  (5% level of significance) therefore we reject  $H_0$  if  $t>1.706$
- Fail to reject - there is not sufficient evidence of a difference in mean between 'before' and 'after' weights

## Wrong Conclusions

- By ignoring the paired structure of the data, we incorrectly conclude that there was no evidence of diet effectiveness.
- When pairing is ignored, the variability is inflated by the subject-to-subject variation.
- The paired analysis eliminates this source of variability from the calculations, whereas the unpaired analysis includes it.
- Take home message: NB to use the right test for your data. If data is paired, use a test that accounts for this.

# Analysis of Variance (ANOVA)

- Many investigations involved a comparison of **more than two** population means
- Need to be able to extend our two sample methods to situations involving more than two samples
- i.e. equivalent of the paired samples t-test, but allows for two or more levels of the categorical variable
- Tests whether the mean of the dependent variable differs by the categorical variable
- Such methods are known collectively as the **analysis of variance**

# Completely Randomised Design/one-way ANOVA

- Equivalent to independent samples design for two populations
- A completely randomised design is frequently referred to as a **one-way ANOVA**
- Used when you have a categorical independent variable (with two or more categories) and a **normally distributed** interval dependent variable (e.g. \$10,000, \$15,000, \$20,000) and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable
- e.g. compare three methods for measuring tablet hardness. 15 tablets are randomly assigned to three groups of 5 and each group is measured by one of these methods

# ANOVA example

Mean of the dependent variable differs significantly among the levels of program type. However, we do not know if the difference is between only two of the levels or all three of the levels.

```
anova write prog
```

```
Number of obs =      200      R-squared      = 0.1776
Root MSE      = 8.63918      Adj R-squared = 0.1693
```

Source	Partial SS	df	MS	F	Prob > F
Model	3175.69786	2	1587.84893	21.27	0.0000
prog	3175.69786	2	1587.84893	21.27	0.0000
Residual	14703.1771	197	74.635417		
Total	17878.875	199	89.843593		

```
tabulate prog, summarize(write)
```

type of program	Summary of writing score Mean	Std. Dev.	Freq.
general	51.333333	9.3977754	45
academic	56.257143	7.9433433	105
vocation	46.76	9.3187544	50
Total	52.775	9.478586	200

See that the students in the academic program have the highest mean writing score, while students in the vocational program have the lowest.

# Example

Compare three methods for measuring tablet hardness.  
15 tablets are randomly assigned to three groups of 5

Method A	Method B	Method C
102	99	103
101	100	100
101	99	99
100	101	104
102	98	102

## Hypothesis Tests: One-way ANOVA

- K populations

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$H_A$  : *at least one  $\mu$  is different*



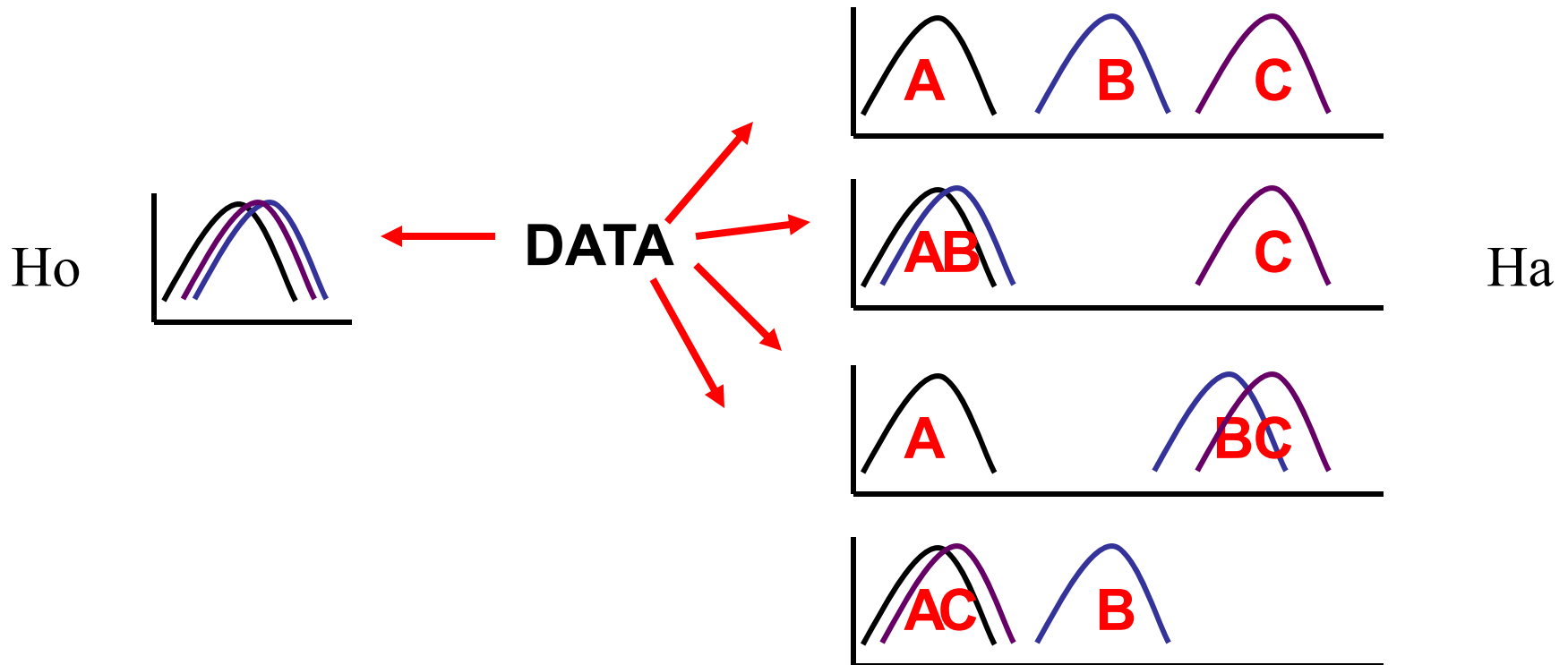
# Do the samples come from different populations?

- Two-sample (t-test)



# Do the samples come from different populations?

- One-way ANOVA (F-test)



# F-test

- The ANOVA extension of the t-test is called the **F-test**
- Basis: We can decompose the total variation in the study into sums of squares
- Tabulate in an **ANOVA table**

# Decomposition of total variability (sum of squares)

Assign subscripts to the data

- $i$  is for treatment (or method in this case)
- $j$  are the observations made within treatment

e.g.

- $y_{11}$  = first observation for Method A i.e. 102
- $y_{1.}$  = average for Method A

## Using algebra

**Total Sum of Squares (SST) = Treatment Sum of Squares (SSX)  
+ Error Sum of Squares (SSE)**

$$\sum (y_{ij} - \bar{y})^2 = \sum (\bar{y}_{i.} - \bar{y})^2 + \sum (y_{ij} - \bar{y}_{i.})^2$$

# ANOVA table

	df	SS	MS	F	P-value
<b>Treatment (between groups)</b>	$df (X)$	$SSX$	$\frac{SSX}{df (X)}$	$\left. \begin{array}{l} MSX \\ MSE \end{array} \right\}$	<i>Look up !</i>
<b>Error (within groups)</b>	$df (E)$	$SSE$	$\frac{SSE}{df (E)}$	$\left. \begin{array}{l} \\ \end{array} \right\}$	
<b>Total</b>	$df (T)$	$SST$			

## Example (Contd)

- Are any of the methods different?
- P-value=0.0735
- At the 5% level of significance, there is no evidence that the 3 methods differ

# Two-Way ANOVA

- Often, we wish to study 2 (or more) independent variables (factors) in a single experiment
- An ANOVA of observations each of which can be classified in two ways is called a *two-way ANOVA*

# Randomised Block Design

- This is an extension of the paired samples situation to more than two populations
- A block consists of homogenous items and is equivalent to a pair in the paired samples design
- The randomised block design is generally more powerful than the completely randomised design (/one way anova) because the variation between blocks is removed from the test statistic



## Decomposition of sums of squares

$$\sum (y_{ij} - \bar{y})^2 = \sum (\bar{y}_{i.} - \bar{y})^2 + \sum (\bar{y}_{.j} - \bar{y})^2 + \sum (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y})^2$$

Total SS = Between Blocks SS + Between Treatments SS + Error SS

- Similar to the one-way ANOVA, we can decompose the overall variability in the data (total SS) into components describing variation relating to the factors (block, treatment) & the error (what's left over)
- We compare Block SS and Treatment SS with the Error SS (a signal-to-noise ratio) to form F-statistics, from which we get a p-value

# Example

- An experiment was conducted to compare the mean bioavailability (as measured by AUC) of three drug products from laboratory rats.
- Eight litters (each consisting of three rats) were used for the experiment. Each litter constitutes a block and the rats within each litter are randomly allocated to the three drug products

## Example (cont'd)

Litter	Product A	Product B	Product C
1	89	83	94
2	93	75	78
3	87	75	89
4	80	76	85
5	80	77	84
6	87	73	84
7	82	80	75
8	68	77	75

Example (cont'd):  
ANOVA table

<b>Source</b>	<b>df</b>	<b>SS</b>	<b>MS</b>	<b>F-ratio</b>	<b>P-value</b>
<b>Product</b>	<b>2</b>	<b>200.333</b>	<b>100.167</b>	<b>3.4569</b>	<b>0.0602</b>
<b>Litter</b>	<b>7</b>	<b>391.833</b>	<b>55.9762</b>	<b>1.9318</b>	<b>0.1394</b>
<b>Error</b>	<b>14</b>	<b>405.667</b>	<b>28.9762</b>		
<b>Total</b>	<b>23</b>	<b>997.833</b>			

# Interactions

- The previous tests for block and treatment are called tests for *main effects*
- ***Interaction effects*** happen when the effects of one factor are different depending on the level (category) of the other factor

# Example

- 24 patients in total randomised to either Placebo or Prozac
- Happiness score recorded
- Also, patients gender may be of interest & recorded
- There are two factors in the experiment: treatment & gender
  - Two-way ANOVA

# Example

- Tests for Main effects:
  - Treatment: are patients happier on placebo or prozac?
  - Gender: do males and females differ in score?
- Tests for Interaction:
  - Treatment x Gender: Males may be happier on prozac than placebo, but females not be happier on prozac than placebo. Also vice versa. Is there any evidence for these scenarios?
  - Include interaction in the model, along with the two factors treatment & gender

# More jargon: factors, levels & cells

Happiness score

— Factor 2 Treatment →

Levels

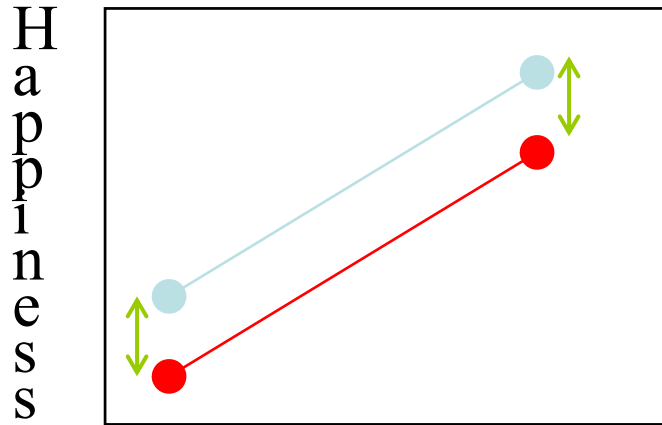
Cells

Factor 1  
Gender  
↓

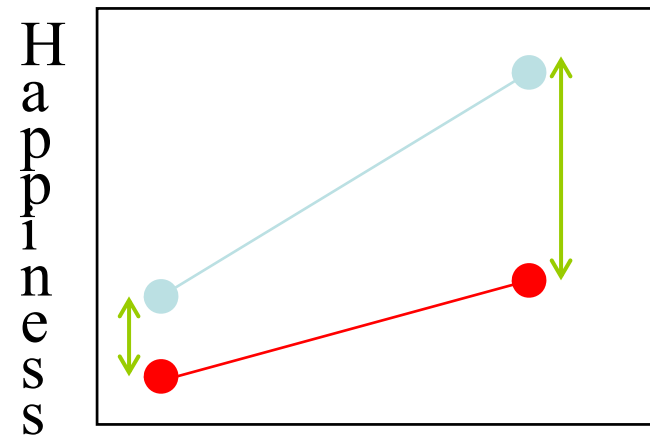
	Placebo	Prozac
Male	3 4 2 3 4 3	7 7 6 5 6 6
Female	4 5 4 6 6 4.5	5 5 5 4 6 6



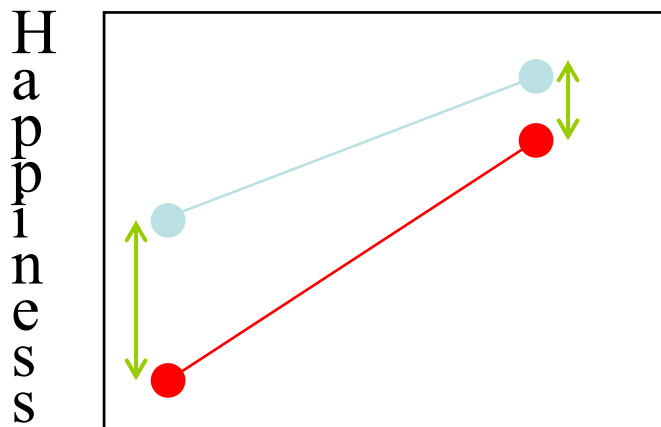
# What do interactions look like?



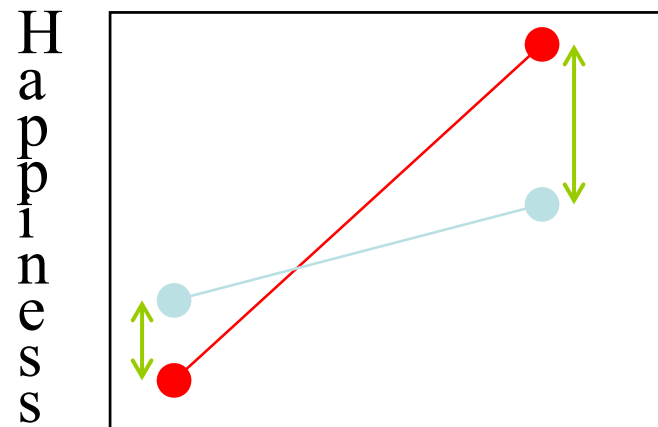
Placebo Prozac  
**NO INTERACTION!**



Placebo Prozac



Placebo Prozac



Placebo Prozac

# Results

## Tests of Between-Subjects Effects

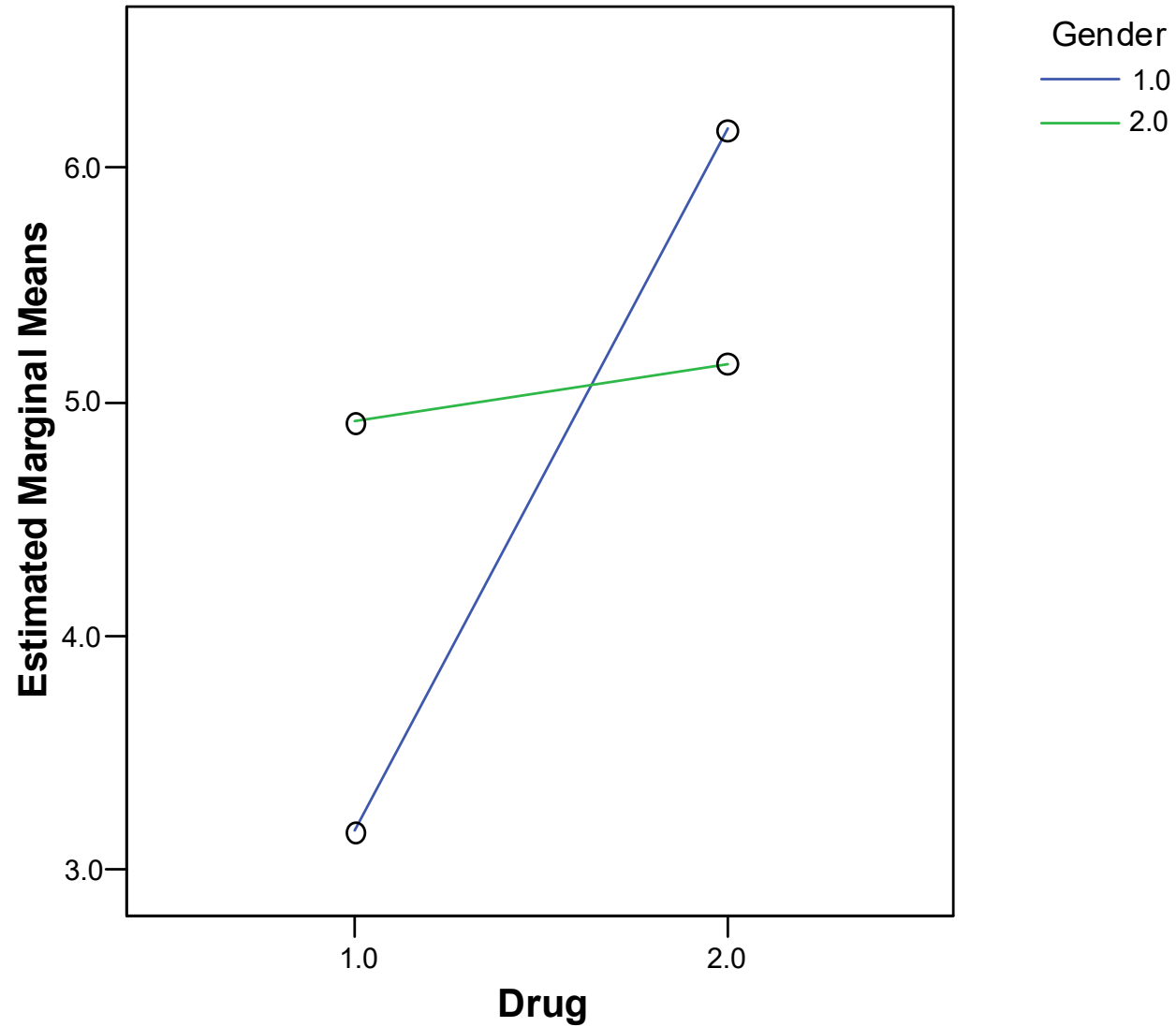
Dependent Variable: Happiness

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	28.031 <sup>a</sup>	3	9.344	14.705	.000
Intercept	565.510	1	565.510	889.984	.000
Drug	15.844	1	15.844	24.934	.000
Gender	.844	1	.844	1.328	.263
Drug * Gender	11.344	1	11.344	17.852	.000
Error	12.708	20	.635		
Total	606.250	24			
Corrected Total	40.740	23			

a. R Squared = .688 (Adjusted R Squared = .641)

# Interaction? Plot the means

Estimated Marginal Means of Happiness



# Example: Conclusions

- Significant evidence that drug treatment affects happiness in depressed patients ( $p < 0.001$ )
  - Prozac is effective, placebo is not
- No significant evidence that gender affects happiness ( $p = 0.263$ )
- Significant evidence of an interaction between gender and treatment ( $p < 0.001$ )
  - Prozac is effective in men but not in women!!\*

# Introduction to Linear Regression and Correlation Analysis

# Goals

**After this, you should be able to:**

- Calculate and interpret the simple correlation between two variables
- Determine whether the correlation is significant
- Calculate and interpret the simple linear regression equation for a set of data
- Understand the assumptions behind regression analysis
- Determine whether a regression model is significant

## Goals

*(continued)*

### **After this, you should be able to:**

- Calculate and interpret confidence intervals for the regression coefficients
- Recognize regression analysis applications for purposes of prediction and description
- Recognize some potential problems if regression analysis is used incorrectly
- Recognize nonlinear relationships between two variables

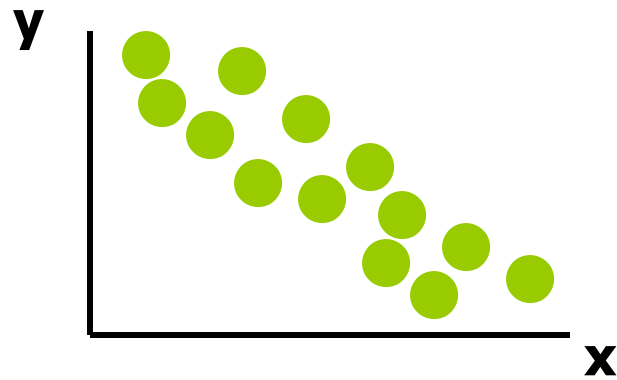
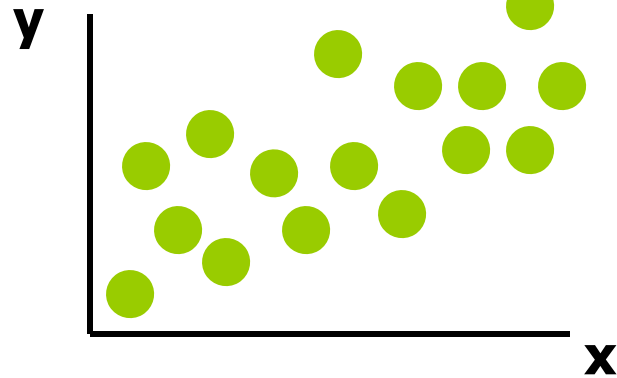
# Scatter Plots and Correlation

- A **scatter plot** (or scatter diagram) is used to show the relationship between two variables
- **Correlation** analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

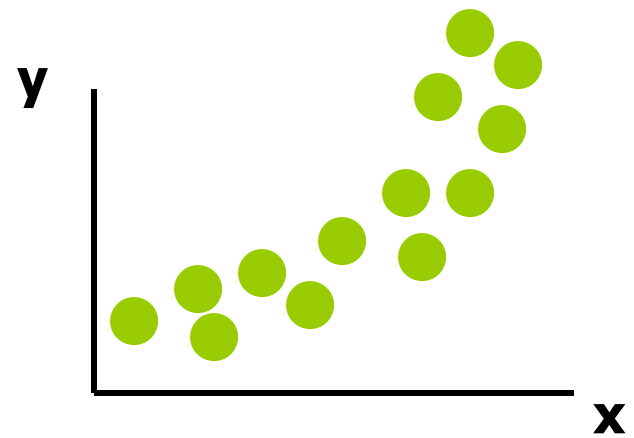
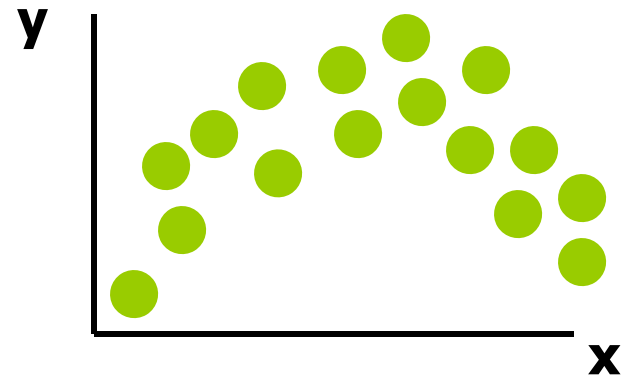


# Scatter Plot Examples

**Linear relationships**

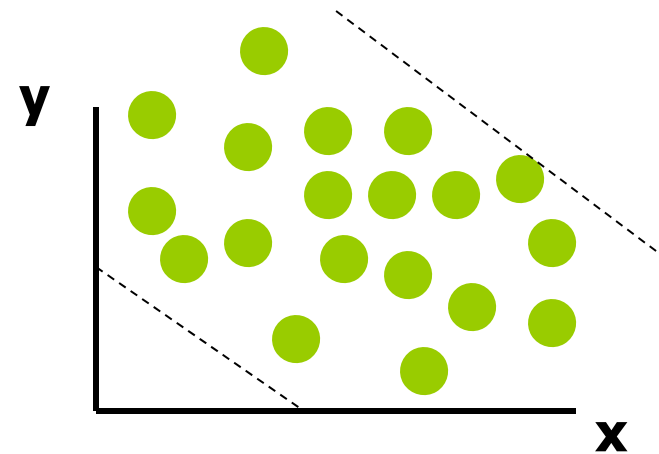
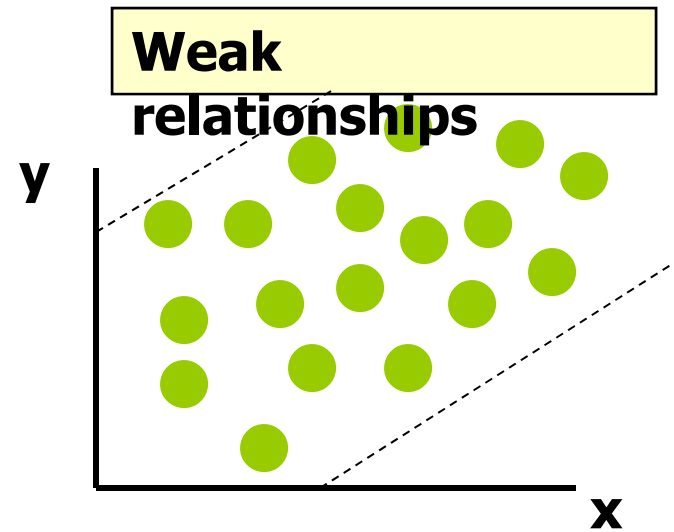
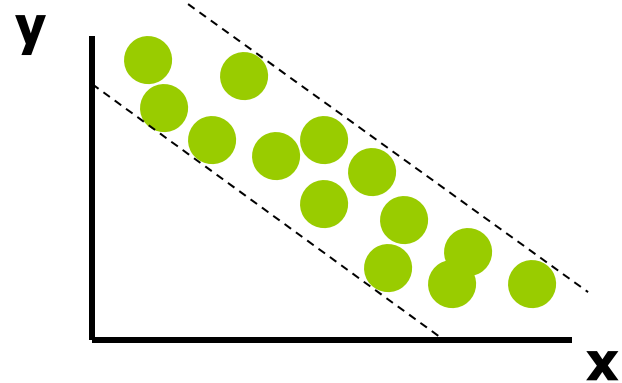
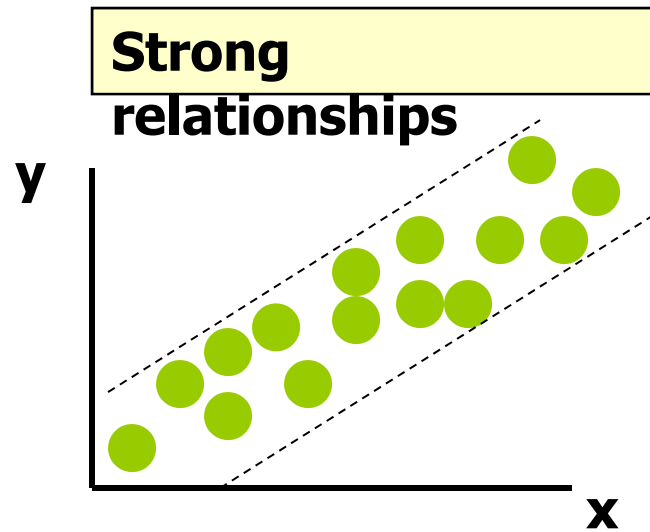


**Curvilinear relationships**



# Scatter Plot Examples

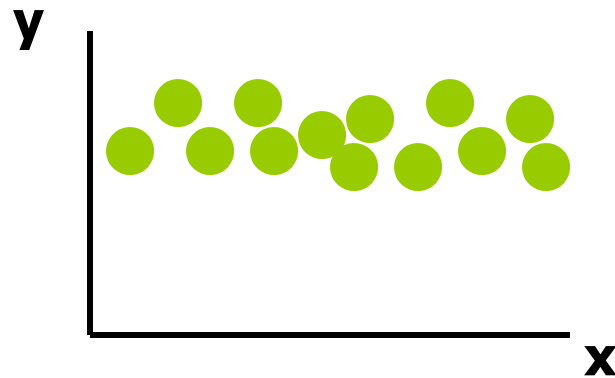
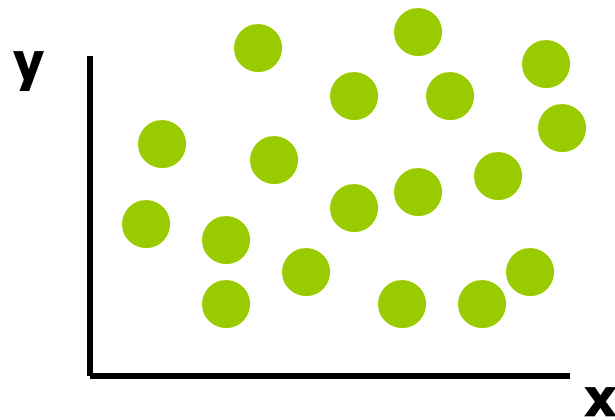
*(continued)*



# Scatter Plot Examples

*(continued)*

**No relationship**



# Correlation Coefficient

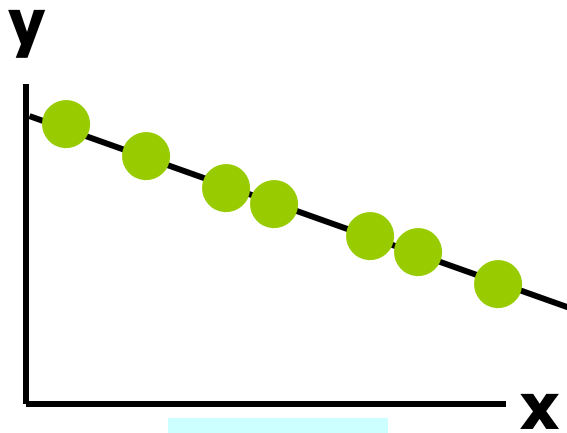
*(continued)*

- The **population correlation coefficient**  $\rho$  (rho) measures the strength of the association between the variables
- The **sample correlation coefficient**  $r$  is an estimate of  $\rho$  and is used to measure the strength of the linear relationship in the sample observations

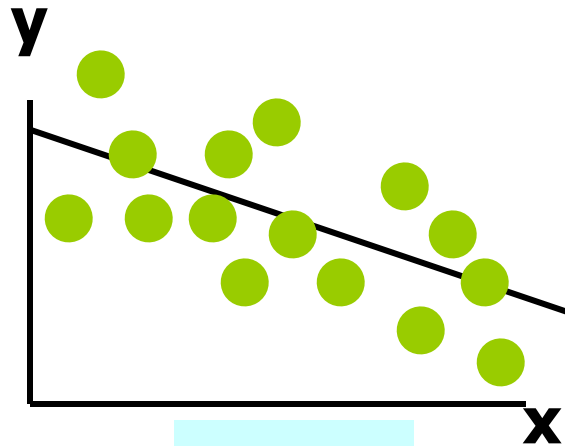
# Features of $\rho$ and $r$

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

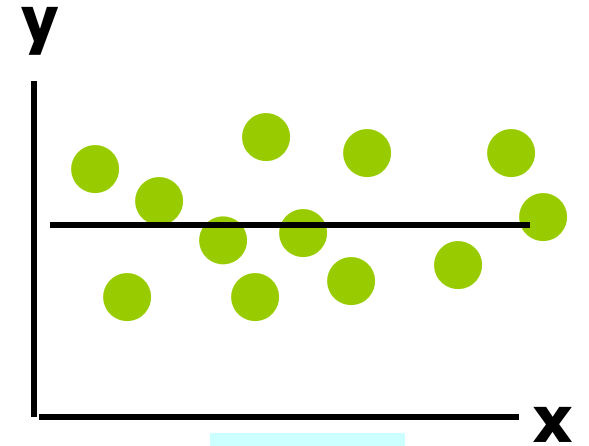
# Examples of Approximate r Values



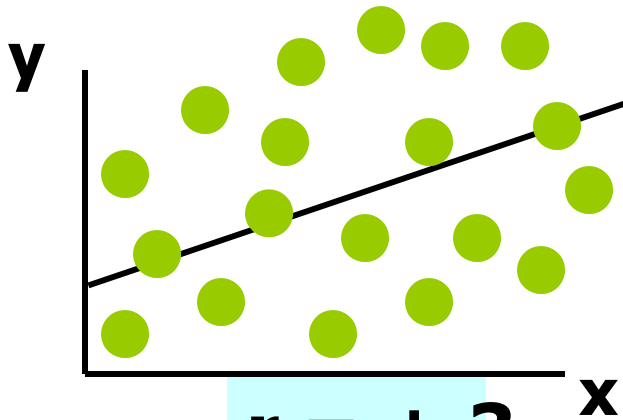
**$r = -1$**



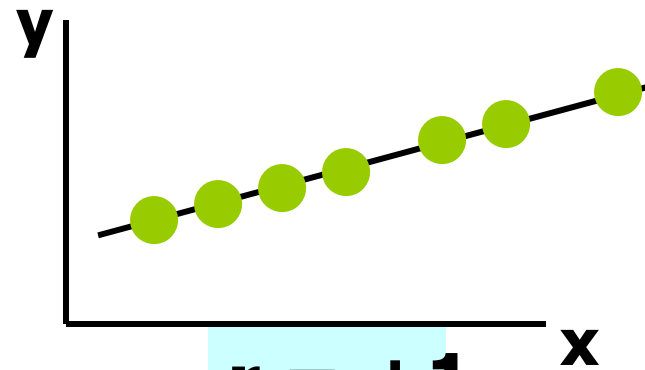
**$r = -.6$**



**$r = 0$**



**$r = +.3$**



**$r = +1$**

# Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

# Calculation Example

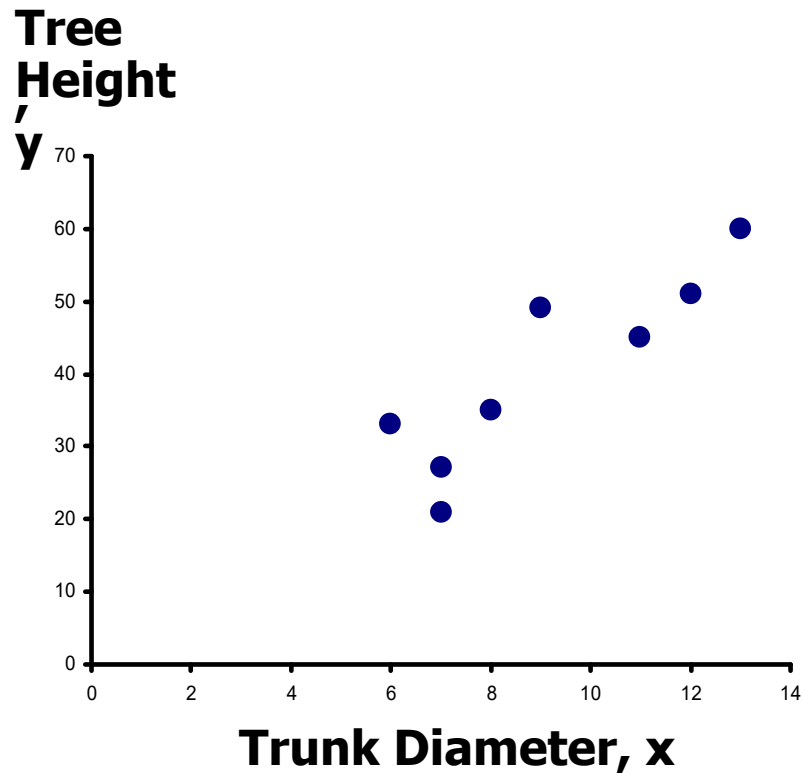


Tree Height	Trunk Diameter			
$y$	$x$	$xy$	$y^2$	$x^2$
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\Sigma=321$	$\Sigma=73$	$\Sigma=3142$	$\Sigma=14111$	$\Sigma=713$



# Calculation Example

*(continued)*



$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$
$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$
$$= 0.886$$

**r = 0.886** → relatively strong positive

linear association between x and y



# Excel Output

## Excel Correlation Output

Tools / data analysis / correlation...

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between  
Tree Height and Trunk Diameter



# Significance Test for Correlation

- Hypotheses

$$H_0: \rho = 0 \quad (\text{no correlation})$$

$$H_A: \rho \neq 0 \quad (\text{correlation exists})$$

- Test statistic

-

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with  $n - 2$  degrees of freedom)



## Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

$H_0: \rho = 0$  (No correlation)

$H_1: \rho \neq 0$  (correlation exists)

$$\alpha = .05, \quad df = 8 - 2 = 6$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$



## Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

$$\text{d.f.} = 8-2 \\ = 6$$



**Decision:**  
Reject  $H_0$

**Conclusion:**  
There **is** **evidence** of a linear relationship at the 5% level of significance

# Introduction to Regression Analysis

- **Regression analysis** is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

**Dependent variable:** the variable we wish to explain

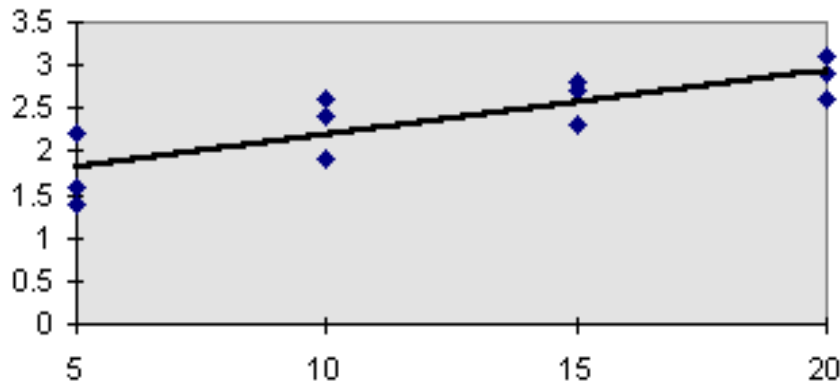
**Independent variable:** the variable used to explain the dependent variable

# Simple Linear Regression Model

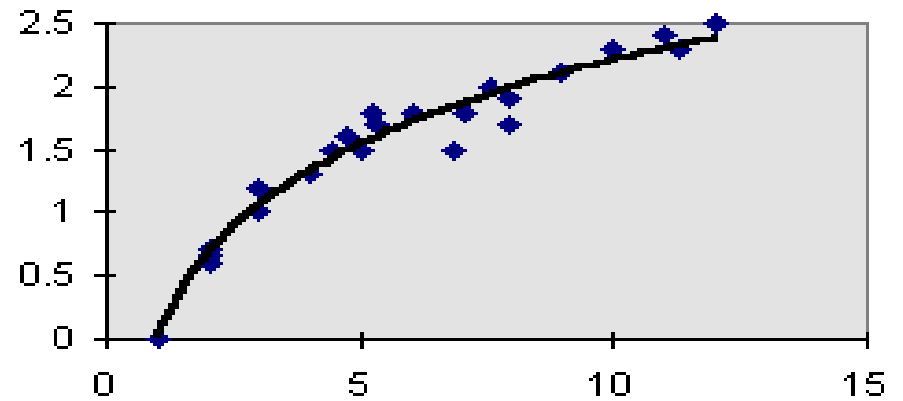
- Only **one independent variable**,  $x$
- Relationship between  $x$  and  $y$  is described by a linear function
- Changes in  $y$  are assumed to be caused by changes in  $x$

# Types of Regression Models

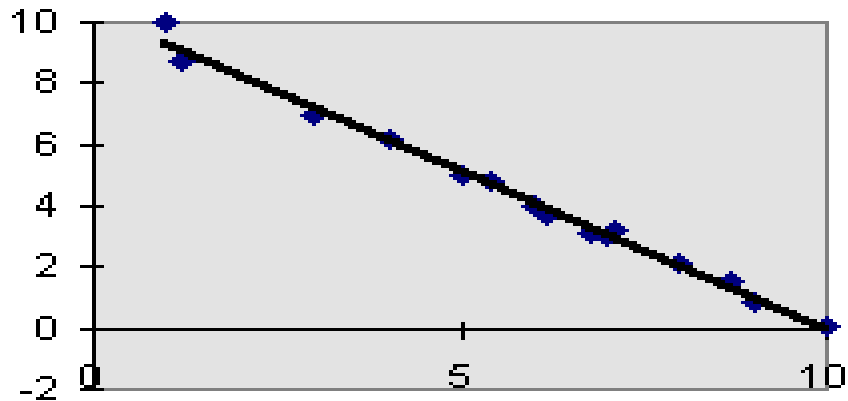
## Positive Linear Relationship



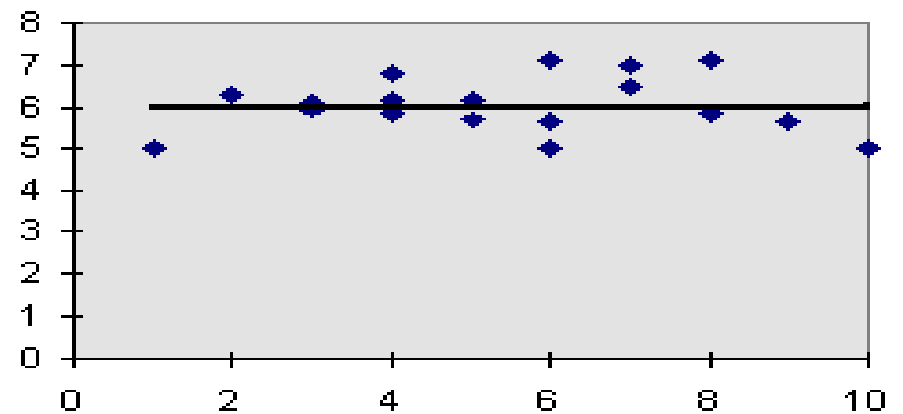
## Relationship NOT Linear



## Negative Linear Relationship



## No Relationship





# Population Linear Regression

The population regression model:

The diagram illustrates the population linear regression model,  $y = \beta_0 + \beta_1 X + \epsilon$ , with the following components labeled:

- Dependent Variable:**  $y$
- Population  $y$  intercept:**  $\beta_0$
- Population Slope Coefficient:**  $\beta_1$
- Independent Variable:**  $X$
- Random Error term, or residual:**  $\epsilon$

The model is also categorized into two main components:

- Linear component:**  $\beta_0 + \beta_1 X$
- Random Error component:**  $\epsilon$

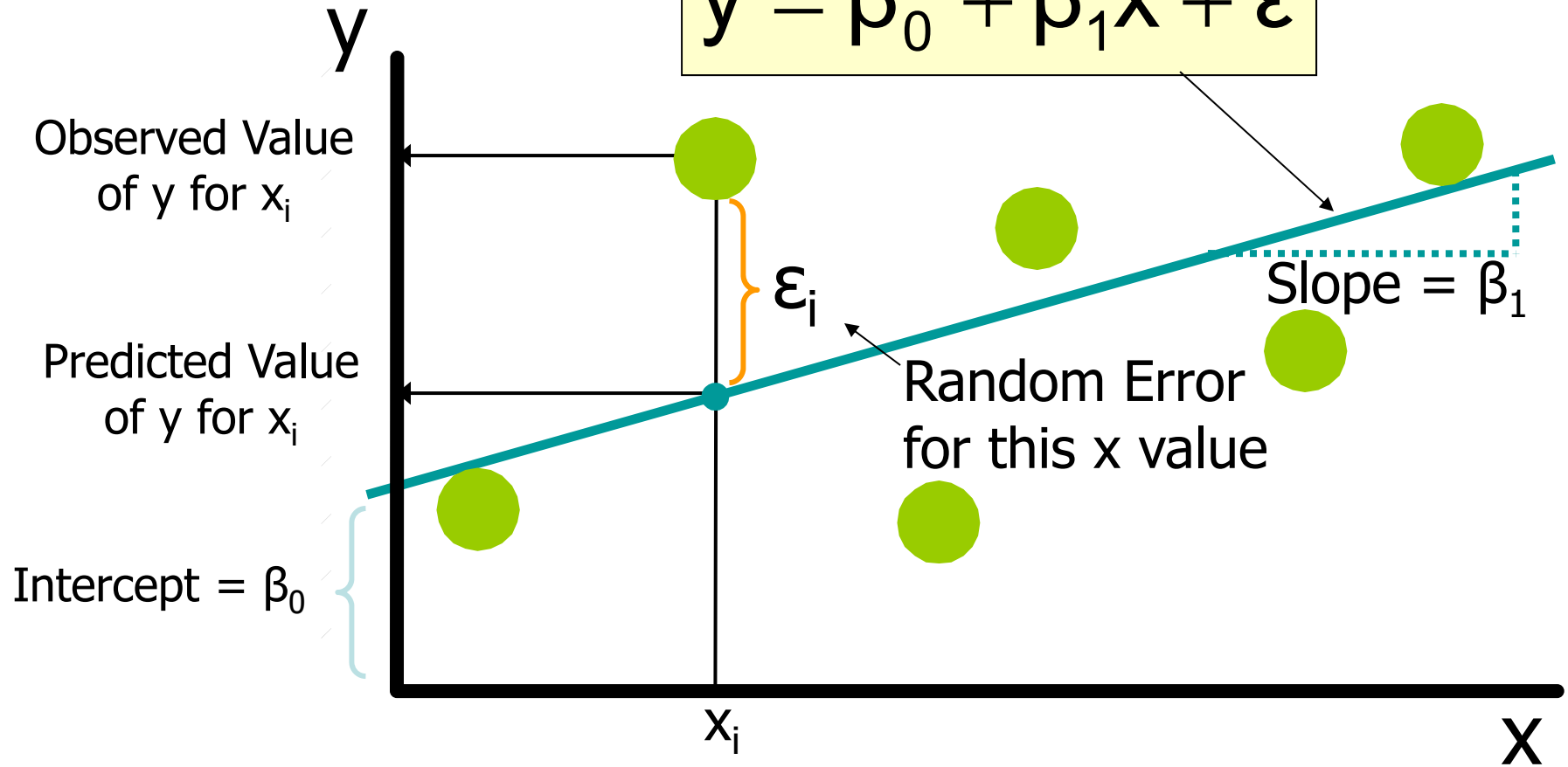
# Linear Regression Assumptions

- Error values ( $\epsilon$ ) are statistically independent
- Error values are normally distributed for any given value of  $x$
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the  $x$  variable and the  $y$  variable is linear

# Population Linear Regression

*(continued)*

$$y = \beta_0 + \beta_1 x + \varepsilon$$



# Estimated Regression Model

The sample regression line provides an **estimate** of the population regression line

Estimated  
(or predicted)  
y value

Estimate of  
the regression  
intercept

Estimate of the  
regression slope

Independent  
variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms  $e_i$  have a mean of zero

# Least Squares Criterion

- $b_0$  and  $b_1$  are obtained by finding the values of  $b_0$  and  $b_1$  that minimize the sum of the squared residuals

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

# The Least Squares Equation

- The formulas for  $b_1$  and  $b_0$  are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

algebraic

equivalent

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

## Interpretation of the Slope and the Intercept

- $b_0$  is the estimated average value of  $y$  when the value of  $x$  is zero
- $b_1$  is the estimated change in the average value of  $y$  as a result of a one-unit change in  $x$

# Finding the Least Squares Equation

- The coefficients  $b_0$  and  $b_1$  will usually be found using computer software, such as Excel or Minitab
- Other regression measures will also be computed as part of computer-based regression analysis



# Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected

– Dependent variable ( $y$ ) = house price in \$1000s

– Independent variable ( $x$ ) = square feet

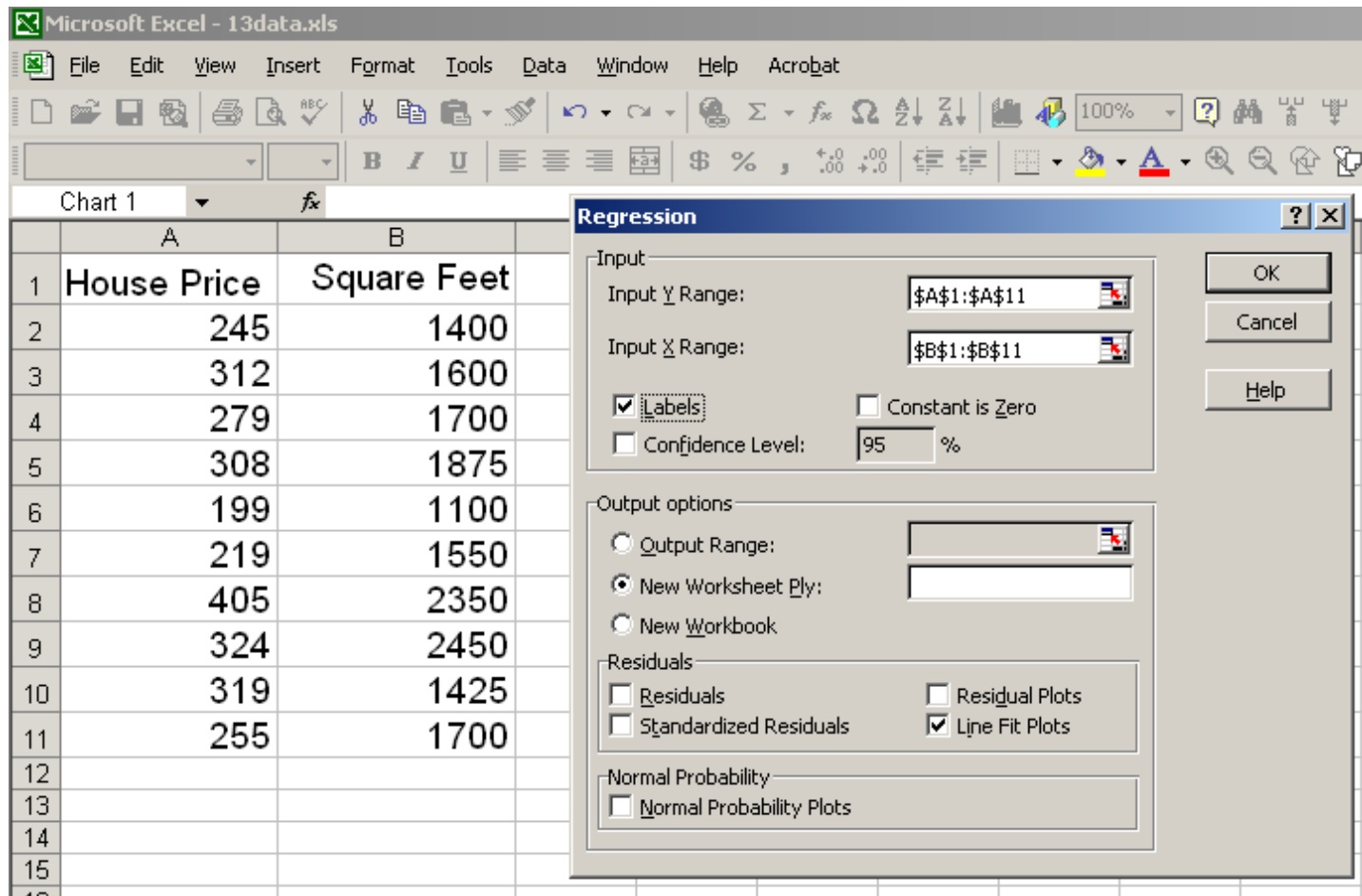
# Sample Data for House Price Model

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



# Regression Using Excel

- Tools / Data Analysis / Regression



The screenshot shows the Microsoft Excel interface with a data table and the Regression dialog box open. The data table has two columns: House Price and Square Feet. The Regression dialog box is configured with the following settings:

Input	Value
Input Y Range:	\$A\$1:\$A\$11
Input X Range:	\$B\$1:\$B\$11
Labels:	<input checked="" type="checkbox"/>
Confidence Level:	95 %
Constant is Zero:	<input type="checkbox"/>

Output options:

- Output Range:
- New Worksheet Ply:
- New Workbook:

Residuals:

- Residuals
- Standardized Residuals
- Residual Plots
- Line Fit Plots

Normal Probability:

- Normal Probability Plots



# Excel Output

## Regression Statistics

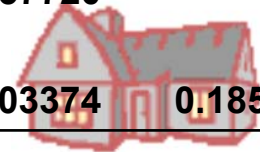
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

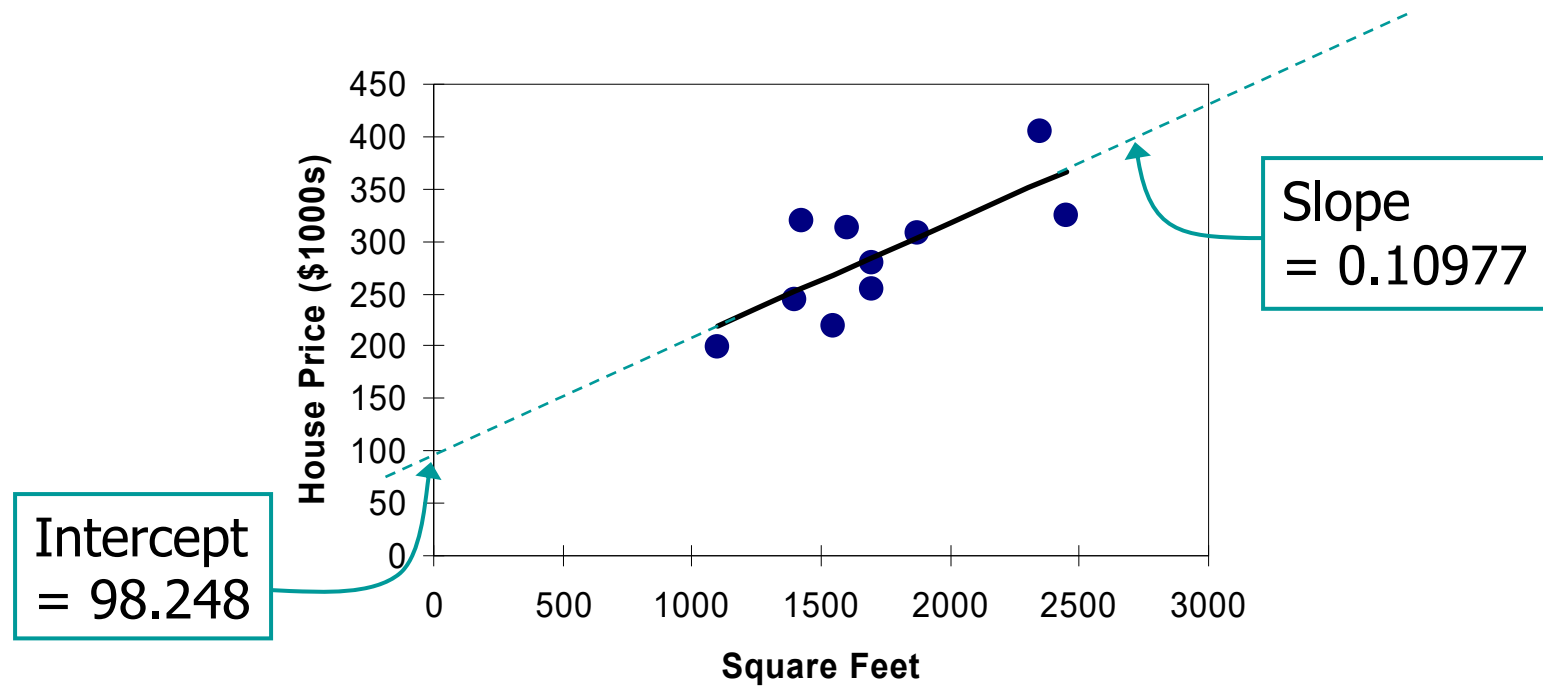
ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.084	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.1289	-35.57720	232.0738
Square Feet	0.10977	0.03297	3.32938	0.0103	0.03374	0.18580



# Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

## Interpretation of the Intercept, $b_0$

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

- $b_0$  is the estimated average value of Y when the value of X is zero (if  $x = 0$  is in the range of observed x values)
  - Here, no houses had 0 square feet, so  $b_0 = 98.24833$  just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet



## Interpretation of the Slope Coefficient, $b_1$

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- $b_1$  measures the estimated change in the average value of  $Y$  as a result of a one-unit change in  $X$

– Here,  $b_1 = .10977$  tells us that the average value of a house increases by  $.10977(\$1000) = \$109.77$ , on average, for each additional one square foot of size



## Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 (  $\sum (y - \hat{y}) = 0$  )
- The sum of the squared residuals is a minimum (minimized  $\sum (y - \hat{y})^2$  )
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of  $\beta_0$  and  $\beta_1$



# Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum  
of Squares

Sum of  
Squares Error

Sum of  
Squares

Regression

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

$$SSR = \sum (\hat{y} - \bar{y})^2$$

where:

$\bar{y}$  = Average value of the dependent variable

$y$  = Observed values of the dependent variable

$\hat{y}$  = Estimated value of  $y$  for the given  $x$  value

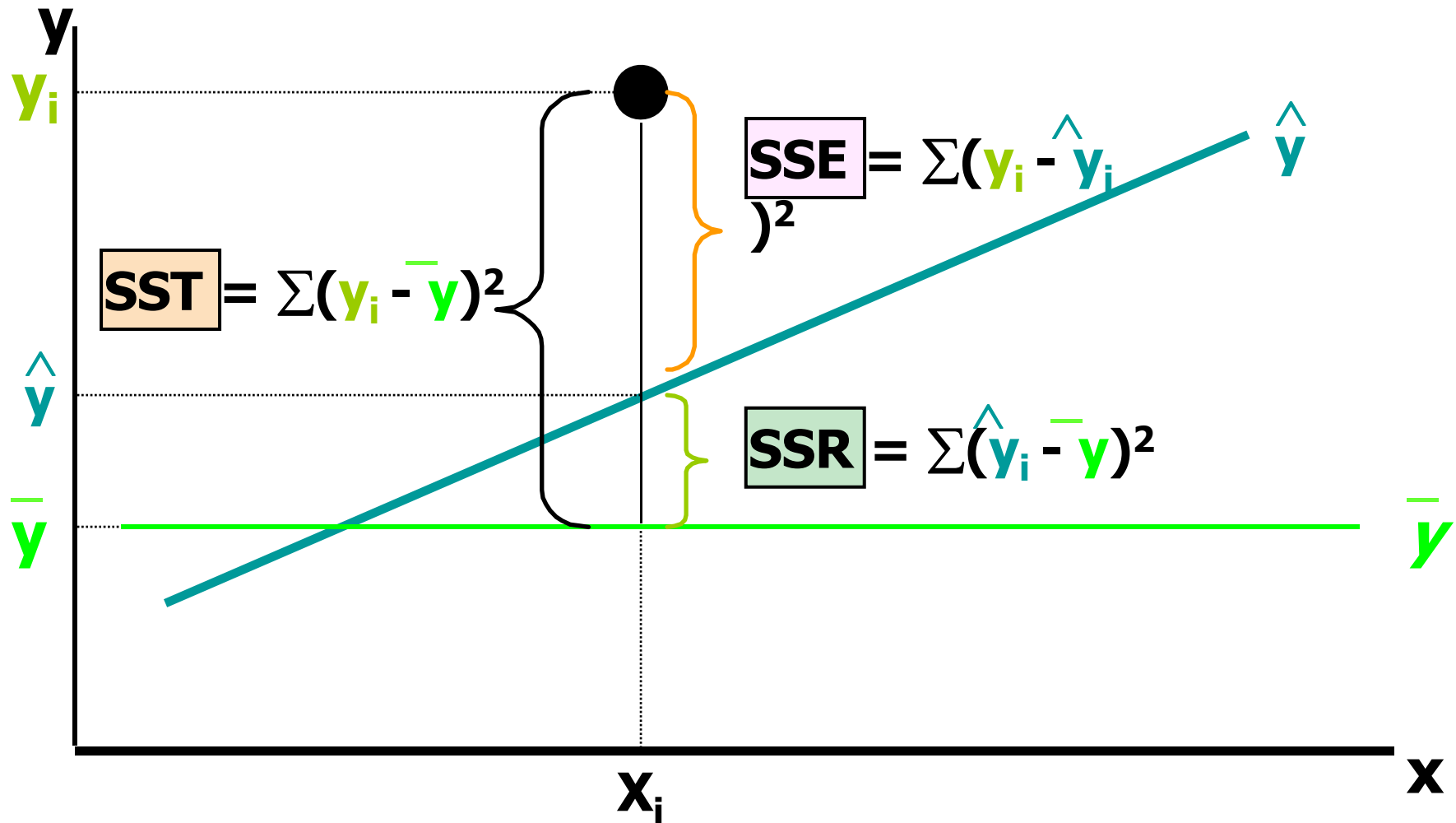
# Explained and Unexplained Variation

*(continued)*

- **SST = total sum of squares**
  - Measures the variation of the  $y_i$  values around their mean  $y$
- **SSE = error sum of squares**
  - Variation attributable to factors other than the relationship between  $x$  and  $y$
- **SSR = regression sum of squares**
  - Explained variation attributable to the relationship between  $x$  and  $y$

# Explained and Unexplained Variation

*(continued)*



# Coefficient of Determination, $R^2$

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **R-squared** and is denoted as  $R^2$

$$R^2 = \frac{SSR}{SST}$$

where

$$0 \leq R^2 \leq 1$$

# Coefficient of Determination, $R^2$

*(continued)*

## Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

**Note:** In the single independent variable case, the coefficient of determination is

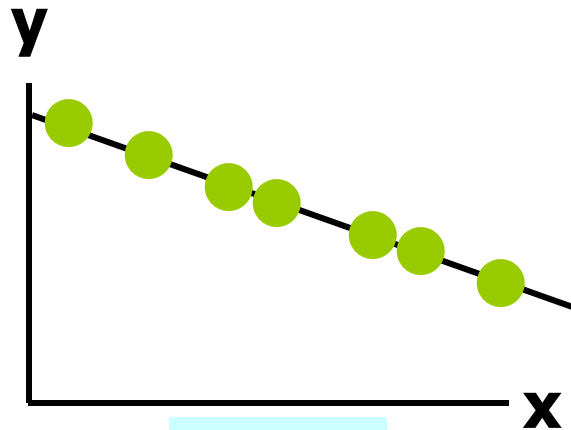
$$R^2 = r^2$$

where:

$R^2$  = Coefficient of determination

$r$  = Simple correlation coefficient

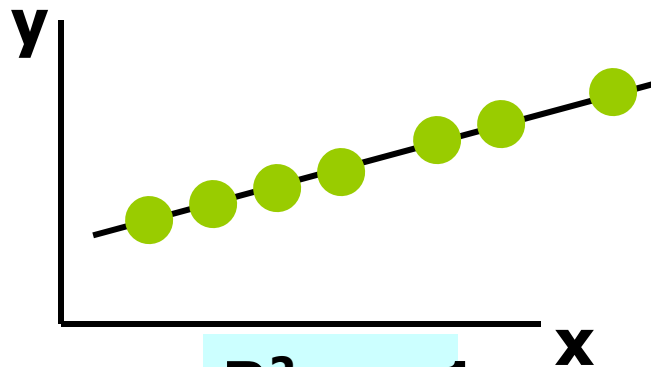
# Examples of Approximate $R^2$ Values



$$R^2 = 1$$

$$R^2 = 1$$

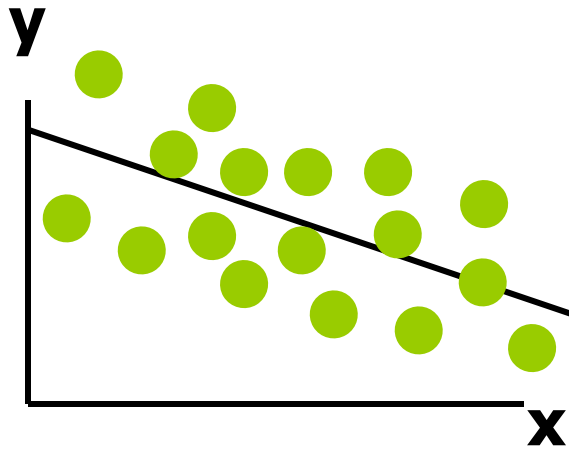
**Perfect linear relationship between x and y:**



$$R^2 = +1$$

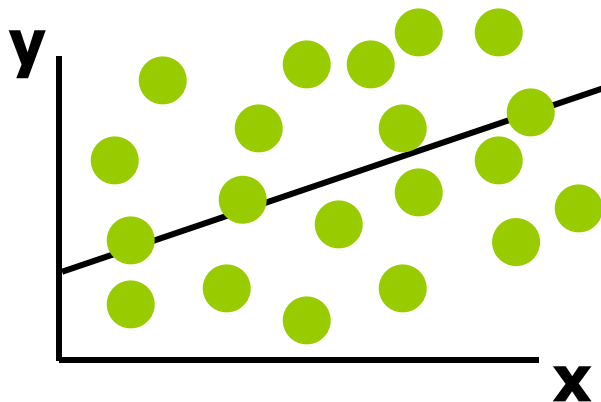
**100% of the variation in y is explained by variation in x**

# Examples of Approximate $R^2$ Values



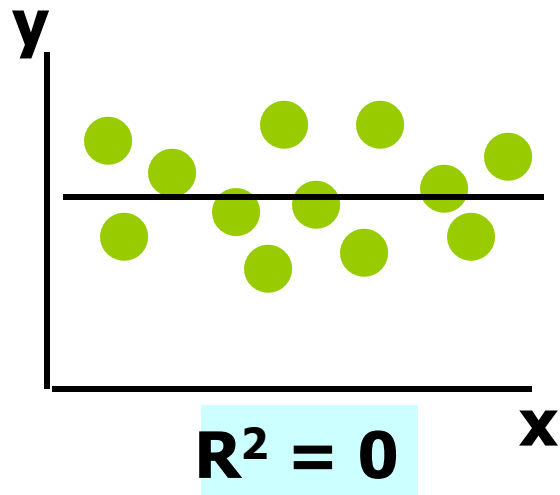
$$0 < R^2 < 1$$

**Weaker linear relationship between x and y:**



**Some but not all of the variation in y is explained by variation in x**

## Examples of Approximate $R^2$ Values



$$R^2 = 0$$

**No linear relationship between x and y:**

**The value of Y does not depend on x. (None of the variation in y is explained by variation in x)**



# Excel Output

## Regression Statistics

Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

## ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

# Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n - k - 1}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

# The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient ( $b_1$ ) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

$s_{b_1}$  = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$  = Sample standard error of the estimate

# Excel Output

## Regression Statistics

Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

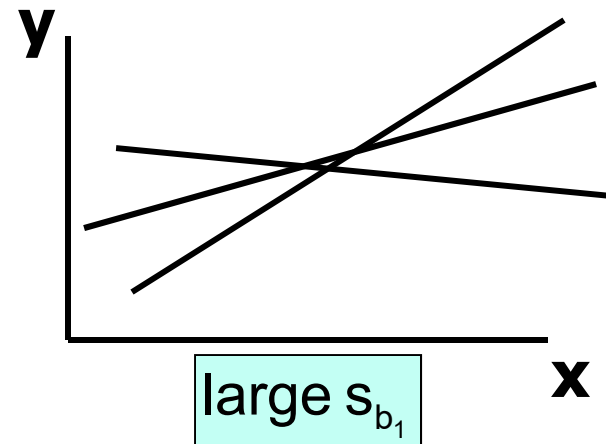
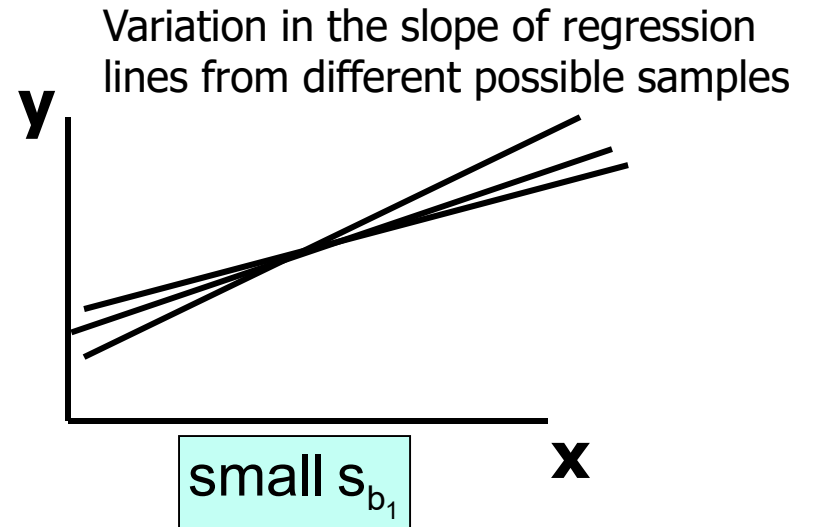
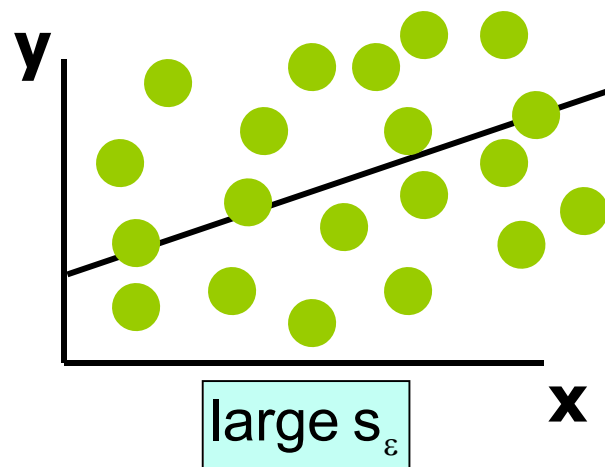
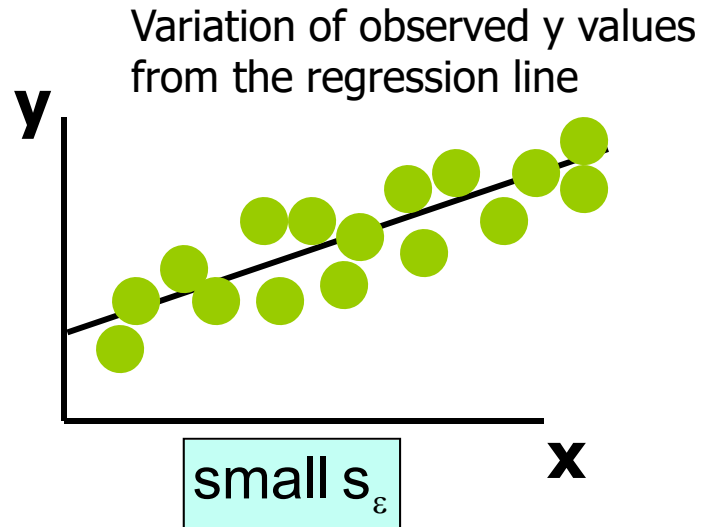
$$s_{\varepsilon} = 41.33032$$

$$s_{b_1} = 0.03297$$

ANOVA	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

# Comparing Standard Errors



# Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - $H_0: \beta_1 = 0$  (no linear relationship)
  - $H_1: \beta_1 \neq 0$  (linear relationship does exist)
- Test statistic

–

–

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$\text{d.f.} = n - 2$$

where:

$b_1$  = Sample regression slope coefficient

$\beta_1$  = Hypothesized slope

$s_{b_1}$  = Estimator of the standard error of the slope

# Inference about the Slope: t Test

*(continued)*

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

## Estimated Regression

### Equation:

$$\text{house price} = 98.25 + 0.1098 (\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house  
affect its sales price?

# Inferences about the Slope: t Test Example

Test Statistic: **t = 3.329**

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

From Excel output:

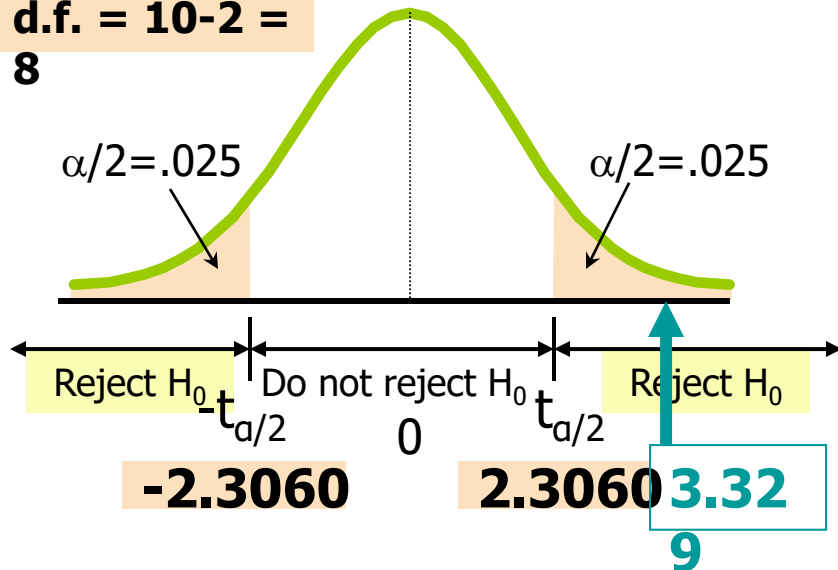
	Coefficients	Standard Error	t Stat	P-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

$b_1$

$s_{b_1}$

t

d.f. = 10-2 = 8



**Decision:**

Reject  $H_0$

**Conclusion:**

There is sufficient evidence that square footage affects house price



# Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

$$d.f. = n - 2$$

Excel Printout for House Prices:

	<i>Coefficient</i> <i>s</i>	<i>Standard</i> <i>Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper</i> <i>95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

## Regression Analysis for Description

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is \$1000s, we are 95% confident that the average impact on sales price is between \$33.70 and \$185.80 per square foot of house size

This 95% confidence interval **does not include 0**.

**Conclusion:** There is a significant relationship between house price and square feet at the .05 level of significance

# Confidence Interval for the Average $y$ , Given $x$

Confidence interval estimate for the **mean of  $y$**  given a particular  $x_p$

Size of interval varies according to distance away from mean,  $\bar{x}$

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

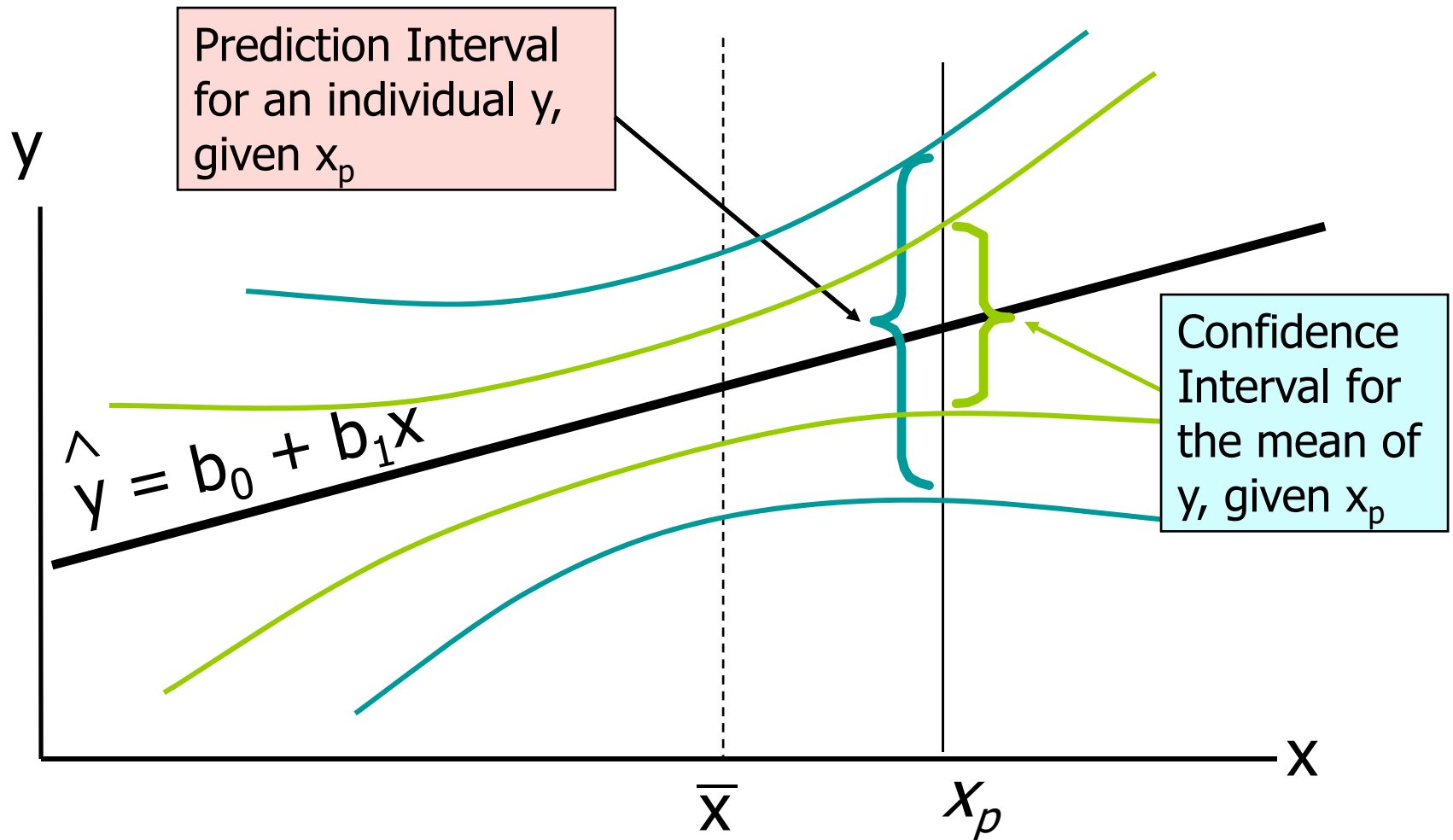
# Confidence Interval for an Individual $y$ , Given $x$

Confidence interval estimate for an **Individual value of  $y$**  given a particular  $x_p$

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

# Interval Estimates for Different Values of $x$



## Example: House Prices

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

### Estimated Regression

#### Equation:

$$\text{house price} = 98.25 + 0.1098 (\text{sq.ft.})$$

Predict the price for a house  
with 2000 square feet

## Example: House Prices

*(continued)*

Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098 (\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is  $317.85(\$1,000\text{s}) = \$317,850$

## Estimation of Mean Values: Example

Confidence Interval Estimate for  $E(y)|x_p$

Find the 95% confidence interval for the average price of 2,000 square-foot houses

Predicted Price  $\hat{Y}_i = 317.85$  (\$1,000s)

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 -- 354.90, or from \$280,660 -- \$354,900



## Estimation of Individual Values: Example

Prediction Interval Estimate for  $y|x_p$

Find the 95% confidence interval for an individual house with 2,000 square feet

Predicted Price  $\hat{Y}_i = 317.85$  (\$1,000s)

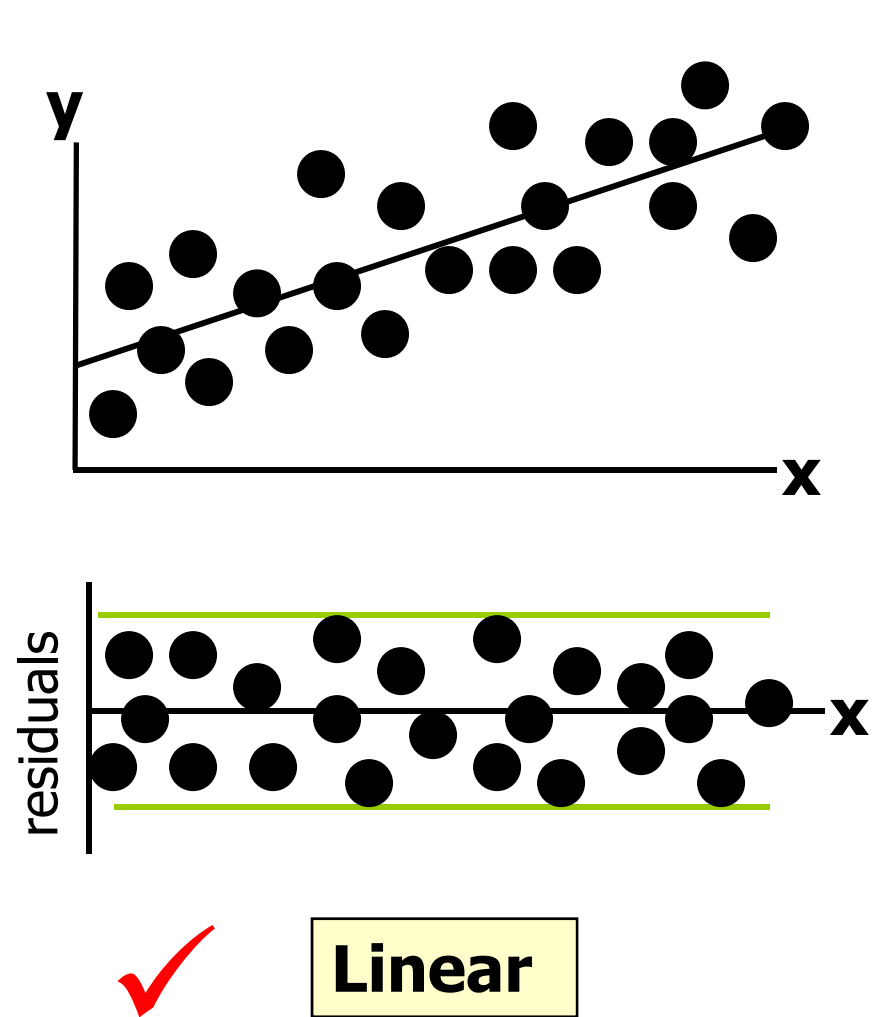
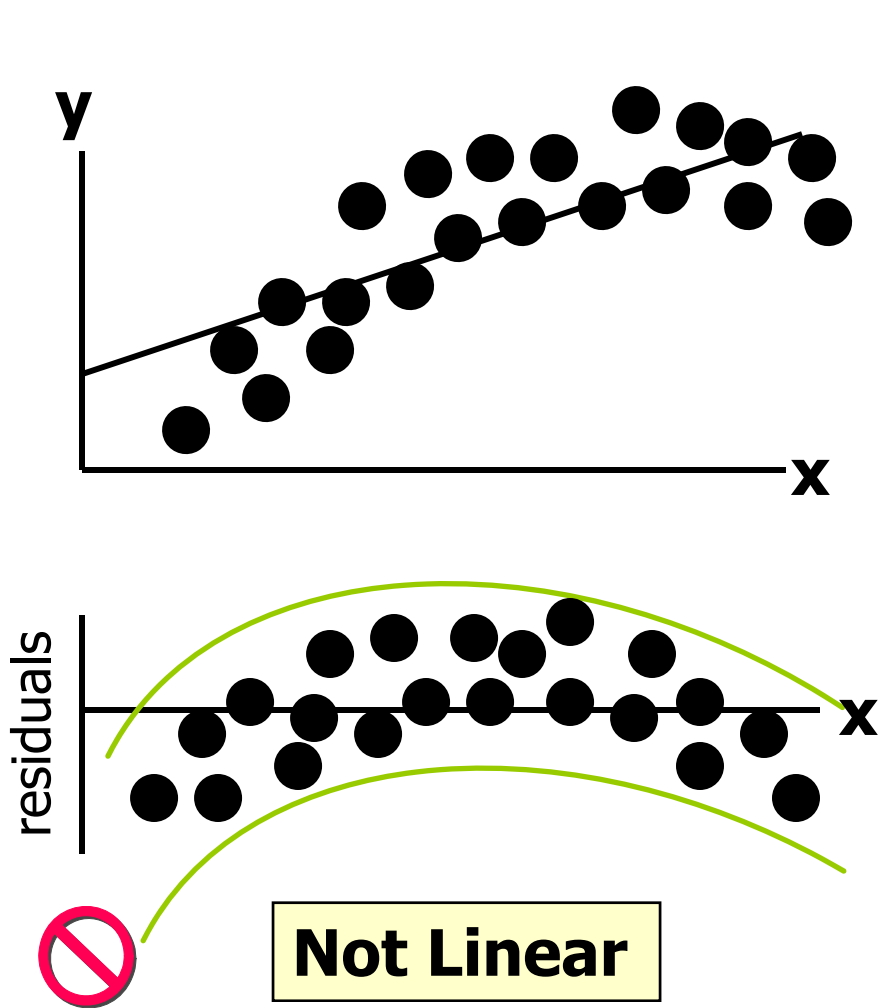
$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 -- 420.07, or from \$215,500 -- \$420,070

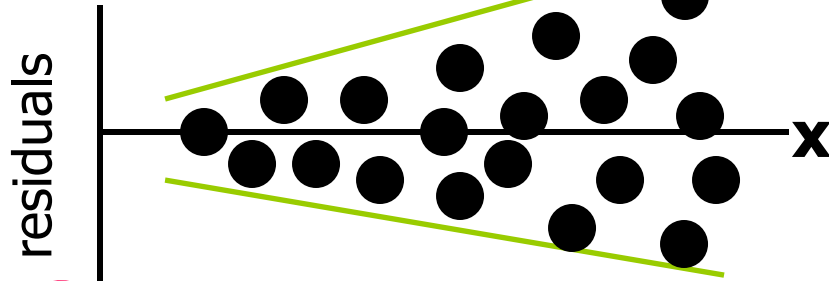
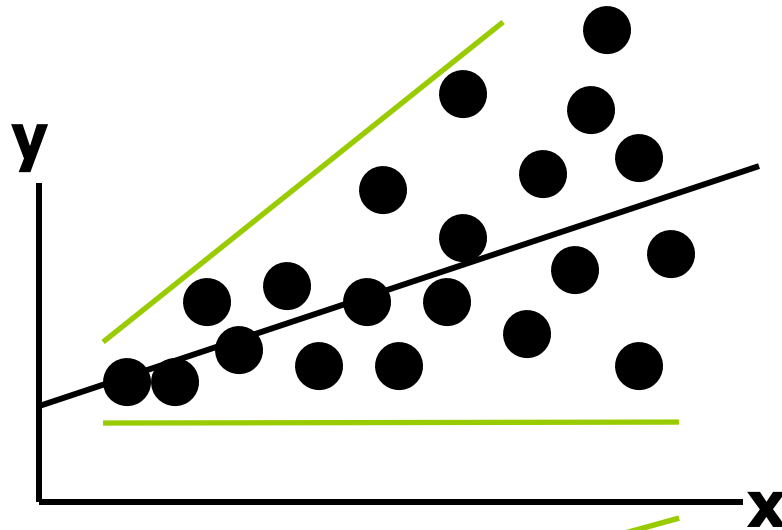
# Residual Analysis

- Purposes
  - Examine for linearity assumption
  - Examine for constant variance for all levels of  $x$
  - Evaluate normal distribution assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs.  $x$
  - Can create histogram of residuals to check for normality

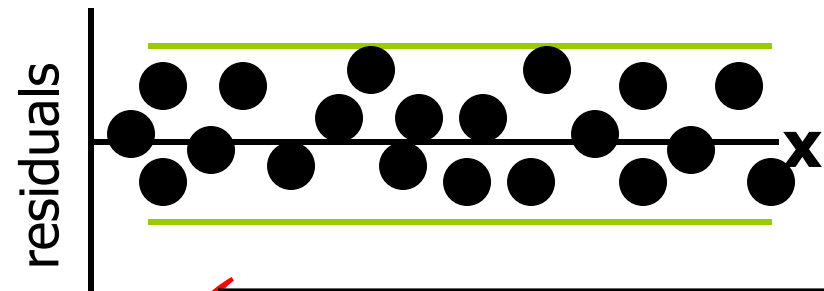
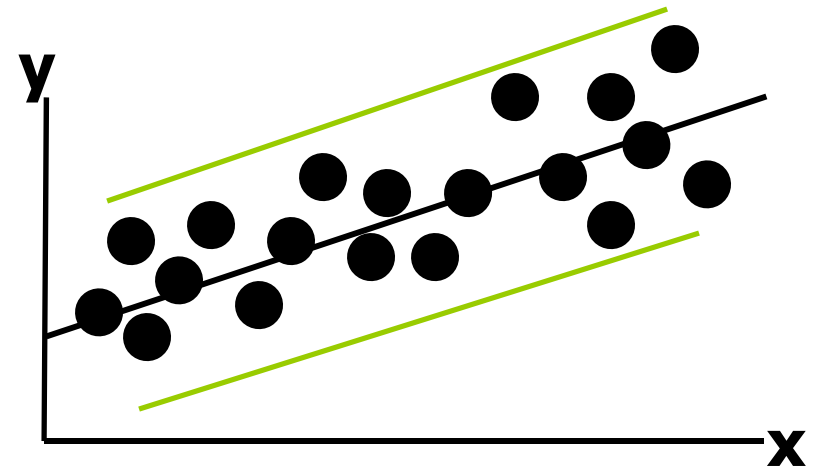
# Residual Analysis for Linearity



# Residual Analysis for Constant Variance



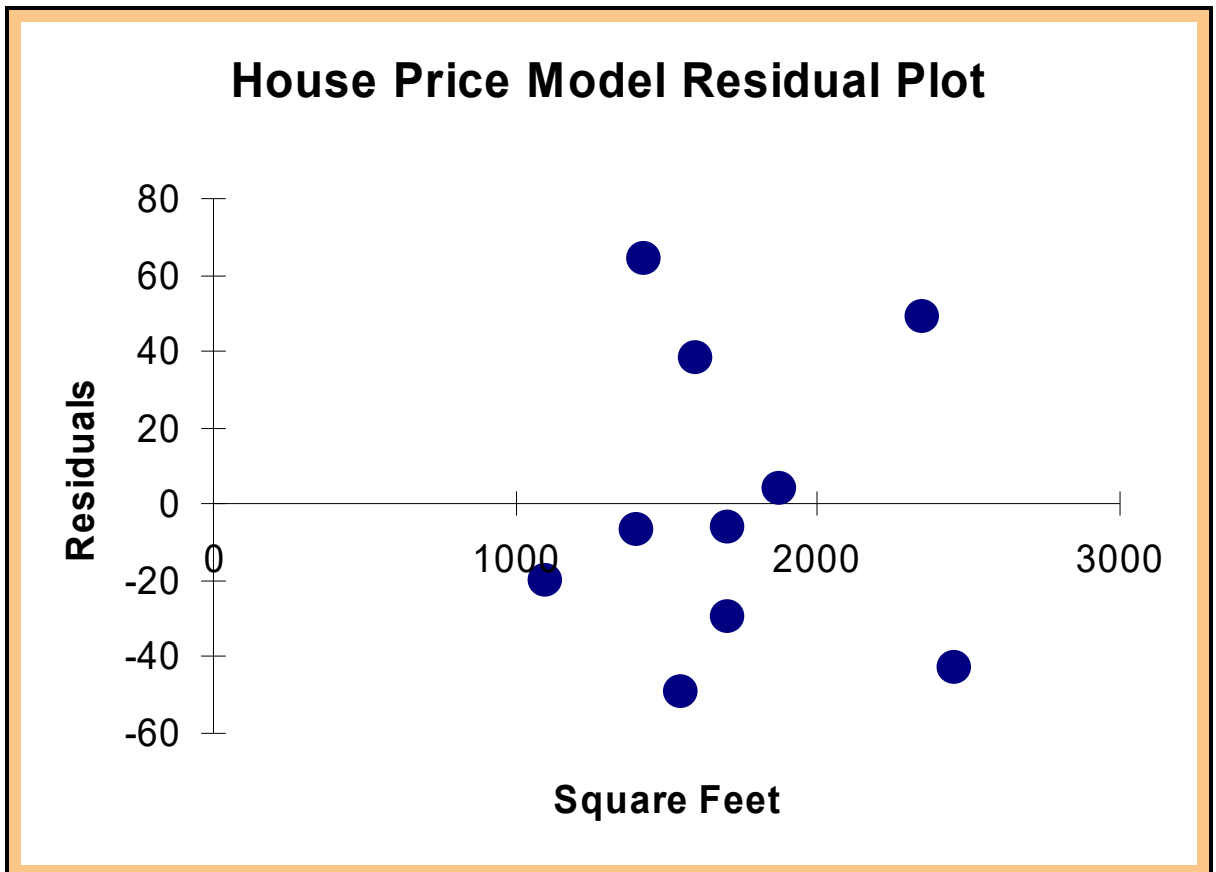
Non-constant variance



Constant variance

# Excel Output

RESIDUAL OUTPUT		
	<i>Predicted House Price</i>	<i>Residuals</i>
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



# Summary

- Introduced correlation analysis
- Discussed correlation to measure the strength of a linear association
- Introduced simple linear regression analysis
- Calculated the coefficients for the simple linear regression equation
- measures of variation ( $R^2$  and  $s_\epsilon$ )
- Addressed assumptions of regression and correlation

# Summary

*(continued)*

- Described inference about the slope
- Addressed estimation of mean values and prediction of individual values
- Discussed residual analysis

## Example

- Daytime SBP (systolic blood pressure) and age collected for 447 hypertensive males.

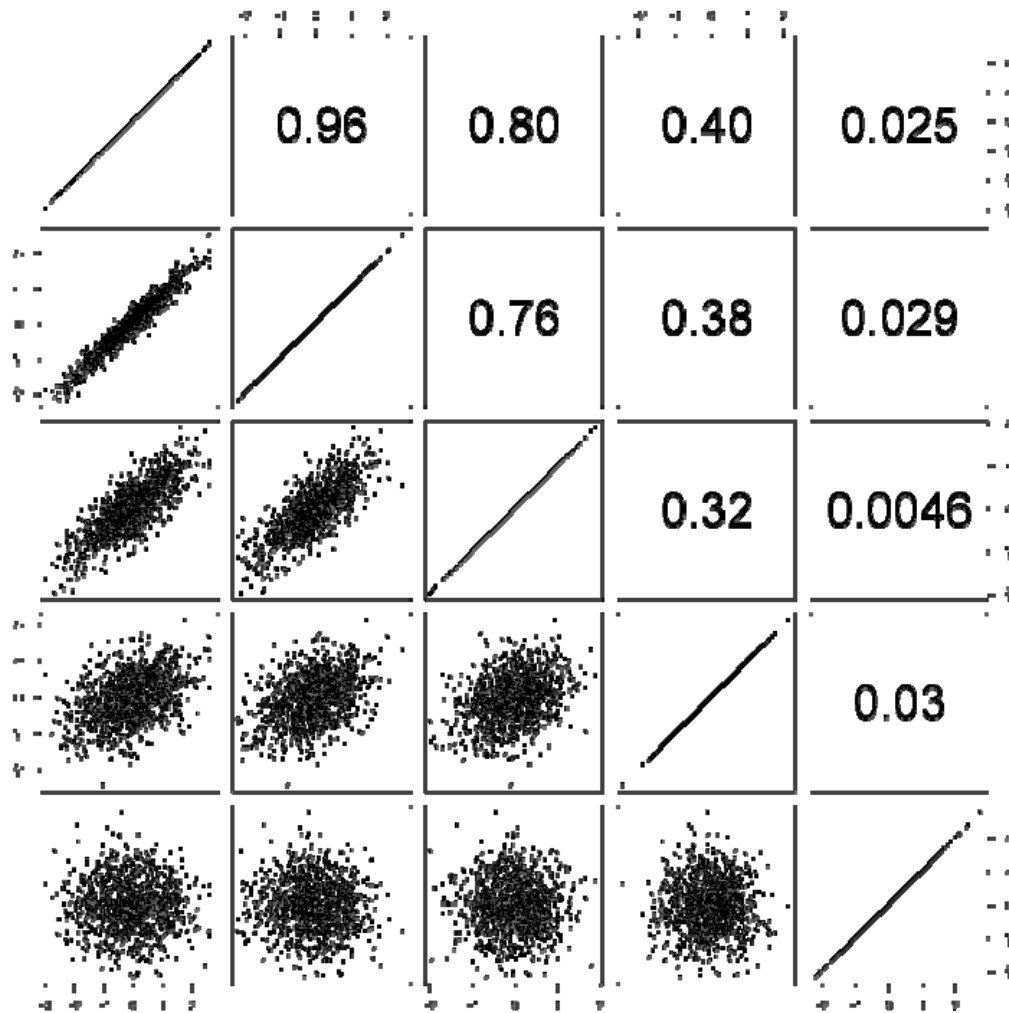
SBP	Age
<b>115</b>	<b>34</b>
<b>130</b>	<b>40</b>
<b>128</b>	<b>28</b>
<b>123</b>	<b>21</b>
<b>126</b>	<b>39</b>
...	...



## Example (contd)

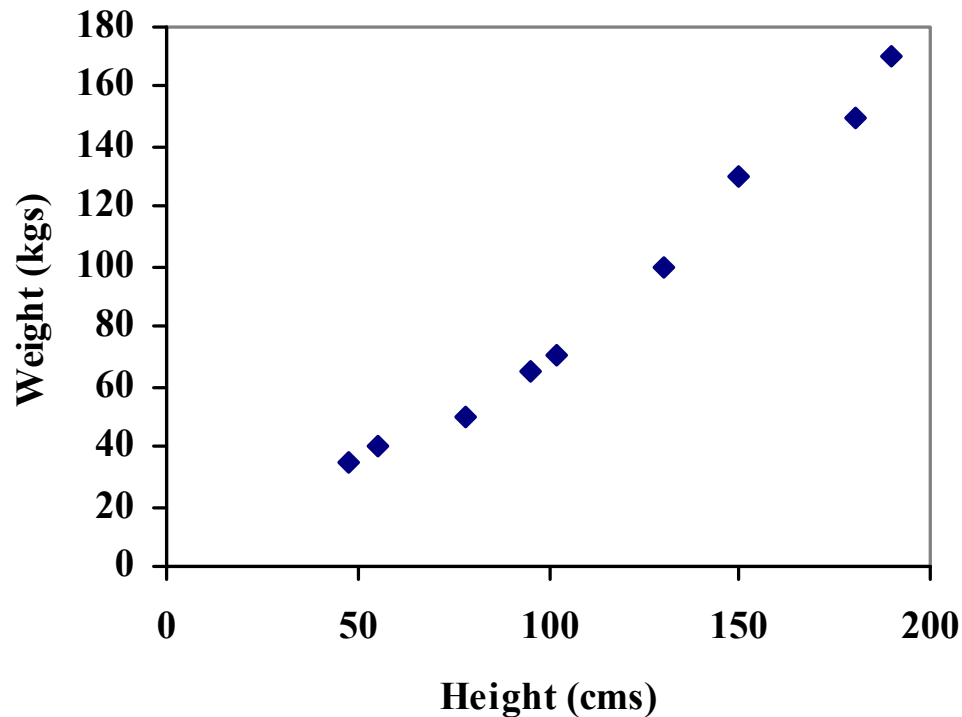
- Is there a linear relationship between SBP and Age?
- $r=0.145 \rightarrow$  weak positive relationship

# Correlation examples



## Example 2: Height vs. Weight

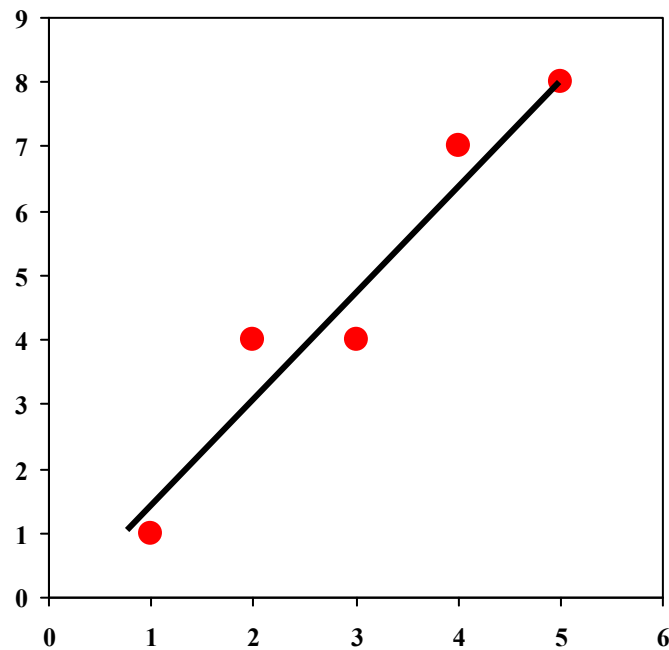
*Graph One: Relationship between Height and Weight*



- Strong positive correlation between height and weight
- Can see how the relationship works, but cannot predict one from the other
- If 120cm tall, then how heavy?

# Regression

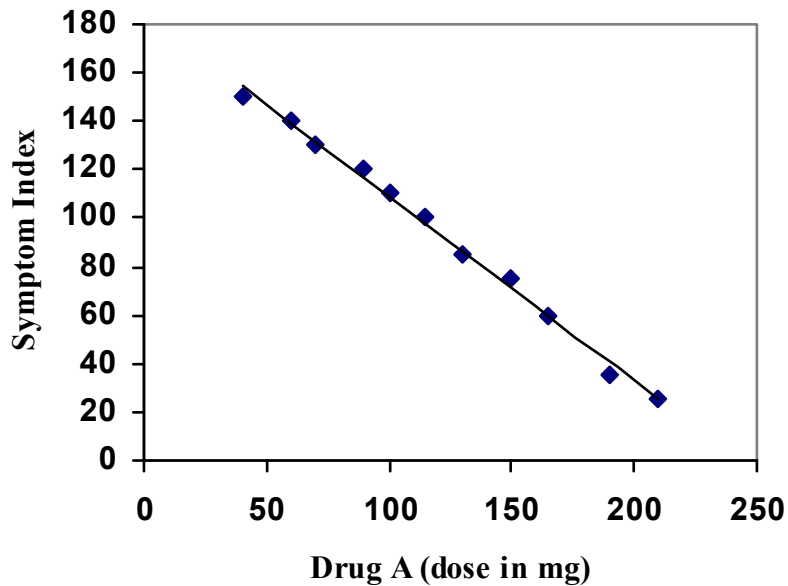
**Problem: to draw a straight line through the points that best explains the variance**



Line can then be used to predict Y from X

# Example: Symptom Index vs Drug A

*Graph Three: Relationship between Symptom Index and Drug A (with best-fit line)*



- “Best fit line”
- allows us to describe relationship between variables more accurately.
- We can now predict specific values of one variable from knowledge of the other
- All points are close to the line

# Simple Linear Regression

- Assume the population regression line:

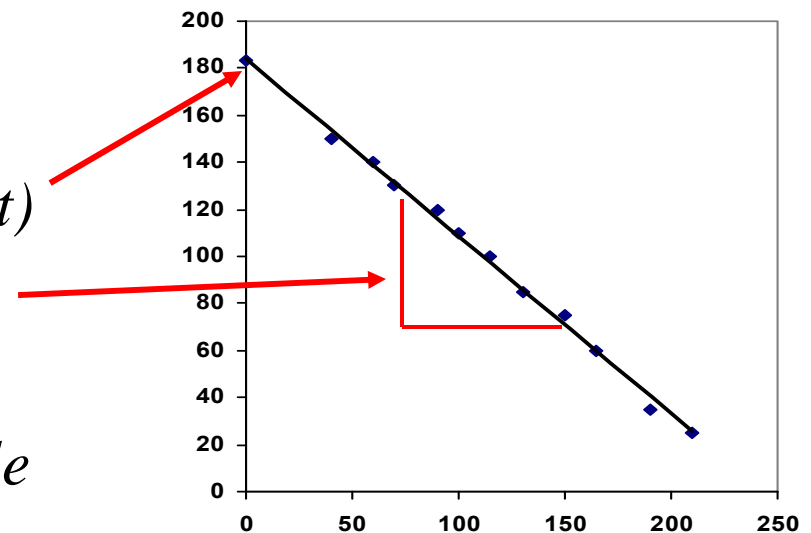
$$y = \alpha + \beta x$$

Where:  $\alpha$  = *y intercept (constant)*

$\beta$  = *slope of line*

$y$  = *dependent variable*

$x$  = *independent variable*



- $y_i = \alpha + \beta x_i + \varepsilon_i$

# Regression

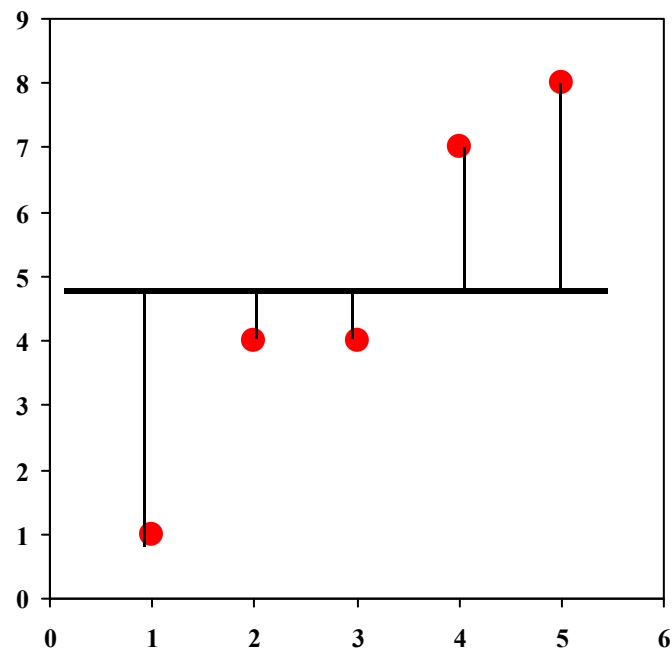
- Establish equation for the **best-fit line**:

$$y = a+bx$$

- Best-fit line same as **regression** line
- b is the **regression coefficient** for **x**
- x is the **predictor** or **regressor** variable for y

## Fit a line to the data:

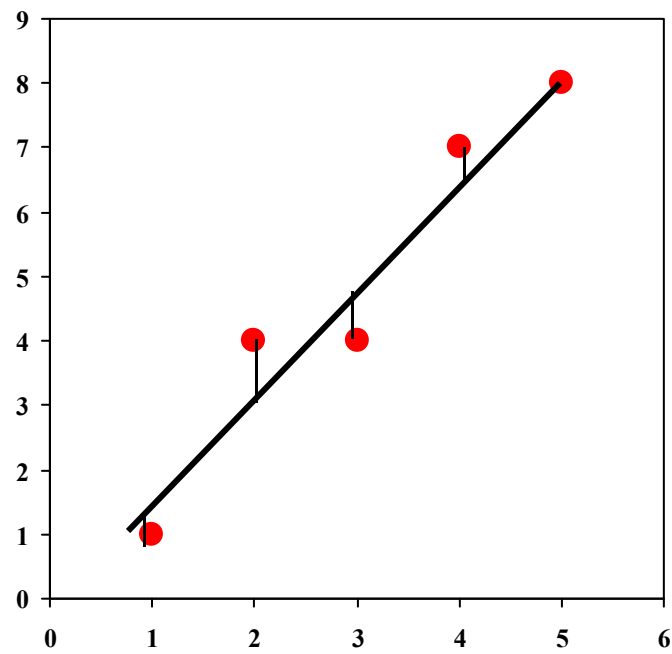
- Not great:





# Fit a line to the data:

- Better:



## Least Squares

- Minimise the (squared) distance between the points and the line
- $a$  and  $b$  are the estimates of  $\alpha$  and  $\beta$  which minimise

$$\sum \{y_i - (\alpha + \beta x_i)\}^2$$

## Least Squares Estimates

- Using calculus (partial derivatives), we get

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- Note  $b$  is related to the correlation coefficient  $r$  (same numerator)- if  $x$  and  $y$  are positively correlated then the slope is positive

## Example from the literature

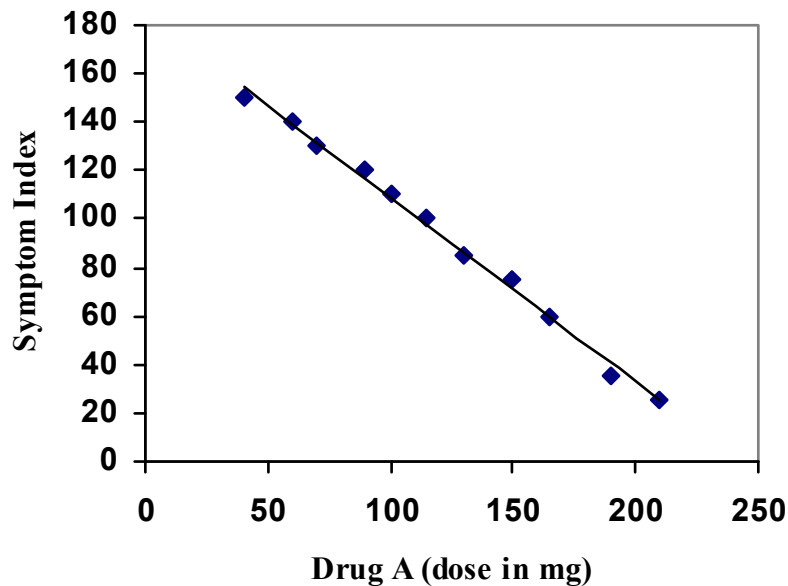
- Predicting tandem repeat variability

Regression coefficients for the VNTR prediction model

Variable	Coef ( $\beta$ )	SE	<i>p</i> value	Odds ratio	95% CI for odds ratio
Copy number <sup>a</sup>	2.69	0.57	<0.0001	14.8	4.784–45.621
Percentage match	0.288	0.068	<0.0001	17.8 <sup>b</sup>	4.732–66.779
Entropy	−7.87	2.91	0.0068	0.455 <sup>c</sup>	0.258–0.805
GC dinucleotide bias <sup>d</sup>	−1.53	0.65	0.0193	0.858 <sup>c</sup>	0.754–0.975
<i>Intercept</i>	−17.0	6.8	—	—	—

# Example: symptom Index versus Drug A dose

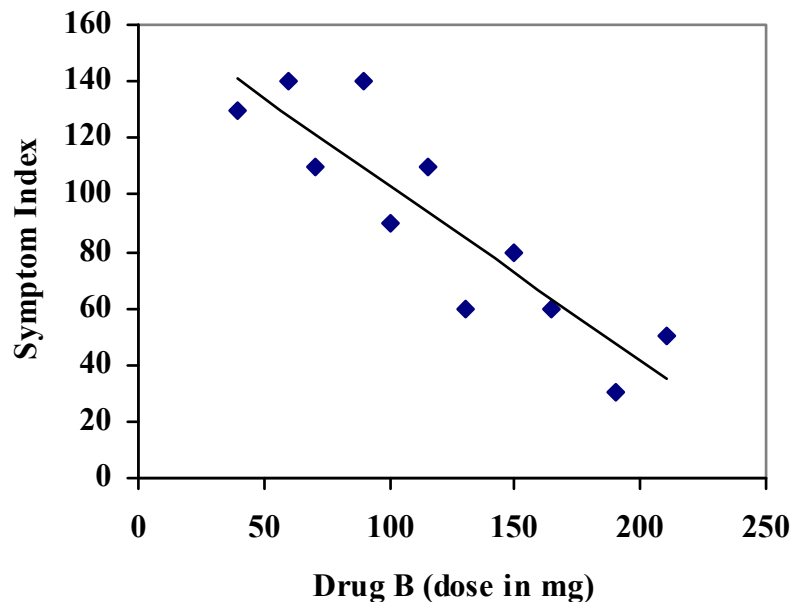
*Graph Three: Relationship between  
Symptom Index and Drug A  
(with best-fit line)*



- “Best fit line”
- Allows us to describe relationship between variables more accurately.
- We can now predict specific values of one variable from knowledge of the other
- All points are close to the line

# Example: Symptom Index versus Drug B dose

*Graph Four: Relationship between Symptom Index and Drug B (with best-fit line)*



- We can still predict specific values of one variable from knowledge of the other
- Will predictions be as accurate?
- Why not?
- Large “residual” variation (random error)
  - = Difference between **observed** data and that **predicted** by the equation

# Regression Hypothesis Tests

- Hypotheses about the intercept

$$H_0: \alpha = 0 \quad H_A: \alpha \neq 0$$

- But most research focuses on the slope

$$H_0: \beta = 0 \quad H_A: \beta \neq 0$$

This addresses the general question "Is X predictive of Y?"

# Regression

- Estimates of a slope ( $b$ ) have a sampling distribution, like any other statistic
  - If certain assumptions are met (NB normality, homogeneity of variance) the sampling distribution approximates the t-distribution
  - Thus, we can assess the probability that a given value of  $b$  would be observed, if  $\beta = 0$
- hypothesis tests & confidence intervals



# Regression

- $R^2$ , the **coefficient of determination**, is the percentage of variation explained by the “regression”.
- $R^2 > 0.6$  is deemed reasonably good.
- Note, the model must also be significant, e.g.

```
regress write female read math science socst
```

Source	SS	df	MS	Number of obs = 200	
Model	10756.9244	5	2151.38488	F( 5, 194)	= 58.60
Residual	7121.9506	194	36.7110855	Prob > F	= 0.0000
-----+-----				R-squared	= 0.6017
Total	17878.875	199	89.843593	Adj R-squared	= 0.5914
-----+-----				Root MSE	= 6.059

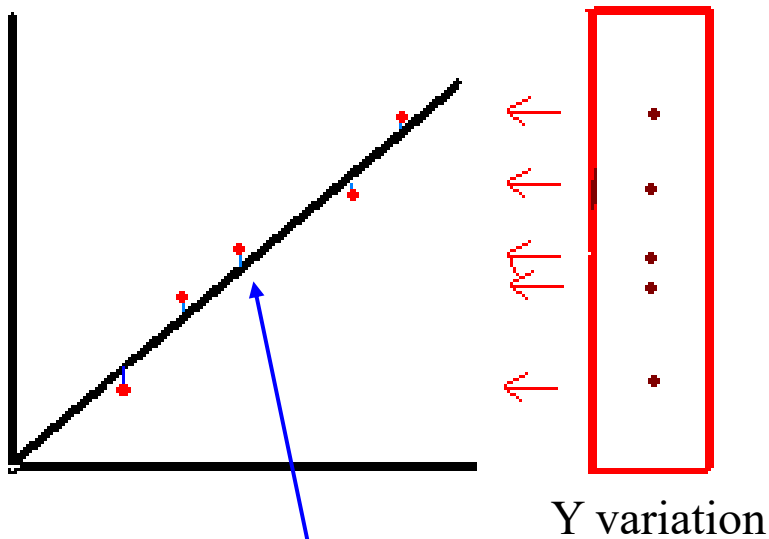
	write	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	5.492502	.8754227	6.27	0.000	3.765935	7.21907	
read	.1254123	.0649598	1.93	0.055	-.0027059	.2535304	
math	.2380748	.0671266	3.55	0.000	.1056832	.3704665	
science	.2419382	.0606997	3.99	0.000	.1222221	.3616542	
socst	.2292644	.0528361	4.34	0.000	.1250575	.3334713	
_cons	6.138759	2.808423	2.19	0.030	.599798	11.67772	

## Back to SBP and Age example

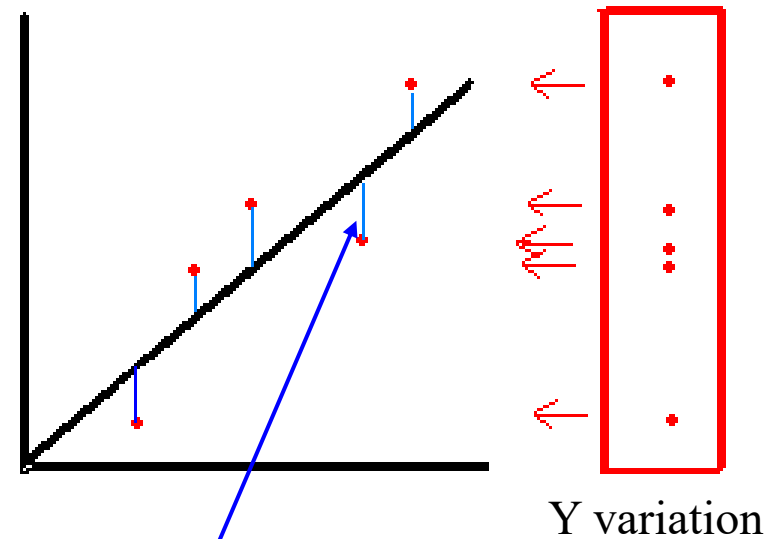
- $a=123$  and  $b=0.159$  approximately
- What does  $b$  mean?
- Is age predictive of BP? i.e. is there evidence that  $b \neq 0$ ?
- How good is the fit of the regression line?

# Regression $R^2$ Interpretation

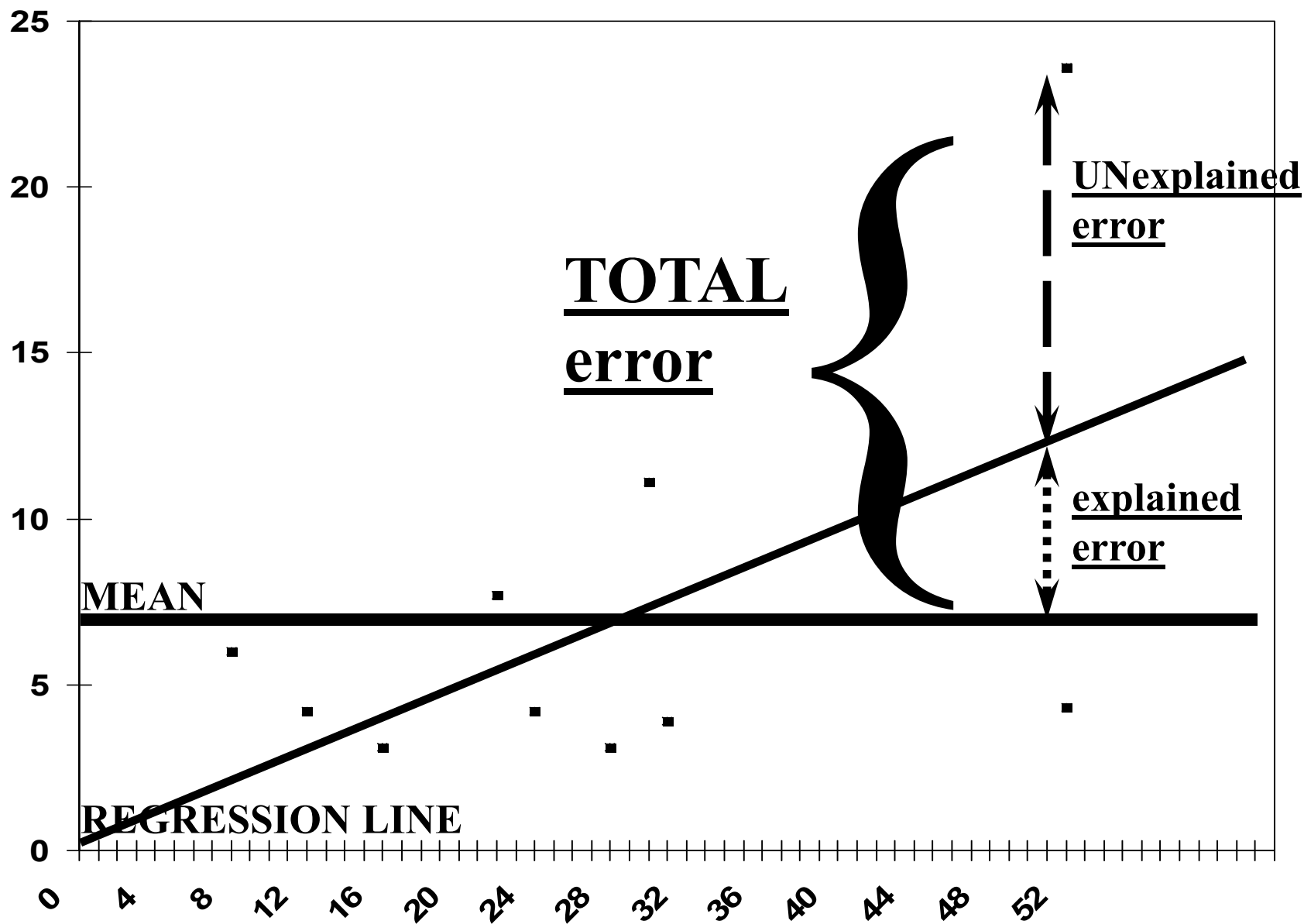
- $R^2$  = proportion of variation explained by (or predictive ability of) the regression



Variation in  $y$  is almost fully explained by  $x$ :  $R^2 \approx 1$



Still some variation in  $y$  left over (not explained by  $x$ ):  $R^2 < 1$

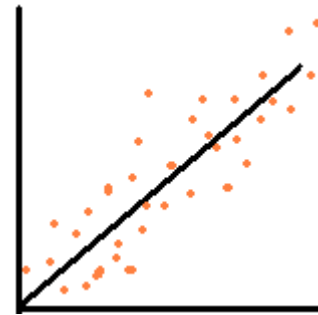


# Regression – four possibilities



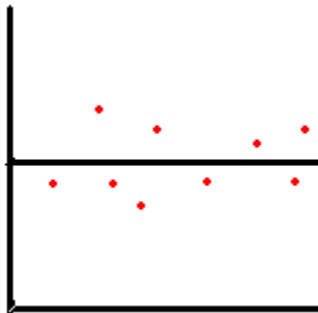
$b \neq 0$   
P-value non-significant

Relationship but not much evidence



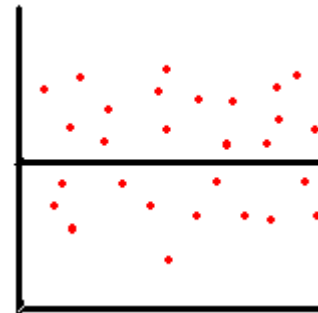
$b \neq 0$   
P-value significant

Plenty of evidence for a relationship



$b \approx 0$   
P-value non-significant

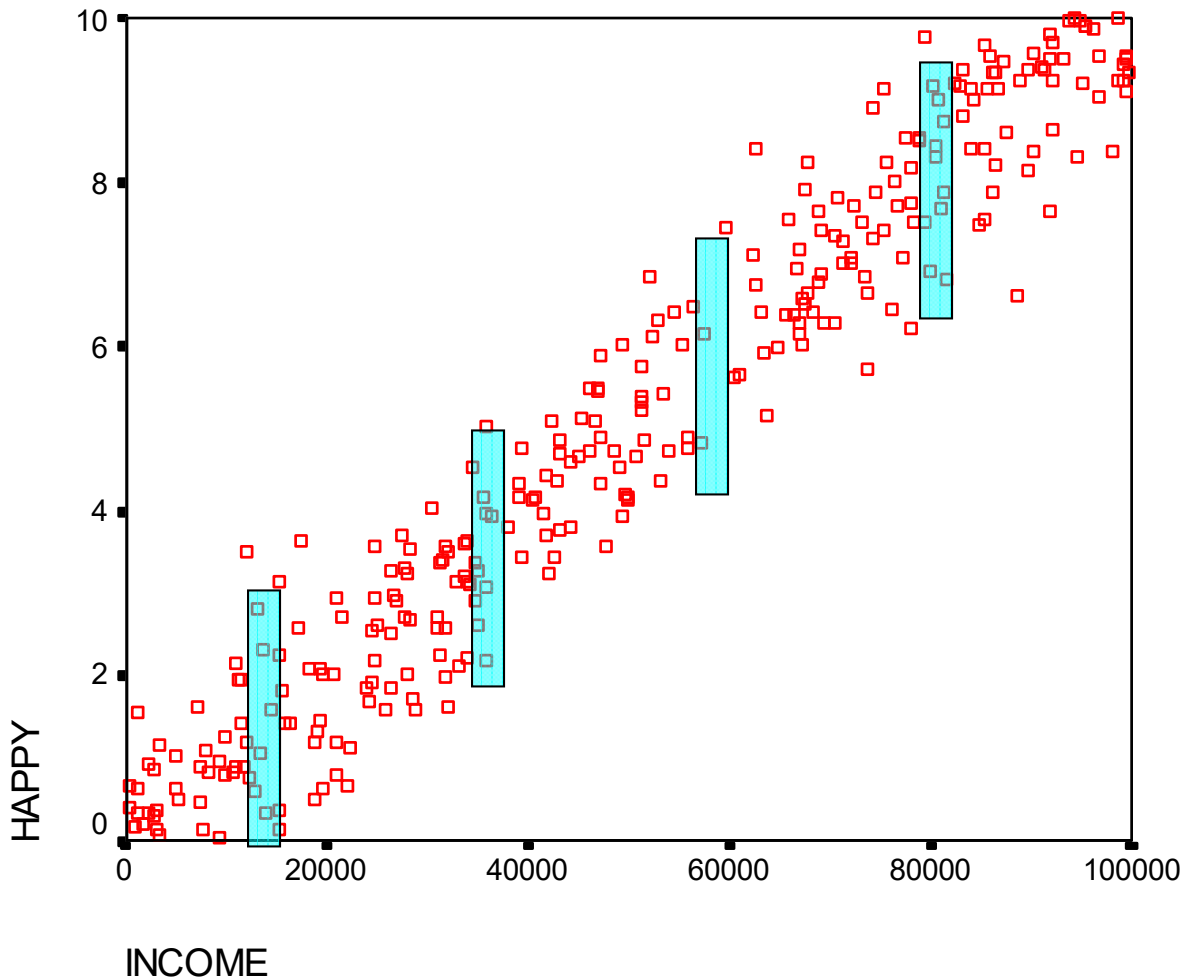
No relationship & not much evidence



$b \approx 0$   
P-value significant

Plenty of evidence for no relationship

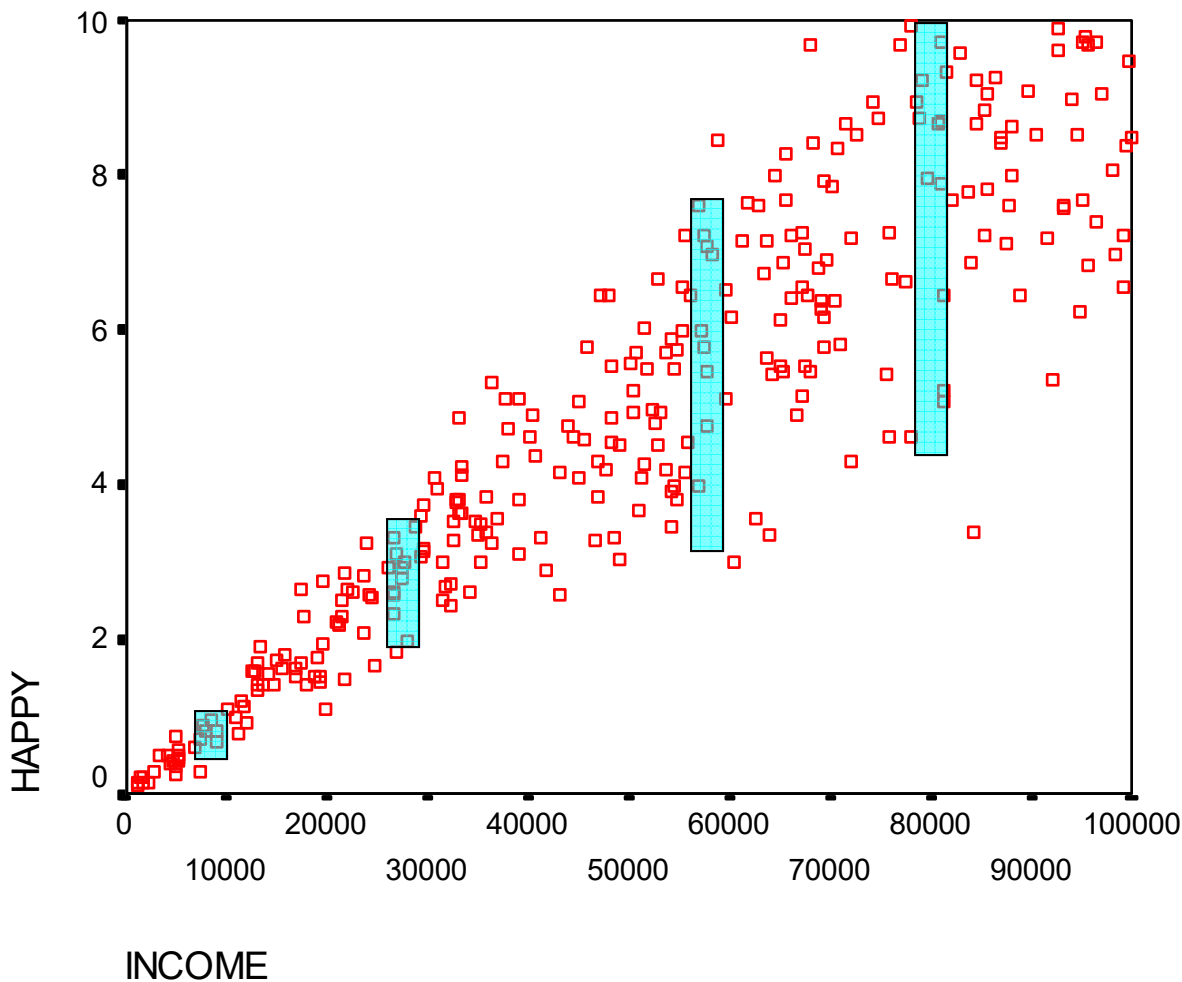
# Regression Assumption: Homoscedasticity (Equal Error Variance)



Examine error at  
different values of X.  
Is it roughly equal?

**Here, things look  
pretty good.**

# Heteroscedasticity: **Unequal Error Variance**



At higher values of X, error variance increases a lot.

**A transformation of data (e.g. log) can remove heteroskedasticity**

# Multiple Regression

- Extension of simple linear regression to more than one (continuous/ordinal) independent variables
- We use least squares in exactly the same way to obtain estimates of the regression coefficients
- e.g. with 2 independent variables x and z, we fit the regression

$$y = a + bx + cz \dots$$

where a, b and c are the regression coefficients. This represents a plane in 3d space

```
regress write female read math science socst
```

Source	SS	df	MS	Number of obs = 200	
Model	10756.9244	5	2151.38488	F( 5, 194)	= 58.60
Residual	7121.9506	194	36.7110855	Prob > F	= 0.0000
Total	17878.875	199	89.843593	R-squared	= 0.6017
				Adj R-squared	= 0.5914
				Root MSE	= 6.059

	write	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female		5.492502	.8754227	6.27	0.000	3.765935	7.21907
read		.1254123	.0649598	1.93	0.055	-.0027059	.2535304
math		.2380748	.0671266	3.55	0.000	.1056832	.3704665
science		.2419382	.0606997	3.99	0.000	.1222221	.3616542
socst		.2292644	.0528361	4.34	0.000	.1250575	.3334713
_cons		6.138759	2.808423	2.19	0.030	.599798	11.67772

Previous example



# Notes on multiple regression

- Make sure variables are normal. If not, transform them. If still not, can split into 2 groups (categories (0/1)) for e.g. high vs. low responders
- Can combine with “stepwise selection”: instead of using every variable and forcing them into a final model, can drop out variables automatically, e.g. petri dish temperature, that are not predictive

# Example

- Study to evaluate the effect of the duration of anesthesia and degree of trauma on percentage depression of lymphocyte transformation
- 35 patients
- Trauma factor classified as 0, 1, 3 and 4, depending upon severity

Duration	Trauma	Depression	Duration	Trauma	Depression
4	3	36.7	3	3	29.9
6	3	51.3	4	3	76.1
1.5	2	40.8	3	3	11.5
4	2	58.3	3	3	19.8
2.5	2	42.2	7	4	64.9
3	2	34.6	6	4	47.8
3	2	77.8	2	2	35
2.5	2	17.2	4	2	1.7
3	3	-38.4	2	2	51.5
3	3	1	1	1	20.2
2	3	53.7	1	1	-9.3
8	3	14.3	2	1	13.9
5	4	65	1	1	-19
2	2	5.6	3	1	-2.3
2.5	2	4.5	4	3	41.6
2	2	1.6	8	4	18.4
1.5	2	6.2	2	2	9.9
1	1	12.2			

## Example (con't)

- Fitted regression line is

$$y = -2.55 + 10.375x + 1.105z$$

or

$$\text{Depression} = -2.55 + 10.375 * \text{Trauma} + 1.105 * \text{Duration}$$

- Both slopes are non-significant (p-value=0.1739 for trauma, 0.7622 for duration)
- $R^2 = 16.6\%$  of the variation in lymphocyte depression is explained by the regression
- Conclusion: Trauma score and duration of anesthesia are poor explanations for lymphocyte depression

# Collinearity

- If two (independent) variables are closely related its difficult to estimate their regression coefficients because they tend to get confused
- This difficulty is called *collinearity*
- Solution is to exclude one of the highly correlated variables

# Example

- Correlation between trauma and duration= 0.762 (quite strong)
- Drop trauma from regression analysis

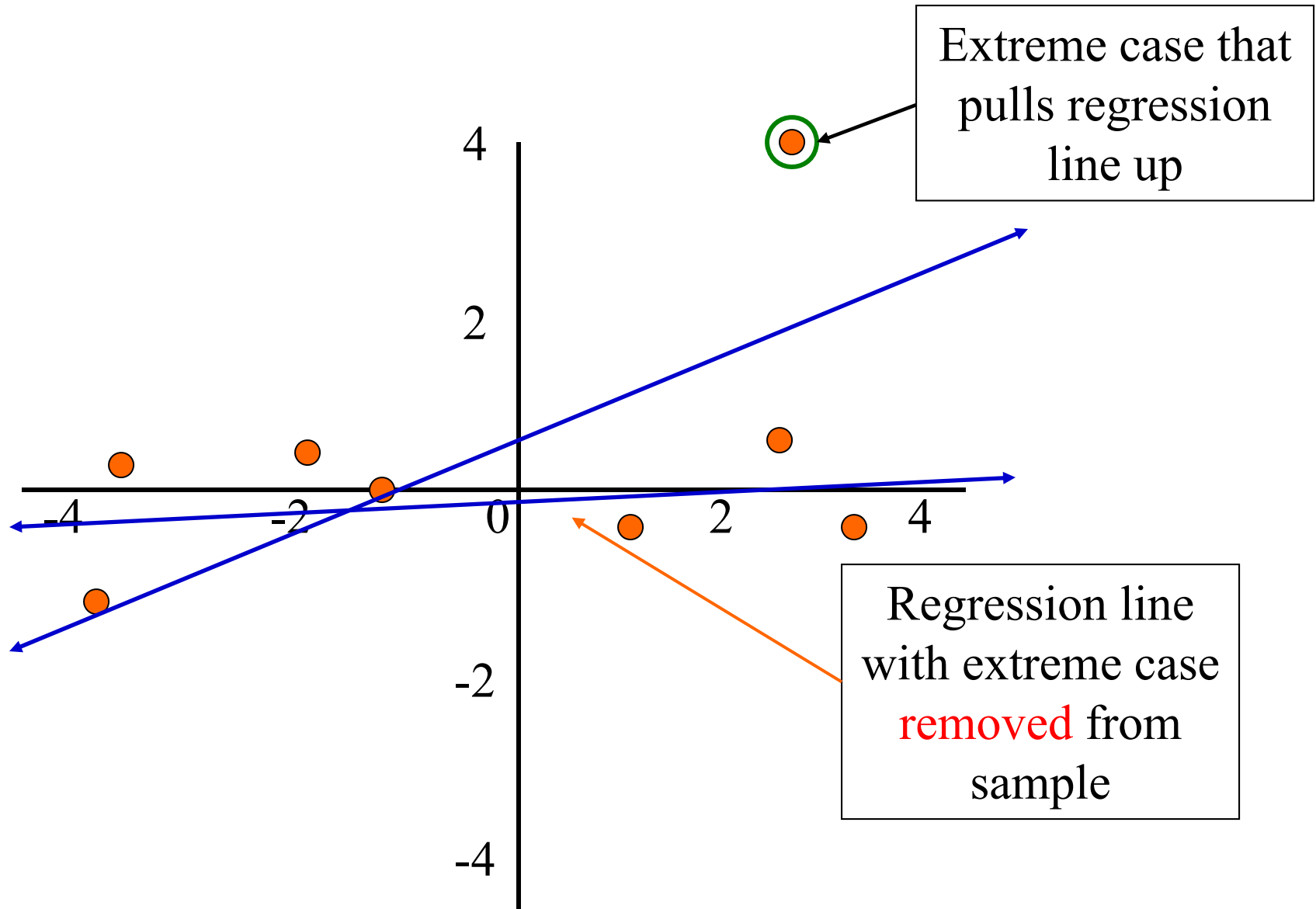
$$\text{Depression} = 9.73 + 4.94 * \text{Duration}$$

- P-value for duration is 0.0457, statistically significant!
- However, the  $R^2$  is still small (11.6%)
- Conclusion: Although there is evidence for a non-zero slope or linear relationship with duration, there is still considerable variation not explained by the regression.

# Outliers in Regression

- Outliers: cases with extreme values that differ greatly from the rest of your sample
- Even a few outliers can dramatically change estimates of the slope ( $b$ )
- Outliers can result from:
  - Errors in coding or data entry (→rectify)
  - Highly unusual cases (→exclude?)
  - Or, sometimes they reflect important “real” variation (→include?)

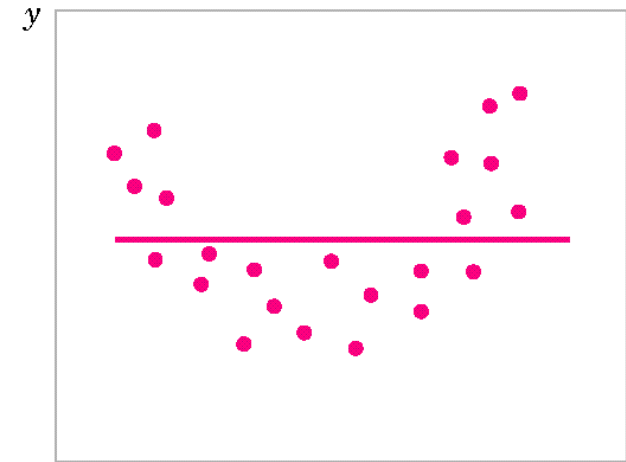
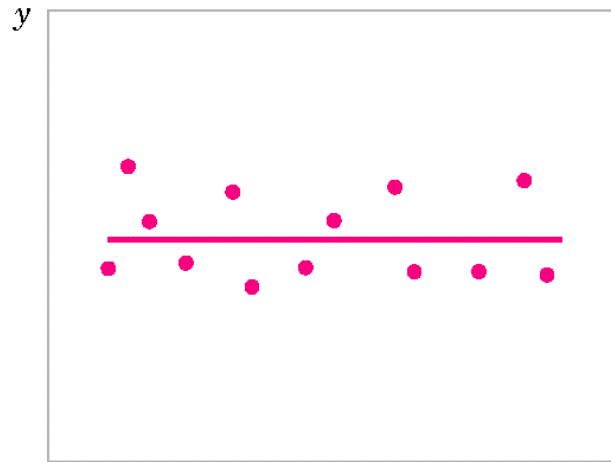
# Outliers: Example



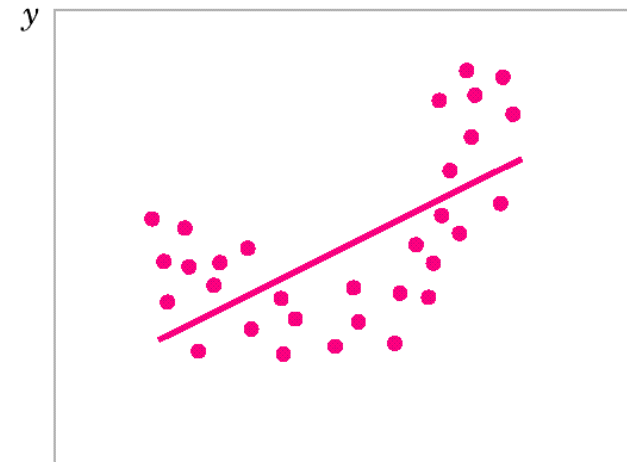


# What about non-linear relationships?

Fail to  
reject  $\beta=0$



Reject  $\beta=0$

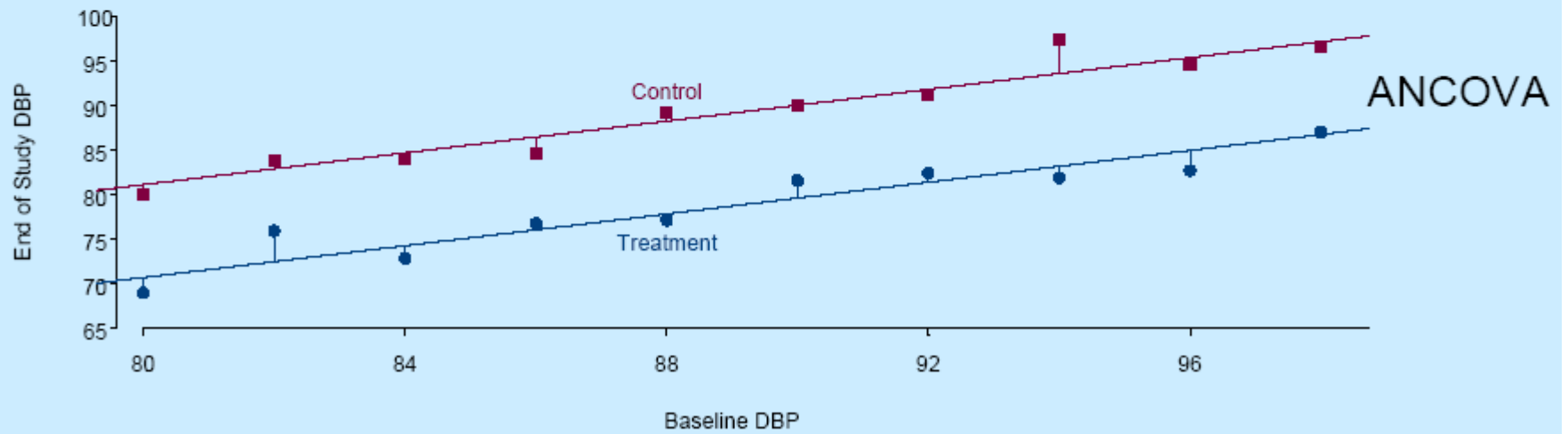
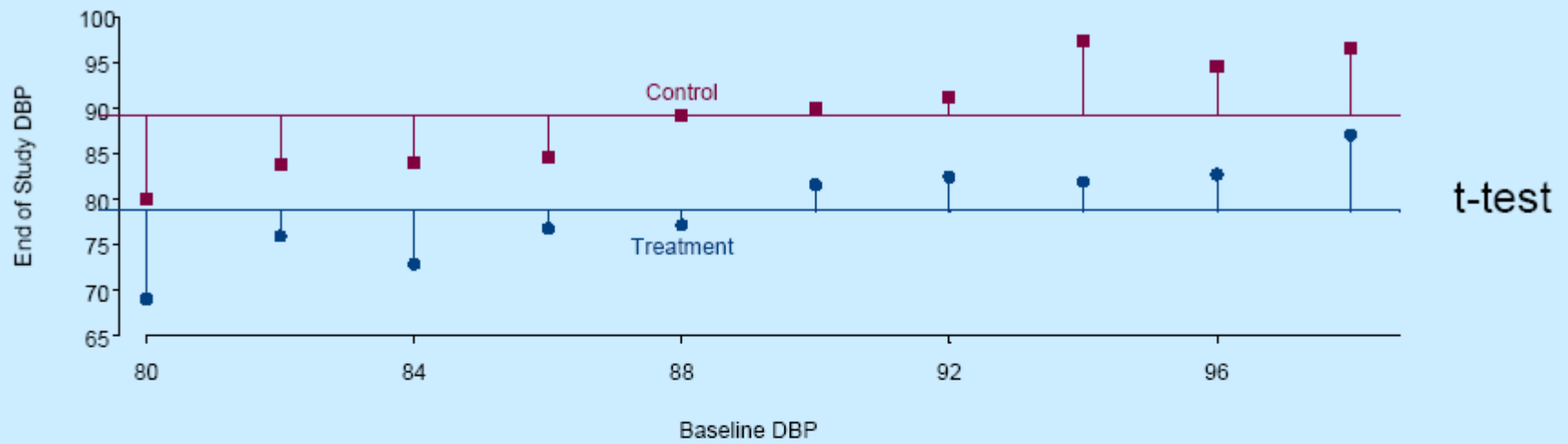


# Non-linear models

- Linear Regression fits a straight line to your data
- Non-linear models may, in some cases, be more appropriate
- Non-linear models are usually used in situations where non-linear relationships have some *biological explanation*
- e.g. non-linear models are commonly used in pharmacokinetics studies and compartmental analysis
- Computationally intensive - estimates may not “converge” and results may be difficult to interpret

# Analysis of Covariance-ANCOVA

- Modelling both categorical and continuous independent variables (*covariates*)
- Justification: Consideration of a covariate may improve the precision (reduce variability) of the comparisons of categories made in ANOVA
- Often used as a correction for baseline imbalances
- ANOVA + regression combined



Within-arm variability smaller with ANCOVA than with t-test

# Example

- Antihypertensive Drug Clinical Trial: 84 patients randomly allocated to either Aliskiren or Losartan. Blood pressure measured at baseline and after several weeks treatment. Age and gender was also recorded for each patient. Is there a significant difference in the two treatments for BP reduction?

Age	Treatment	Ba_daySBP	Red_daySBP	gender	Age	Treatment	Ba_daySBP	Red_daySBP	gender
66	SPP75	131.1	0.6	Female	45	Los100	156	17	Male
62	SPP75	160.4	-0.7	Male	68	Los100	149.9	6.8	Female
48	Los100	147.3	17.9	Male	48	Los100	147	16	Female
32	Los100	144.8	-2.4	Male	69	SPP75	145.3	-1.2	Male
61	Los100	150.7	21.6	Female	64	Los100	142.5	20.6	Female
68	SPP75	152.4	6	Male	40	SPP75	168.6	2.6	Male
60	Los100	143.6	15.3	Male	61	SPP75	165.6	8.1	Male
33	SPP75	143.2	5.1	Male	47	Los100	156.3	2.3	Female
69	Los100	166.6	24.6	Male	60	Los100	147.7	11.5	Female
53	SPP75	147.6		Female	35	SPP75	157.4	12.5	Male
63	SPP75	163.7	-0.2	Male	61	SPP75	143.7	-1.8	Male
64	Los100	145.7	15.5	Male	62	Los100	148.6	6.6	Male
58	SPP75	168.3	0.5	Male	54	Los100	164.6	39.6	Female
52	SPP75	156.8	0.6	Male	45	Los100	145.3	-6.1	Female
54	SPP75	154.9	7.3	Male	57	SPP75	143.9	7	Female
43	SPP75	170.5	-2.7	Female	48	SPP75	144.3	2.4	Male
46	SPP75	155.5	18.3	Female	59	Los100	147.8	1.1	Female
46	Los100	173.1	26.7	Male	47	SPP75	150.4	-2.3	Male
66	Los100	151.2	-1.7	Male	54	Los100	143.9	-0.2	Male
29	SPP75	139.8	-1.5	Male	45	SPP75	145.1	6.6	Male
50	SPP75	162.6	13	Male	61	SPP75	158	3	Male
49	SPP75	178.8	17.2	Female	69	Los100	154.8		Male
40	SPP75	146.8	0	Male	21	SPP75	142.1		Male
52	Los100	157.1	-0.2	Female	69	SPP75	171.3	-2	Male
68	Los100	152	8.9	Male	66	Los100	140.3	2.8	Male
35	SPP75	145.4	2.8	Male	42	SPP75	146	5.7	Male
49	Los100	153.7	14.9	Male	47	SPP75	159.3	14.6	Female
47	SPP75	139.2	10	Male	60	SPP75	157.8	6.8	Male

Age	Treatment	Ba_daySBP	Red_daySBP	gender	Age	Treatment	Ba_daySBP	Red_daySBP	gender
45	SPP75	162.9	10.7	Male	58	Los100	153	-10.3	Male
62	SPP75	173.4	55.9	Female	41	SPP75	149.9	4.3	Male
57	SPP75	141.1	0.2	Female	42	Los100	165.8	35.5	Male
54	Los100	147.6	11.1	Male	57	SPP75	154.8	16.1	Female
54	Los100	140.5	-16.5	Male	56	Los100	146.6	18.1	Female
63	Los100	156	19.3	Female					
35	SPP75	150.9	26.8	Male					
52	SPP75	143.8	-9.7	Female					
66	Los100	150.9	0.6	Male					
55	SPP75	155.8	-3	Female					
61	Los100	162.1	21	Male					
35	Los100	149.1	21.3	Male					
52	Los100	177	-2.7	Male					
60	SPP75	157.8	6.8	Male					
37	SPP75	143.4	-3.9	Male					
54	Los100	163.9	26.6	Male					
62	Los100	147.4		Female					
55	Los100	141.9	8.7	Female					
57	Los100	163.1	3.6	Male					
40	SPP75	153.3	0.9	Male					
56	SPP75	165.9	11.9	Male					
53	Los100	144.1	-1.9	Female					
52	Los100	144.1	18.6	Male					
46	SPP75	147.7	-5.6	Male					

## Analysis without covariates

- Since treatment has only 2 levels (Losartan & Aliskiren), the ANOVA is equivalent to the two-sample t-test
- Treatment difference (Losartan-Aliskiren)=5.06 with P-value=0.0554
- Borderline non-significant at the 5% level of significance



# ANCOVA analysis

- We have as factors (categorical independent variables)
  - Treatment
  - Gender
- As covariates (continuous independent variables)
  - Age
  - Baseline BP

## Results: ANOVA table

Source	df	SS	MS	F	P-value
Baseline BP	1	1572.90	1572.90	14.36	0.0005
Treatment	1	651.73	651.73	5.58	0.0208
Age	1	71.57	71.57	0.61	0.4363
Gender	1	279.02	279.02	2.39	0.1264

# Classification of statistical methods based on distributional assumptions

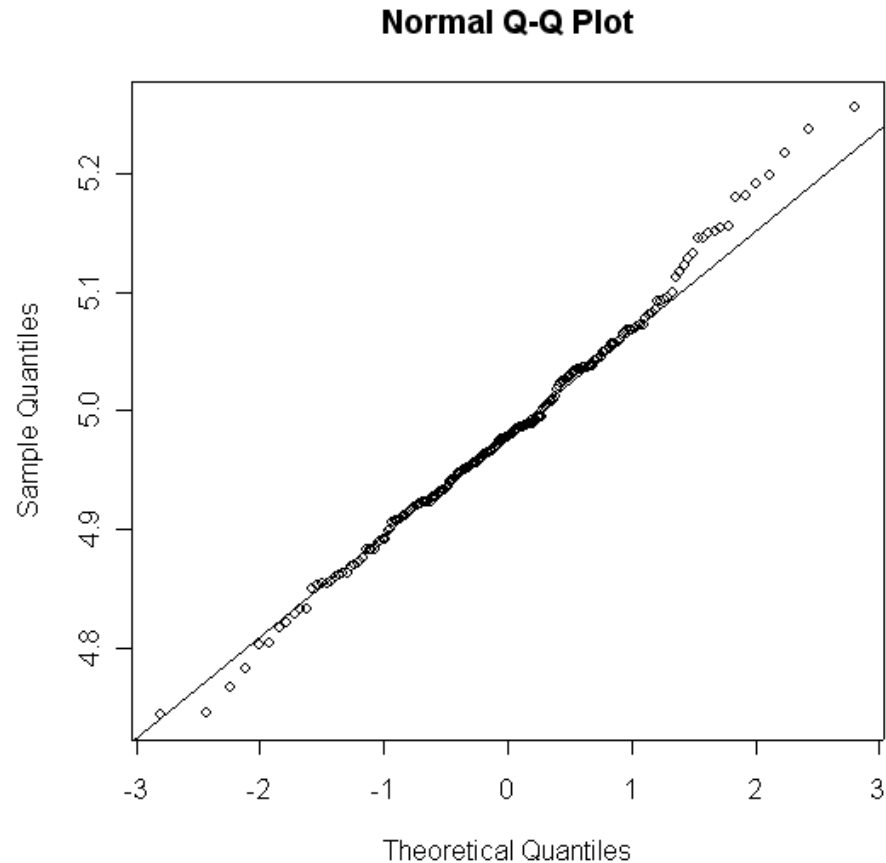
**Distributional  
assumptions**



- Likelihood Based Methods
  - Assume a distribution of data explicitly at the outset
  - Can fit very complex models e.g. model correlation structures with the data, allow for unequal variances, etc.
- T-tests, ANOVA, regression, etc. rely on normality of data or that the data be of a sufficiently large size (Central Limit Theorem)
- Non-parametric methods
  - Relaxes assumptions about the shape of the distribution
  - Methods are based on ranking of the data points

# So how normal is your data?

- Difficult to see visually if a histogram looks normal
- Use normal probability (quantile-quantile) plots
- Points must lie along a line
- Also useful for detecting outliers



# Significant Deviations from Normality: Alternatives

- Transform your data to make it more “normal” and use standard parametric tests
  - e.g. log transform (eliminates skew)
  - Difficulty - may wish to reverse-transform results back (e.g. exponentiate parameter estimates)
- Use non-parametric methods
  - Make less assumptions on distribution shape
  - These may have less power than parametric methods

## Non-parametric Equivalents

Parametric	Non-parametric
Paired T-test	Wilcoxon Signed Rank Test
Two-sample T-test	Wilcoxon Rank-Sum
ANOVA	Kruskal Wallis Analysis
Pearson's Correlation	Spearman's Rank Correlation
ANCOVA	ANCOVA on ranked data

# Other Multivariate Methods

- Multivariate Analysis of Variance (MANOVA)
  - ANOVA with more than one dependent variable or response
  - Also MANCOVA
  - Low power is a problem
  - Not so widely used
- The objectives behind multivariate analyses can be quite different (to those presented), namely
  - Discriminant Analysis
  - Classification
  - Clustering
  - Pattern Recognition (principal components analysis)

# Multivariate analysis: mineral water

Pattern

Pattern recogni

U.L.S.S. n° 18 A.N.P.A.V.  
Sezione Chimico  
Ambientale - PADOVA

**Analisi Chimica e Chimico - Fisica**

Temperatura dell'acqua al prelievo	16,7° C
pH	7,68
Conducibilità a 20° C	400 µS/cm
Residuo fisso a 180° C	250 mg/l

**Gas disciolti in un litro d'acqua al prelievo**

Anidride carbonica libera	mg 9,6
Ossigeno	mg 7,1

**Sostanze disciolte in un litro d'acqua espresse in ioni e mg**

Sodio	6,8
Potassio	1,1
Magnesio	30
Calcio	46
Idrocarbonico	293
Cloridrico	2,8
Nitrico	6,8
Solfonico	4,9
Silice (come SiO <sub>2</sub> )	17
Fluoridrico	<0,1

**MENO DELLO 0,0007% DI SODIO**  
**MICROBIOLOGICAMENTE PURA**

Padova, 13/10/2000. Autorizzazioni: D.R. Veneto n° 558 del 09/12/1998; Ministero della Sanità n° 3207-126 del 25/11/1999 e n°3399-126 del 27/07/2001.

**SAN BENEDETTO**  
*Acqua Minerale Naturale*  
OLIGOMINERALE  
*Leggermente Frizzante*  
**MENO DELLO 0,0007% DI SODIO**



# Correlation coefficients in data

- The square root of coefficient R is C.
- Example: we measure 5 variables of a population of peaches :
  - Total acidity, anthocyanin, brix, carotene e chlorophyll
  - We want to know the correlation between them
- The correlation can be shown as a matrix 5\*5

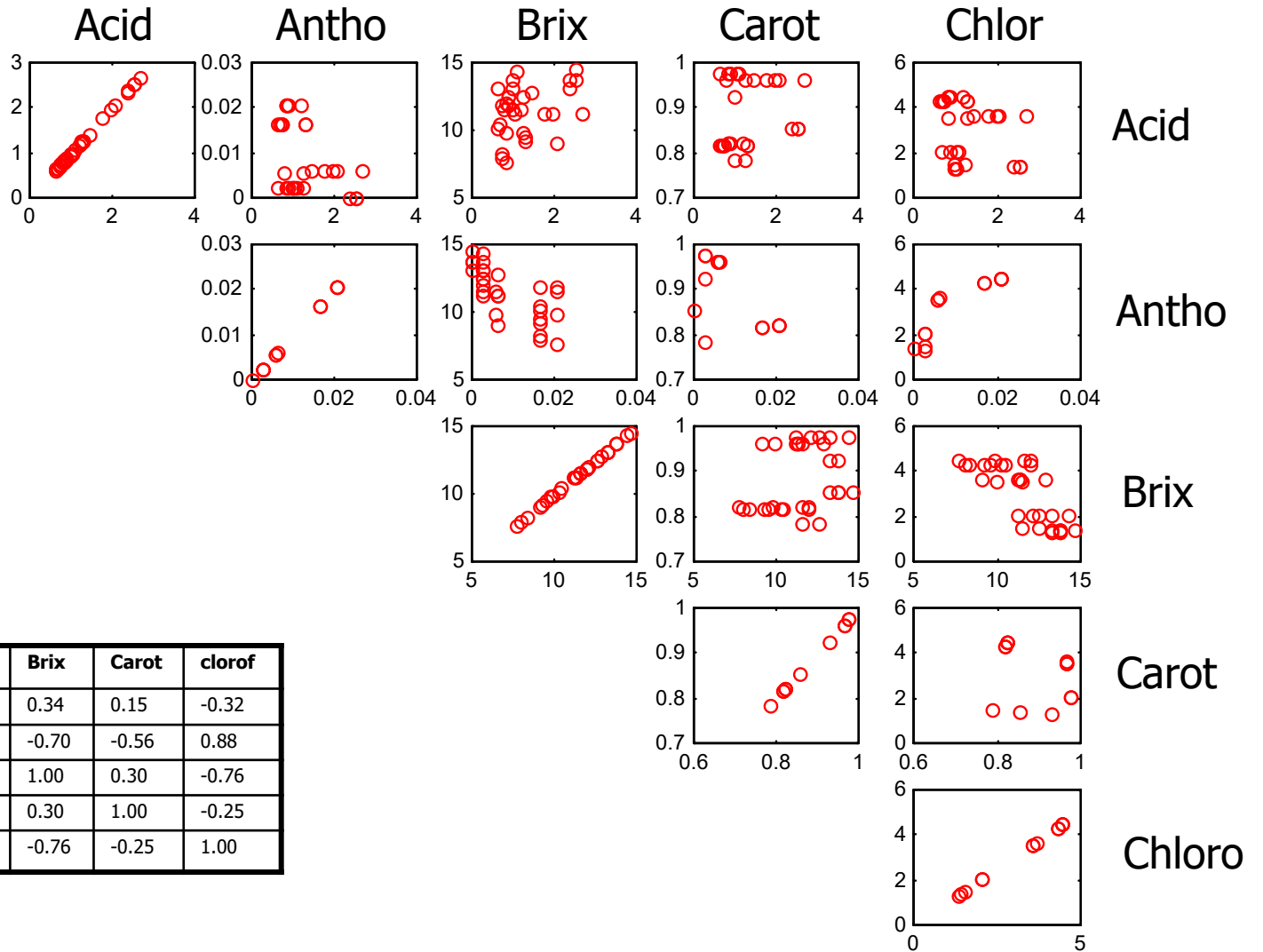
	<b>Acid</b>	<b>Anth</b>	<b>Brix</b>	<b>Carot</b>	<b>chlo</b>
<b>acid</b>	1.00	-0.48	0.34	0.15	-0.32
<b>anth</b>	-0.48	1.00	-0.70	-0.56	0.88
<b>brix</b>	0.34	-0.70	1.00	0.30	-0.76
<b>carot</b>	0.15	-0.56	0.30	1.00	-0.25
<b>chlor</b>	-0.32	0.88	-0.76	-0.25	1.00

# correlation matrix

	<b>Acid</b>	<b>Antho</b>	<b>Brix</b>	<b>Carot</b>	<b>chlo</b>
<b>acid</b>	1.00	-0.48	0.34	0.15	-0.32
<b>antoc</b>	-0.48	1.00	-0.70	-0.56	0.88
<b>brix</b>	0.34	-0.70	1.00	0.30	-0.76
<b>carot</b>	0.15	-0.56	0.30	1.00	-0.25
<b>clorof</b>	-0.32	0.88	-0.76	-0.25	1.00

- The matrix is symmetric
  - Y to X has the same correlation of X to Y
- The values are between -1 (anticorrelation) and 1 (correlation)
- Usually two variables are not 100% correlated ( $c=1$ ) or not at all correlated ( $c=0$ ) there are always a partial correlation between variables

# Correlation graph



# Multivariate probability

- What is the probability that a peach has at the same time a concentration of carotene of  $0.40 \pm 0.02$  and of chlorophyll of  $4.31 \pm 0.23$ ?
- To answer this question we have to know the joint probability.
  - There are two possibilities :
    - Y and X are independent  $\Rightarrow P(X,Y) = P(X) \cdot P(Y)$
    - Y and X are dependent on each other  $\Rightarrow P(X,Y)$
  - In the first case we have the product of the PDF univariate functions (PDF)
  - In the second case we have to introduce the bivariate PDF

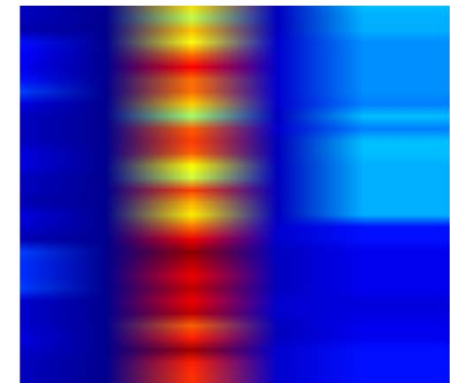
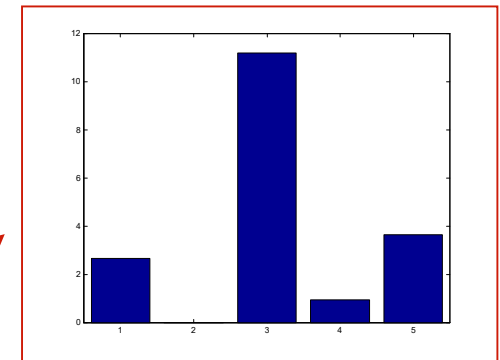
# Multivariate data

- The multivariate value is a vector
- The multivariate data are shown as a matrix
- Example :
  - A population of 20 peaches having measured 5 variables:
  - Total acidity, anthocyanin, brix, carotene e chlorophyll

**variabili** →

↓ **campioni**

Acidity	anthocyanin	Brix	Carotene	chlorophyll
0.8200	0.0206	9.8000	0.8231	4.4600
0.7300	0.0165	8.0000	0.8179	4.2900
0.6000	0.0165	10.2000	0.8179	4.2900
2.0400	0.0060	9.1000	0.9642	3.6600
1.7600	0.0060	11.3000	0.9642	3.6600
1.4200	0.0060	12.8000	0.9642	3.6600
1.9700	0.0060	11.3000	0.9642	3.6600
2.6800	0.0060	11.2000	0.9642	3.6600
1.2440	0.0057	9.9000	0.9634	3.5700
0.8300	0.0206	7.7000	0.8231	4.4600
0.7880	0.0057	11.5000	0.9634	3.5700
0.8600	0.0206	11.9000	0.8231	4.4600
1.1800	0.0206	11.6000	0.8231	4.4600
1.2700	0.0165	9.2000	0.8179	4.2900
0.7300	0.0165	8.3000	0.8179	4.2900
0.7200	0.0165	11.9000	0.8179	4.2900
0.6600	0.0165	10.4000	0.8179	4.2900
1.2600	0.0165	9.5000	0.8179	4.2900
1.0000	0.0025	11.2000	0.9756	2.0300
0.6400	0.0025	13.2000	0.9756	2.0300



# Multivariate average

- The average multivariate data is a vector made by the average of the single variable (the column).

Acidity	anthocyanin	Brix	Carotene	chlorophyll
0.8200	0.0206	9.8000	0.8231	4.4600
0.7300	0.0165	8.0000	0.8179	4.2900
0.6000	0.0165	10.2000	0.8179	4.2900
2.0400	0.0060	9.1000	0.9642	3.6600
1.7600	0.0060	11.3000	0.9642	3.6600
1.4200	0.0060	12.8000	0.9642	3.6600
1.9700	0.0060	11.3000	0.9642	3.6600
2.6800	0.0060	11.2000	0.9642	3.6600
1.2440	0.0057	9.9000	0.9634	3.5700
0.8300	0.0206	7.7000	0.8231	4.4600
0.7880	0.0057	11.5000	0.9634	3.5700
0.8600	0.0206	11.9000	0.8231	4.4600
1.1800	0.0206	11.6000	0.8231	4.4600
1.2700	0.0165	9.2000	0.8179	4.2900
0.7300	0.0165	8.3000	0.8179	4.2900
0.7200	0.0165	11.9000	0.8179	4.2900
0.6600	0.0165	10.4000	0.8179	4.2900
1.2600	0.0165	9.5000	0.8179	4.2900
1.0000	0.0025	11.2000	0.9756	2.0300
0.6400	0.0025	13.2000	0.9756	2.0300

↓ ↓ ↓ ↓ ↓

<b>1.2874</b>	<b>0.0084</b>	<b>11.4516</b>	<b>0.8867</b>	<b>3.0586</b>
---------------	---------------	----------------	---------------	---------------

# variance in Multivariate data

- In multivariate data the variance is a matrix called covariance matrix
- The covariance matrix is linked to correlation and correlation matrix
- The covariance matrix is defined as:

$$\text{cov}(X) = \Sigma = E\left[(x - m)^T \cdot (x - m)\right]$$

- The covariance matrix is symmetric and quadratic. The dimensions are equal to the variables measured
- each element on the principal diagonal of the covariance matrix is just the variance of each of the elements in the vector
- The other elements of the covariance matrix are proportional to the correlation coefficients ( $\rho$ )

$$\Sigma_{ii} = \sigma_i^2 \quad ; \quad \Sigma_{ik} = \rho_{ik} \sigma_i \sigma_k$$

<b>0.4167</b>	-0.0023	0.4175	0.0072	-0.2635
-0.0023	<b>0.0001</b>	-0.0099	-0.0003	0.0084
0.4175	-0.0099	<b>3.5179</b>	0.0409	-1.8080
0.0072	-0.0003	0.0409	<b>0.0053</b>	-0.0236
-0.2635	0.0084	-1.8080	-0.0236	<b>1.5868</b>

# Covariance and Correlation

- The covariance matrix can be written as:

$$\Sigma = \Gamma \cdot R \cdot \Gamma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix} \cdot \begin{bmatrix} 1 & \rho_{21} & \dots & \rho_{n1} \\ \rho_{12} & 1 & & \\ \dots & & \dots & \\ \rho_{1n} & & & 1 \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix}$$

- Where R is the correlation matrix



# Correlation and independence

- Two variables are independent if :

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

- This condition is also called linear independence

- Two variables are independent if:

$$P[X \cdot Y] = P[X] \cdot P[Y]$$

# PDF multivariate 1

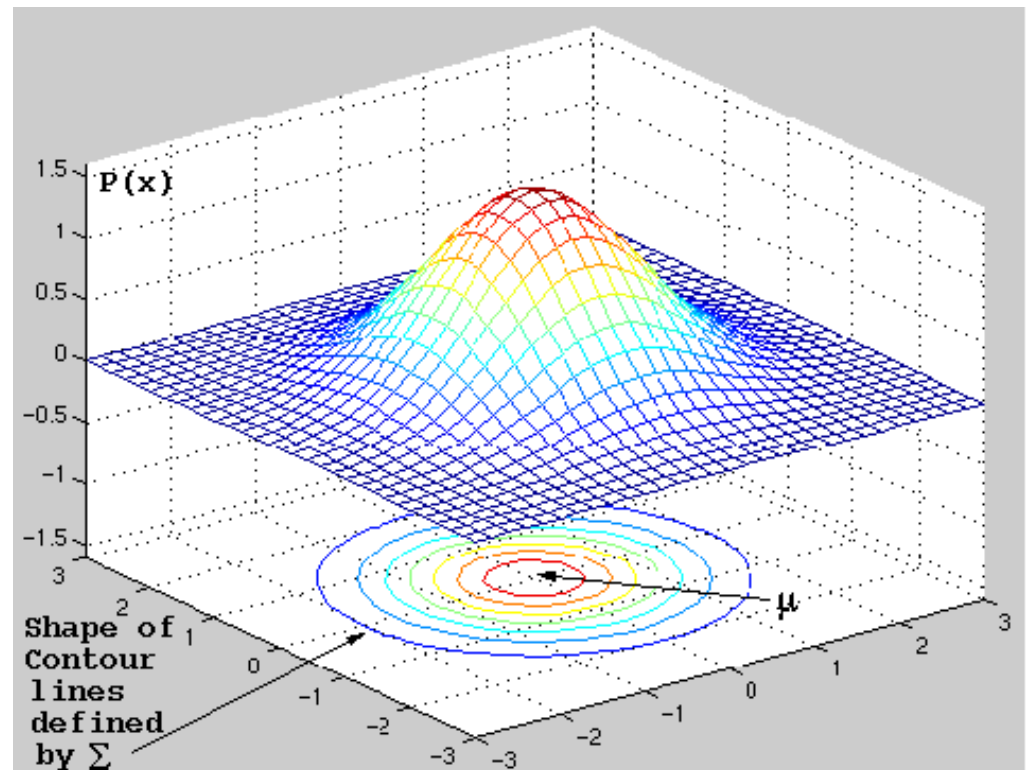
$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

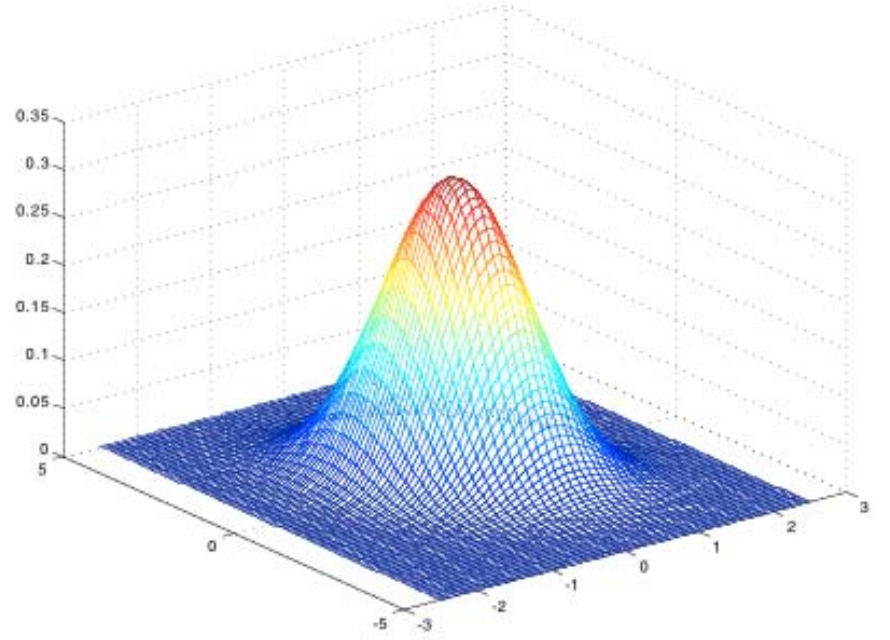
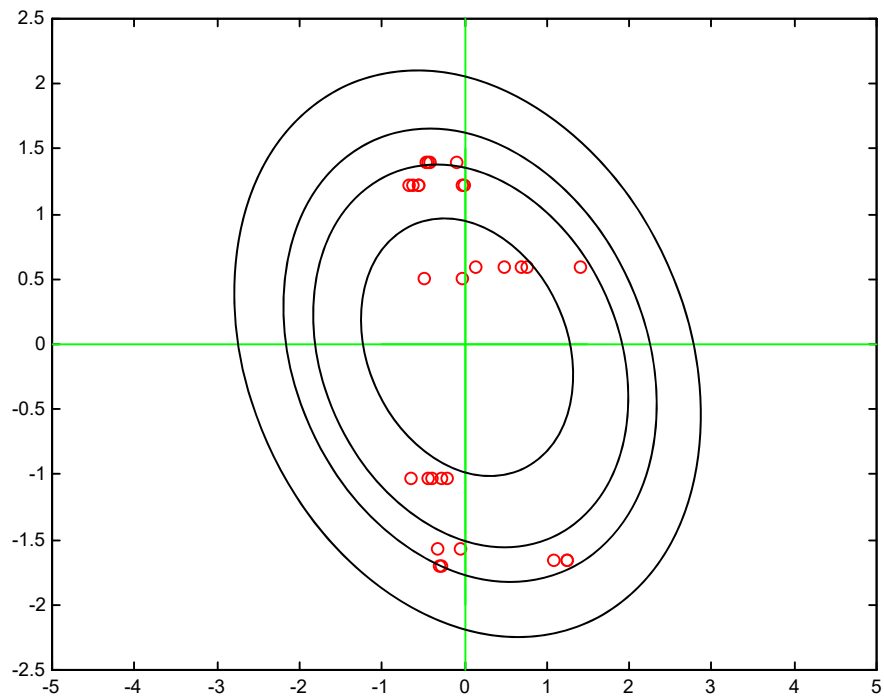
$$f_{\vec{x}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

where n is the dimensionality of the space under consideration.

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi} \sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \vec{\mu})^T \Sigma^{-1}(\mathbf{x} - \vec{\mu})\right]$$

The probability points are quadratic forms. **Ellipse** for the PDF bivariate.





# Instruments review

Instrument Converts information stored in the physical or chemical characteristics of the analyte into useful information

Require a source of energy to stimulate measurable response from analyte

## Data domains

- Methods of encoding information electrically

- Nonelectrical domains

- Electrical domains

  - Analog, Time, Digital

## Detector

Device that indicates a change in one variable in its environment (eg., pressure, temp, particles)

Can be mechanical, electrical, or chemical

## Sensor

Analytical device capable of monitoring specific chemical species continuously and reversibly

## Transducer

Devices that convert information in nonelectrical domains to electrical domains and the converse

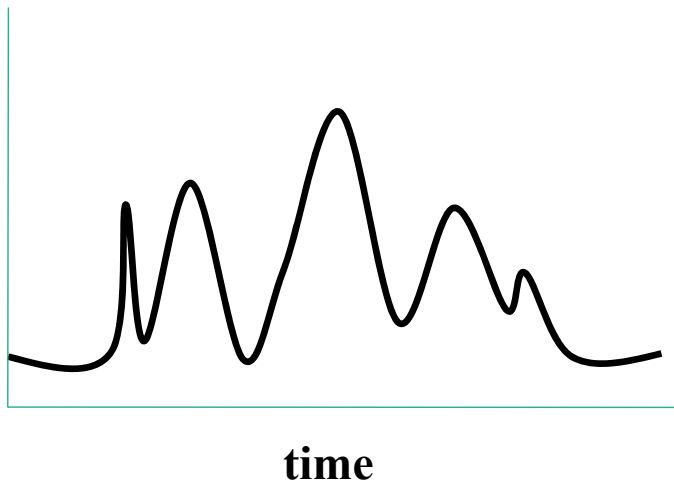
# Method Validation

- Specificity
- Linearity
- Accuracy
- Precision
- Range
- Limits of Detection and Quantitation

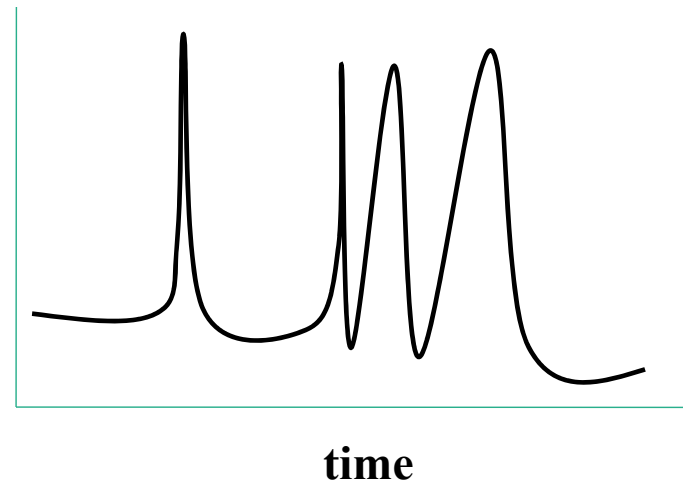


# Method Validation - Specificity

- How well an analytical method distinguishes the analyte from everything else in the sample.
- Baseline separation



**vs.**



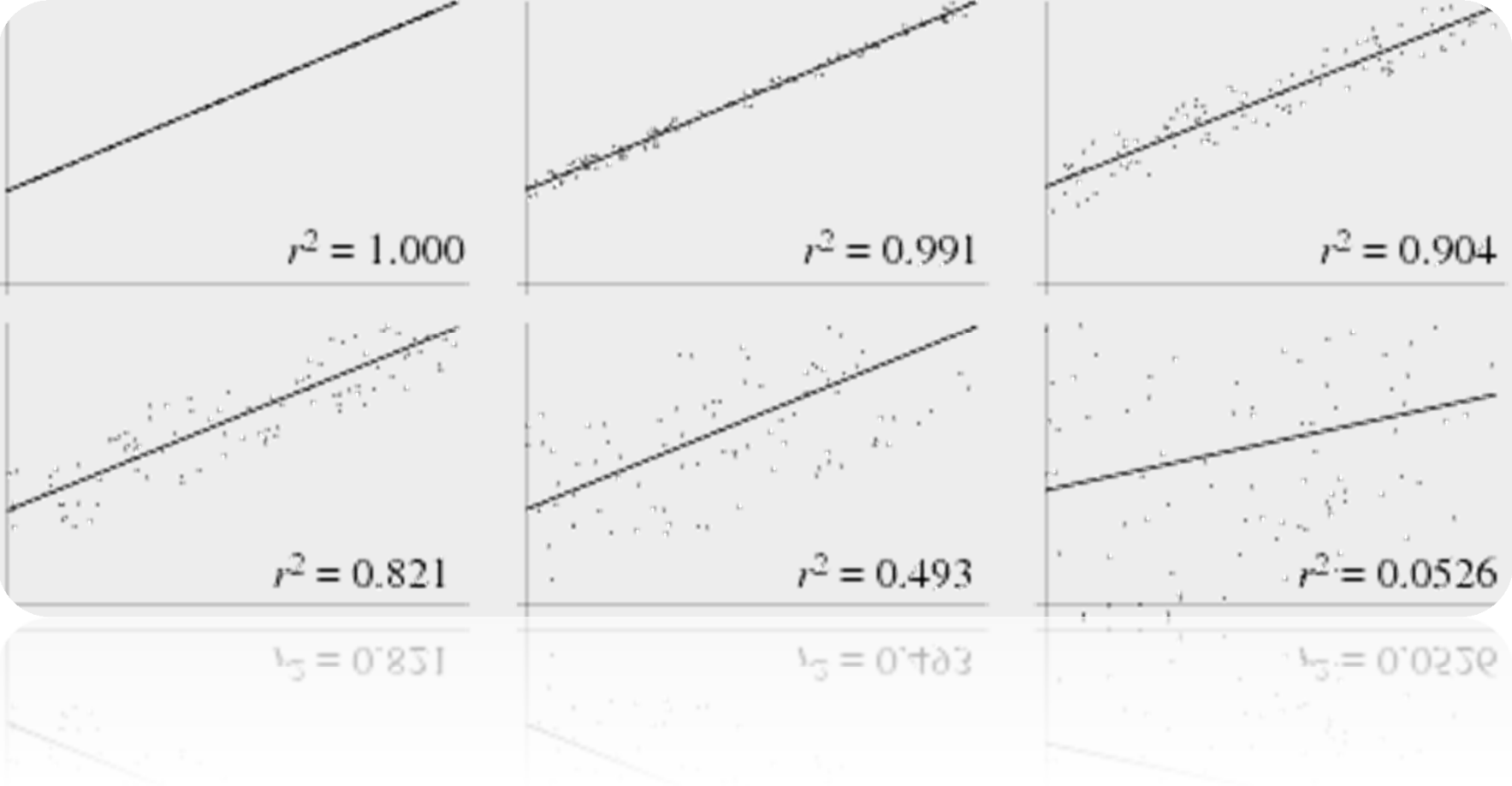
# Method Validation- Linearity

- How well a calibration curve follows a straight line.
- $R^2$  (Square of the correlation coefficient)

$$R^2 = \frac{[\sum(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$$



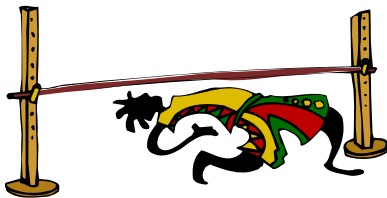
# Method Validation- Linearity



# Method Validation- LOD and LOQ

## Sensitivity

- Limit of detection (LOD) – “the lowest content that can be measured with reasonable statistical certainty.”
- Limit of quantitative measurement (LOQ) – “the lowest concentration of an analyte that can be determined with acceptable precision (repeatability) and accuracy under the stated conditions of the test.”

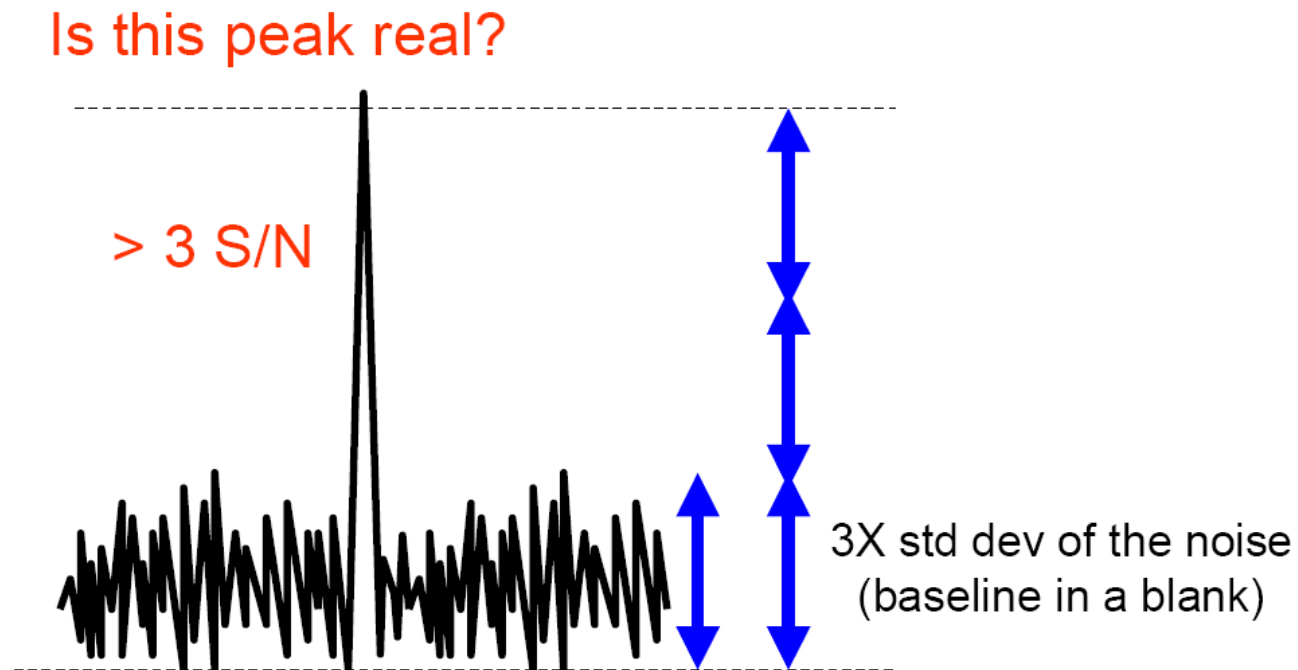


• How low can you go?



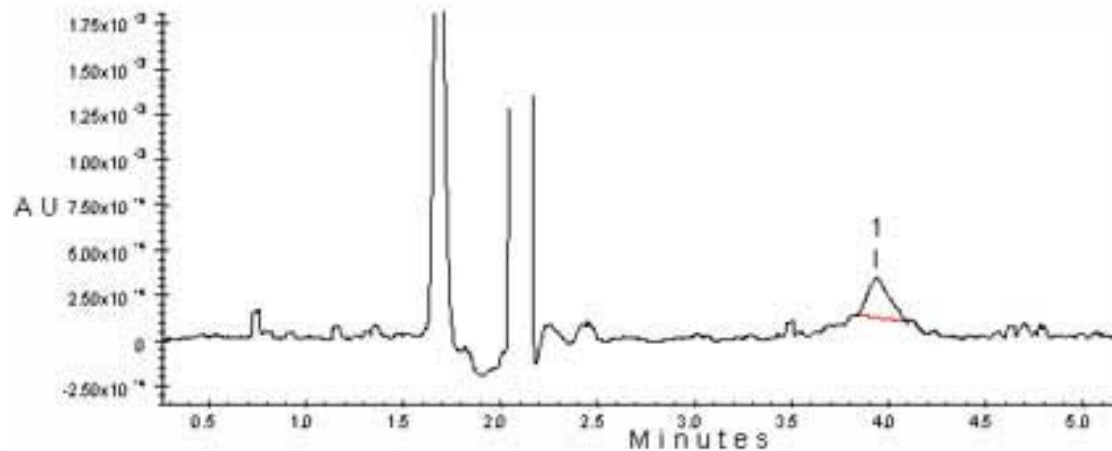
# Limit of Detection (LOD)

- Typically 3 times the signal-to-noise (based on standard deviation of the noise)

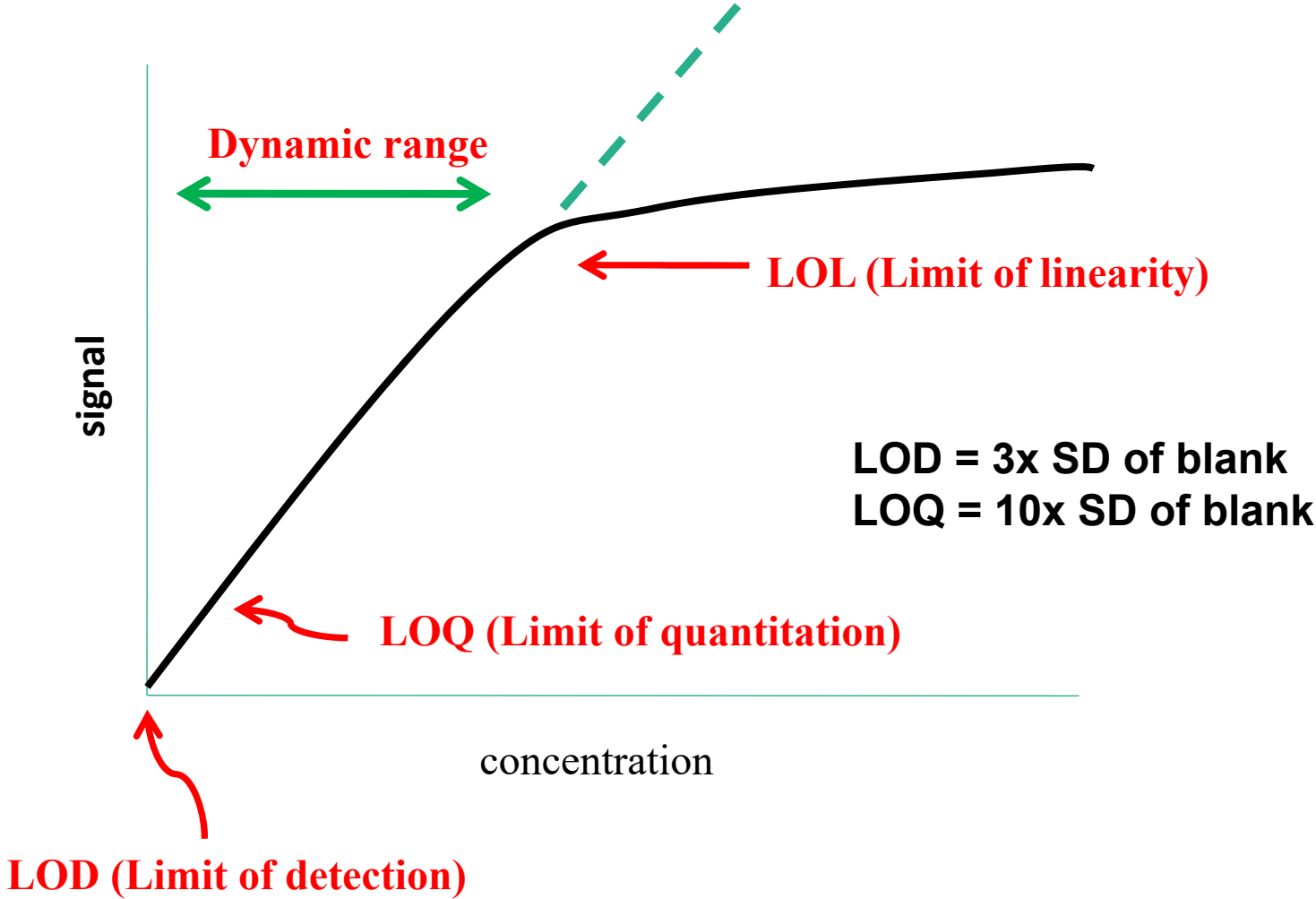


# Limit of Linear Response (LOL)

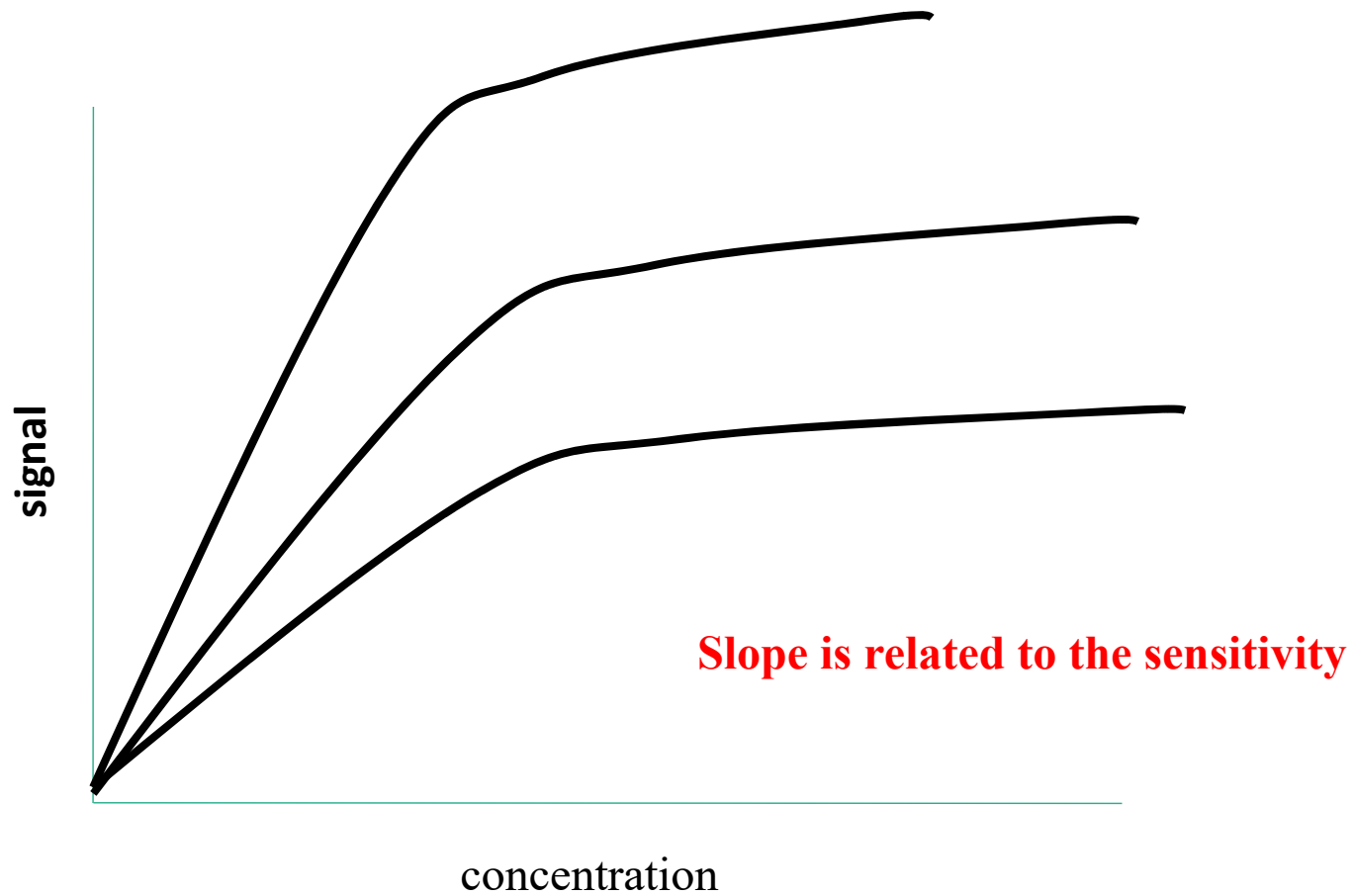
- Point of saturation for an instrument detector so that higher amounts of analyte do not produce a linear response in signal.



# Useful Range of an Analytical Method



# Method Validation- Linearity



# Method Validation- Accuracy and Precision

- Accuracy – nearness to the truth
- Compare results from more than one analytical technique
- Analyze a blank spiked with known amounts of analyte.

Precision - reproducibility

# Method Validation- LOD and LOQ

- Detection limit (lower limit of detection – smallest quantity of analyte that is “statistically” different from the blank.
- HOW TO:
  - Measure signal from n replicate samples ( $n > 7$ )
  - Compute the standard deviation of the measurements
  - Signal detection limit:  $y_{dl} = y_{blank} + 3s$
  - $y_{sample} - y_{blank} = m \cdot \text{sample concentration}$
- Detection limit:  $3s/m$
- Lower limit of quantitation (LOQ) :  $10s/m$

**Example: sample concentrations: 5.0, 5.0, 5.2, 4.2, 4.6, 6.0, 4.9 nA**

**Blanks: 1.4, 2.2, 1.7, 0.9, 0.4, 1.5, 0.7 nA**

**The slope of the calibration curve for high conc.  $m = 0.229 \text{ nA}/\mu\text{M}$**

**What is the signal detection limit and the minimum detectable concentration?**

**What is the lower limit of quantitation?**



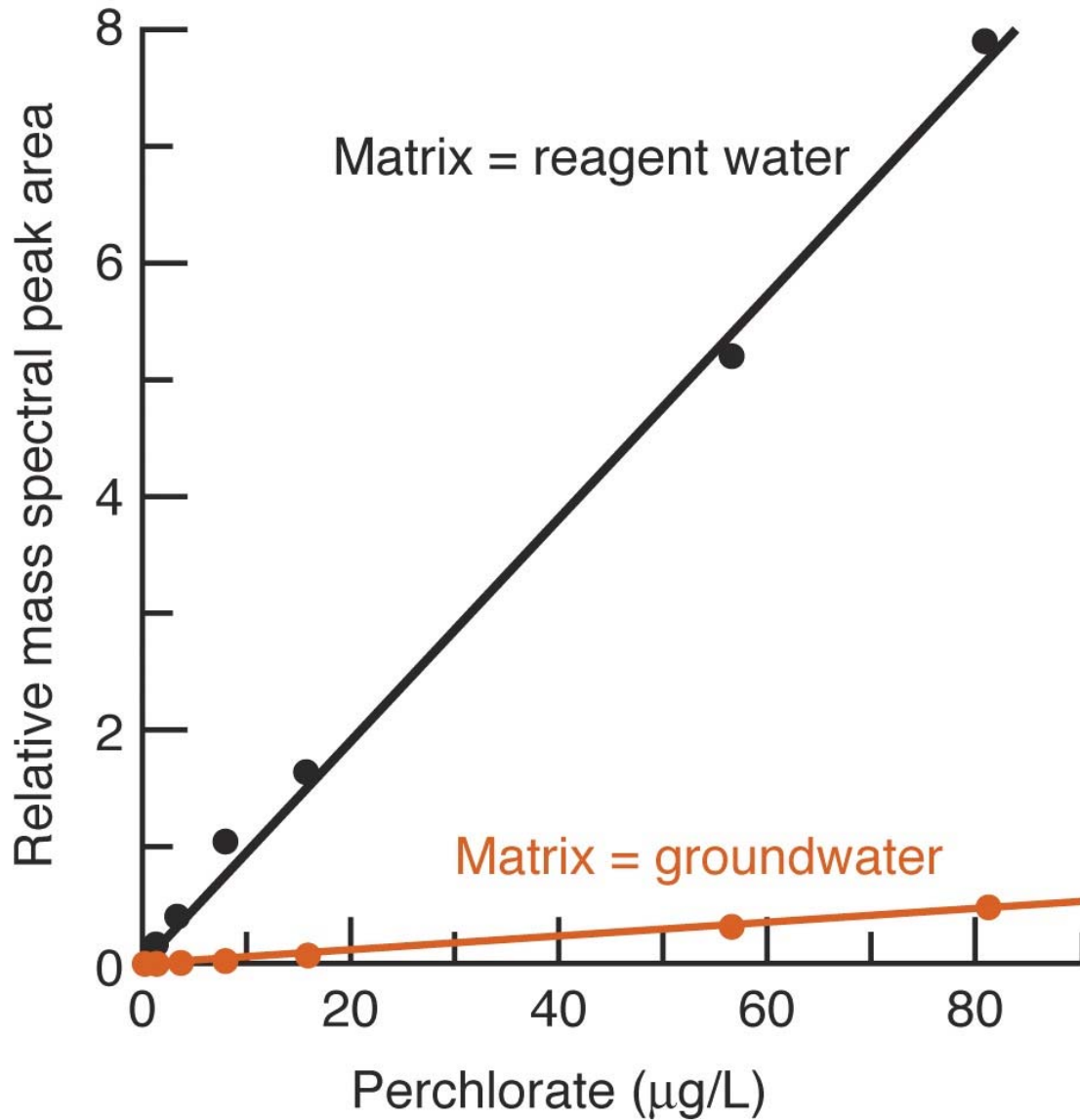
# Standard Addition

- Standard addition is a method to determine the amount of analyte in an unknown.
  - In standard addition, known quantities of analyte are added to an unknown.
  - We determine the analyte concentration from the increase in signal.
- Standard addition is often used when the sample is unknown or complex and when species other than the analyte affect the signal.
  - The **matrix** is everything in the sample other than the analyte and its affect on the response is called the **matrix effect**

# The Matrix Effect

- The matrix effect problem occurs when the unknown sample contains many impurities.
- If impurities present in the unknown interact with the analyte to change the instrumental response or themselves produce an instrumental response, then a calibration curve based on pure analyte samples will give an incorrect determination

# Calibration Curve for Perchlorate with Different Matrices



Perchlorate ( $\text{ClO}_4^-$ ) in drinking water affects production of thyroid hormone.  $\text{ClO}_4^-$  is usually detected by mass spectrometry (Ch. 22), but the response of the analyte is affected by other species, so you can see the response of calibration standards is very different from real samples

# Calculation of Standard Addition

- The formula for a standard addition is:

$$\frac{[X]_i}{[S]_f + [X]_f} = \frac{I_x}{I_{S+X}}$$

[X] is the concentration of analyte in the initial (i) and final (f) solutions, [S] is the concentration of standard in the final solution, and I is the response of the detector to each solution.

- But,

$$[X]_f = [X]_i \left( \frac{V_0}{V_f} \right) \quad \text{and} \quad [S]_f = [S]_i \left( \frac{V_s}{V_f} \right)$$

If we express the diluted concentration of analyte in terms of the original concentration, we can solve the problem because we know everything else.

# Standard Addition Example

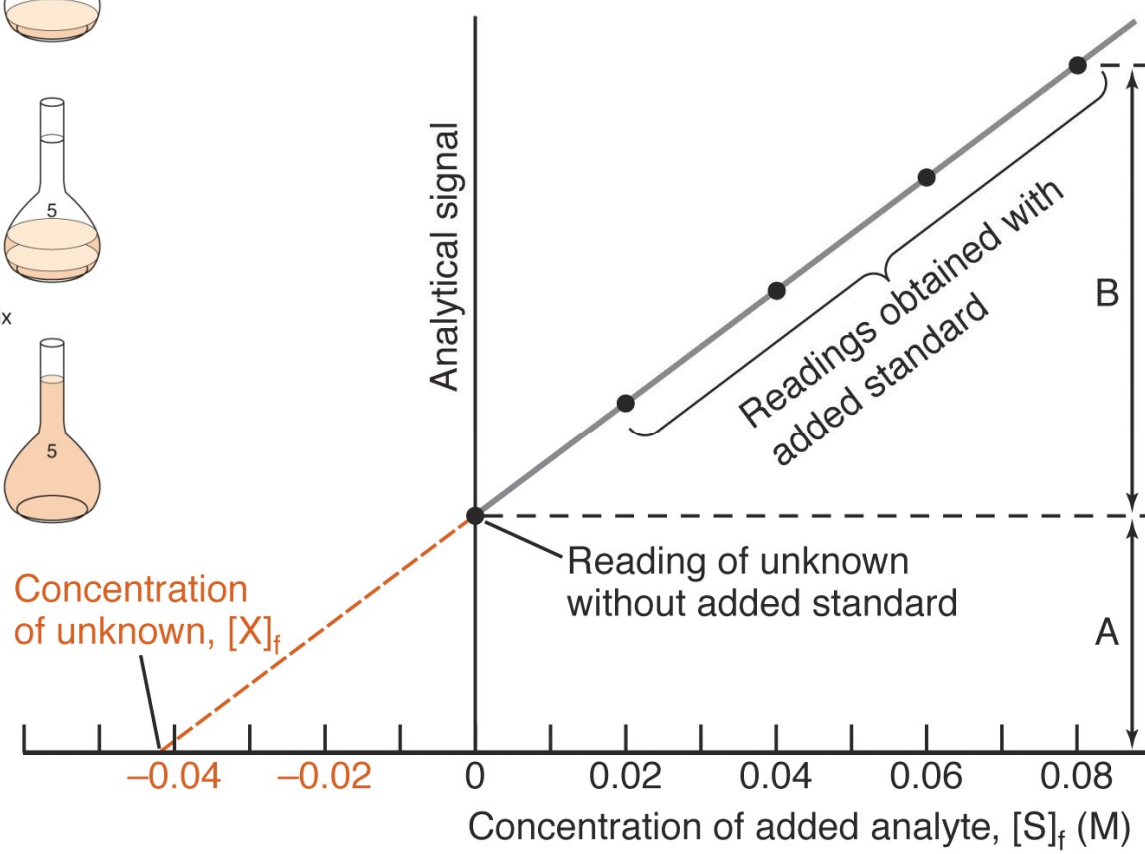
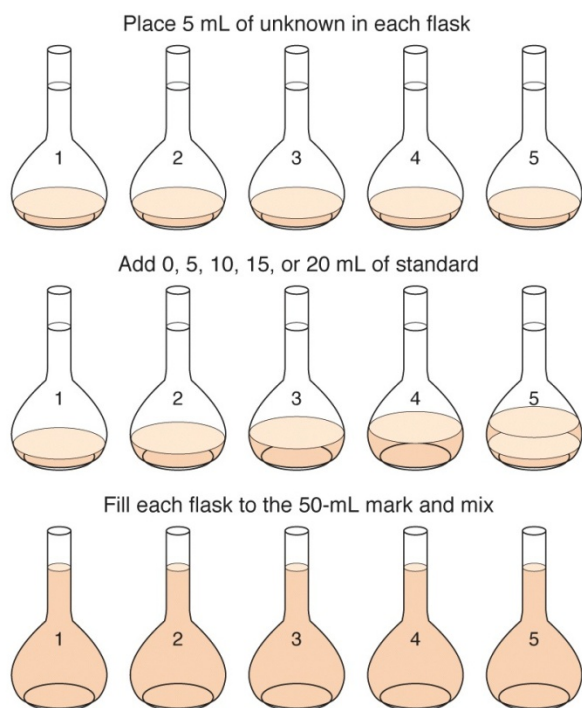
- Serum containing  $\text{Na}^+$  gave a signal of 4.27 mV in an atomic emission analysis. 5.00 mL of 2.08 M NaCl were added to 95.0 mL of serum. The spiked serum gave a signal of 7.98 mV. How much  $\text{Na}^+$  was in the original sample?

$$[\text{X}]_{\text{f}} = [\text{X}]_{\text{i}} \left( \frac{95.0 \text{ mL}}{100.0 \text{ mL}} \right) = 0.950[\text{X}]_{\text{i}}$$

$$[\text{S}]_{\text{f}} = [\text{S}]_{\text{i}} \left( \frac{V_{\text{s}}}{V_{\text{f}}} \right) = (2.08 \text{ M}) \frac{5.00 \text{ mL}}{100.0 \text{ mL}} = 0.104 \text{ M}$$

$$\frac{[\text{Na}^+]_{\text{i}}}{0.104 \text{ M} + 0.950[\text{Na}^+]_{\text{f}}} = \frac{4.27 \text{ mV}}{7.98 \text{ mV}} \quad [\text{Na}^+]_{\text{i}} = 0.113 \text{ M}$$

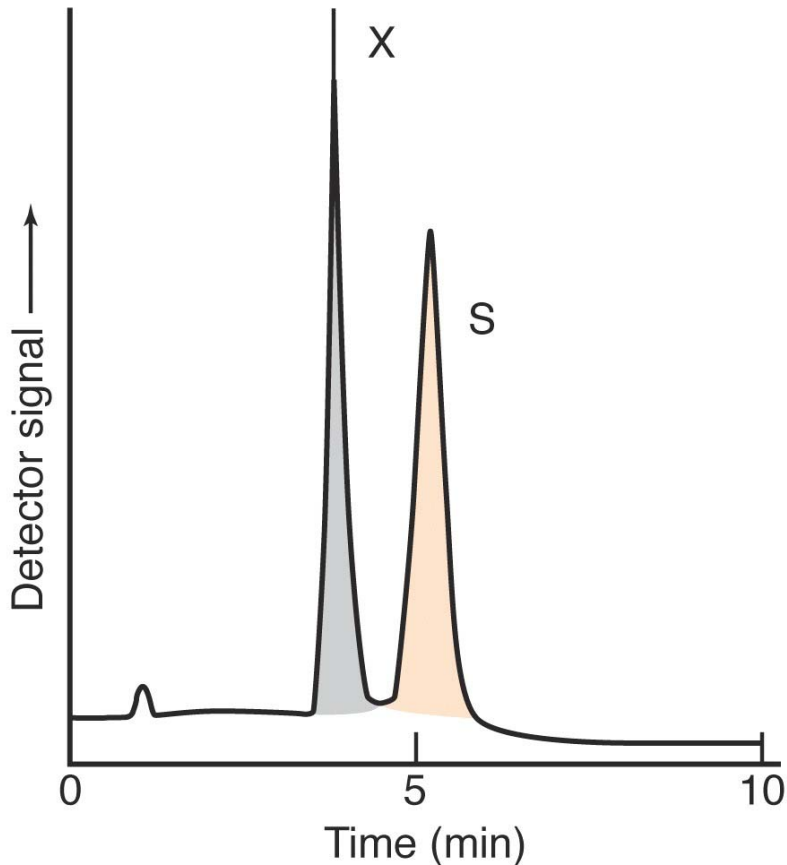
# Standard Additions Graphically



# Internal Standards

- An **internal standard** is a known amount of a compound, different from the analyte, added to the unknown sample.
- Internal standards are used when the detector response varies slightly from run to run because of hard to control parameters.
  - *e.g.* Flow rate in a chromatograph
- But even if absolute response varies, as long as the *relative* response of analyte and standard is the same, we can find the analyte concentration

# Response Factors



For an internal standard, we prepare a mixture with a known amount of analyte and standard. The detector usually has a different response for each species, so we determine a **response factor** for the analyte:

$$\frac{A_X}{[X]} = F \left( \frac{A_S}{[S]} \right)$$

[X] and [S] are the concentrations of analyte and standard after they have been mixed together.

$$\frac{\text{Area of analyte signal}}{\text{Concentration of analyte}} = F \left( \frac{\text{area of standard signal}}{\text{Concentration of standard}} \right)$$



# Internal Standard Example

- In an experiment, a solution containing 0.0837 M Na<sup>+</sup> and 0.0666 M K<sup>+</sup> gave chromatographic peaks of 423 and 347 (arbitrary units) respectively. To analyze the unknown, 10.0 mL of 0.146 M K<sup>+</sup> were added to 10.0 mL of unknown, and diluted to 25.0 mL with a volumetric flask. The peaks measured 553 and 582 units respectively. What is [Na<sup>+</sup>] in the unknown?
- First find the response factor,  $F$

$$\frac{A_{\text{Na}}}{[\text{Na}^+]} = F \left( \frac{A_{\text{K}}}{[\text{K}^+]} \right)$$

$$F = \left( \frac{A_{\text{Na}}}{[\text{Na}^+]} \right) / \left( \frac{A_{\text{K}}}{[\text{K}^+]} \right) = \frac{423}{0.0837} / \frac{347}{0.0666} = 0.970$$

# Internal Standard Example (Cont.)

- Now, what is the concentration of  $K^+$  in the mixture of unknown and standard?

$$[K^+] = (0.146M) \left( \frac{10 \text{ mL}}{25.0 \text{ mL}} \right) = 0.05484 \text{ M}$$

- Now, you know the response factor,  $F$ , and you know how much standard,  $K^+$  is in the mixture, so we can find the concentration of  $Na^+$  in the mixture.

$$\frac{A_{Na}}{[Na^+]} = F \left( \frac{A_K}{[K^+]} \right) \quad \frac{553}{[Na^+]} = (0.970) \left( \frac{582}{0.0584 \text{ M}} \right) \quad [Na^+] = 0.0572 \text{ M}$$

- $Na^+$  unknown was diluted in the mixture by  $K^+$ , so the  $Na^+$  concentration in the unknown was:

$$[Na^+] = (0.0572 \text{ M}) \left( \frac{25 \text{ mL}}{10.0 \text{ mL}} \right) = 0.143 \text{ M}$$

Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali  
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY  
I YEAR

Course:

**EXPERIMENTAL DESIGN AND  
CHEMOMETRICS IN FOOD**  
(5 credits – 38 hours)

Teacher: Marcello Mascini  
([mmascini@unite.it](mailto:mmascini@unite.it))

The Teacher is available to answer questions at the end of the lesson, or on request by mail

# The course is split in 4 units

## UNIT 1: Univariate analysis

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the  $\chi^2$  method, Validation of the model

## UNIT 2: Multivariate analysis

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

## UNIT 3: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

## UNIT 4: Elements of Pattern recognition

cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

# **UNIT 2: Multivariate Analysis**

Correlation

Multiple linear regression

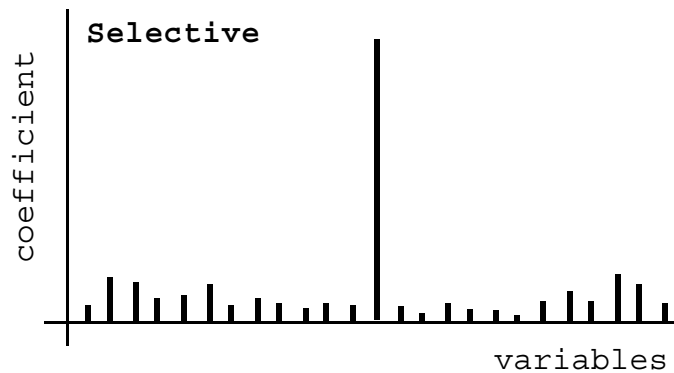
Principal component analysis (PCA)

Principal component regression (PCR) and

Partial least squares regression - (PLS)

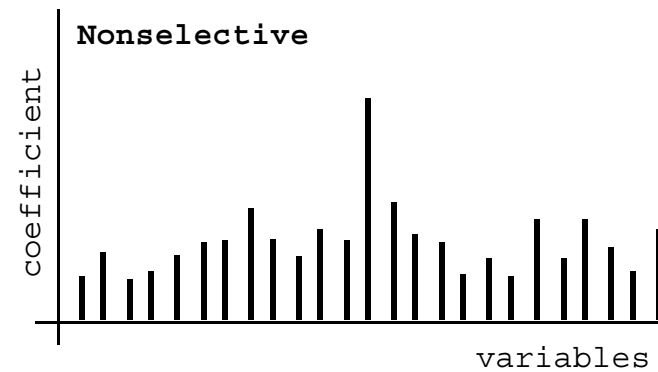
# selective and non-selective measurements

- The measurements can be selective or non-selective
  - Selective: the observation is driven by one variable
  - Non Selective: The observation is driven by many variables
- The non selective measurements are the objects of the multivariate analysis



↓

$$z \cong k_j \cdot C_j$$



↓

$$z = \sum_i k_i \cdot C_i$$

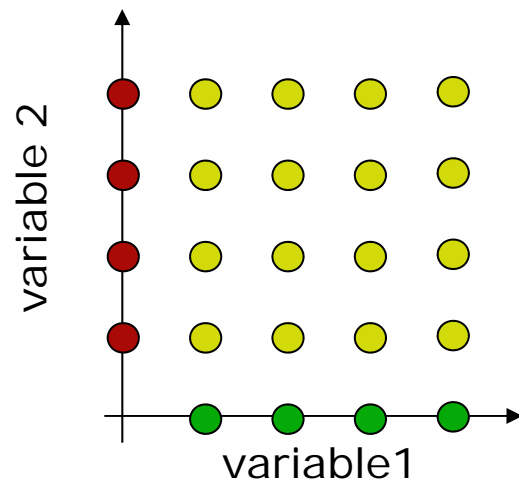
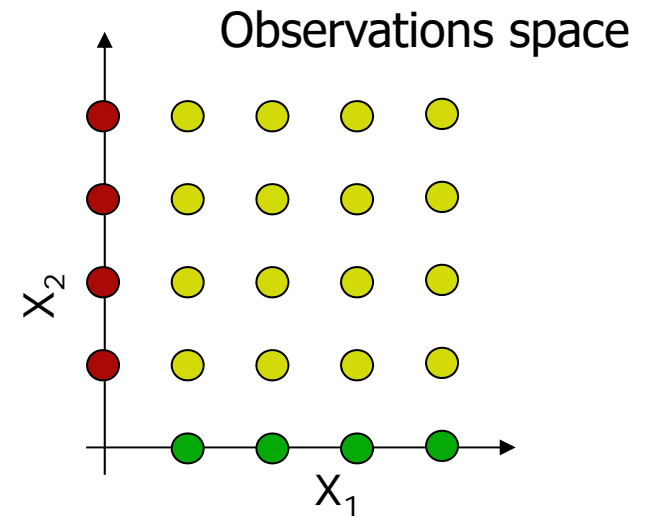
# Non selective measurements

- Example:
  - Spectroscopy
    - At a given frequency the absorbance is influenced by more than one molecule
  - Gas chromatography
    - Compounds with similar elution time can contribute to chromatographic peak
  - Sensors
    - The sensor response is given by the combination of different compounds that interfere with the sensors depending on concentration and affinity

## Variables and observations space : selective measurements

$$Y_1 = aX_1 + bX_2$$

$$Y_2 = cX_1 + dX_2$$

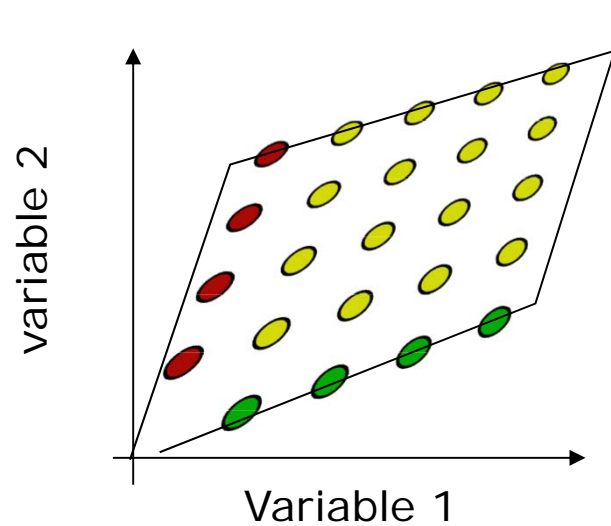


$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

**Correlation = 1 - det(K) = 0**



# Variables and observations space : selective measurements

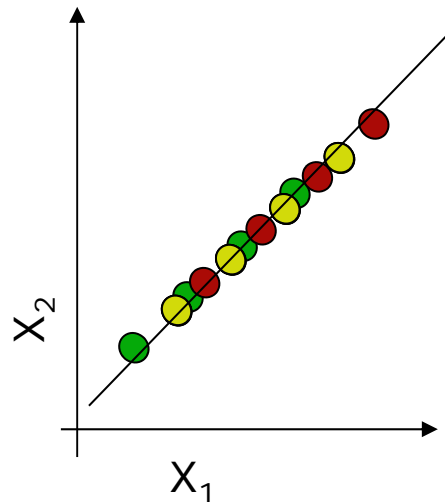


Partial correlation

$$0 < c < 1$$

$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

$$a, b, c, d \neq 0$$



Total correlation

$$c = 1$$

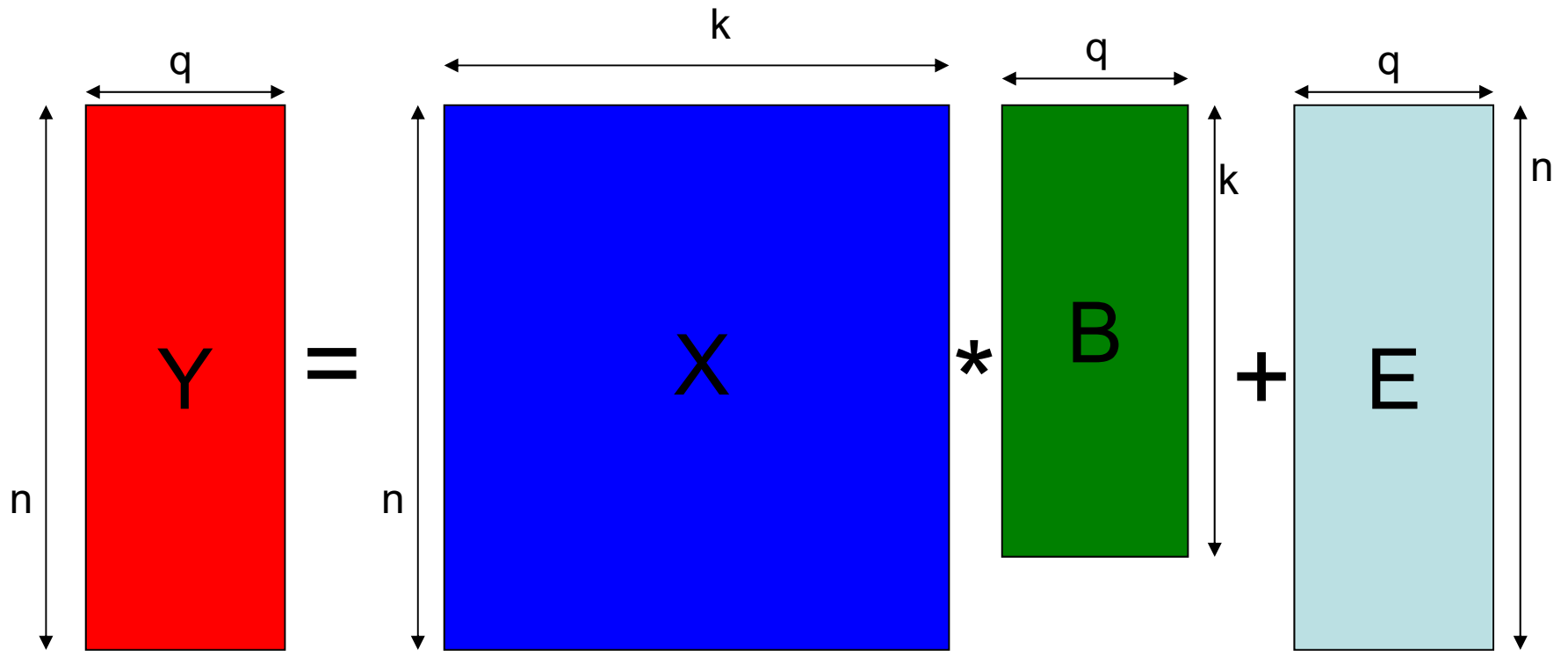
$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

$$a \cdot d - b \cdot c = 0$$

# Multiple Linear Regression

- Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable from two or more independent variables. The independent variables can be continuous or categorical .
- Multiple linear regression analysis makes several key assumptions:
- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity (The variance around the regression line is the same for all values of the predictor variable (X)).

# Multiple Linear Regression



$k = n^\circ$  observations  
 $n = n^\circ$  measurements  
 $q = n^\circ$  variables

$$Y = XB + E$$

# Multiple Linear Regression

- as for the univariate event we use two steps :
  - Calibration: using known Y and X we determine the matrix B (the slope) B
  - Procedure: known the matrix B we can have an optimized estimation of X by measuring Y
- Calibration:
  - Known X and Y the best estimation of B is given by the Gauss-Markov theorem :

$$B_{MLR} = X^+ \cdot Y$$

- If the matrix X has the maximum rank we can calculate the pseudoinverse in this way :

$$B_{MLR} = \left( X^T \cdot X \right)^{-1} \cdot X^T \cdot Y$$

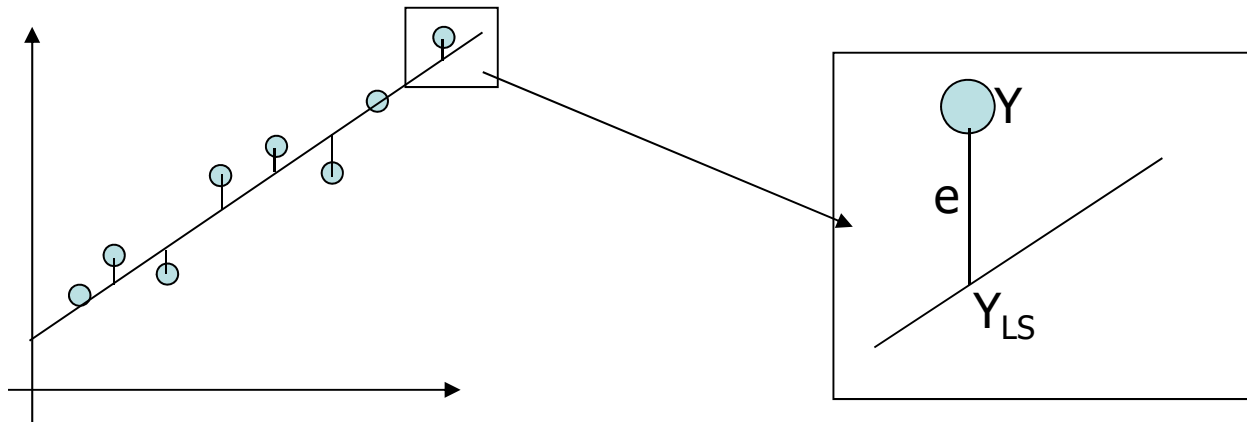
- **It means that every observation is independent from each other**

# MLR meaning

In a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator.

- In practice  $B_{MLR}$  maximize the correlation between  $X$  and  $Y$
- Geometrically the  $Y$  orthogonal projection In a subspace of  $X$
- $\Omega$  is a matrix in a subspace of  $X$

$$Y_{MLR} = X \cdot B_{MLR} = X \cdot (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \Omega \cdot Y$$



# MLR Limitations

Regression analysis is concerned with developing the linear regression equation by which the value of a dependent variable  $Y$  can be estimated given a value of an independent variable  $X$ . If simple regression analysis is used, the assumptions for this technique should be satisfied. The assumption required to develop the linear regression equation and to estimate the value of dependent variable by point estimation is: 1. The relationship between the two variables is linear. 2. The value of the independent variable is a set at various values, while the dependent variable is a random variable. 3. The conditional distributions of the dependent variable have equal variances.

If any interval estimation or hypothesis testing is done, additional required assumptions are: 1. **Successive observations of the dependent variable are uncorrelated.**

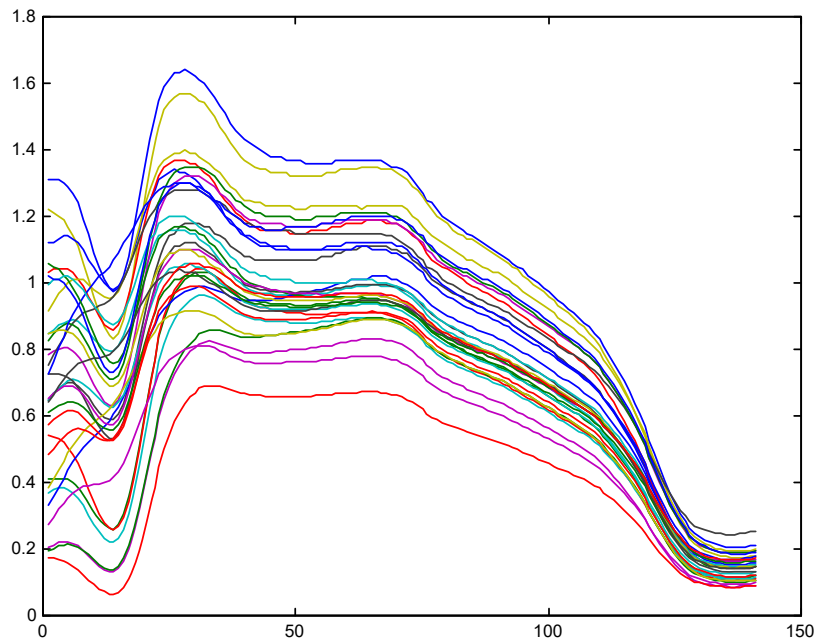
2. The conditional distributions of the dependent variable are normal distributions.

- **IF the observations of the dependent variable are correlated we have to find a method to transform them in uncorrelated observations**

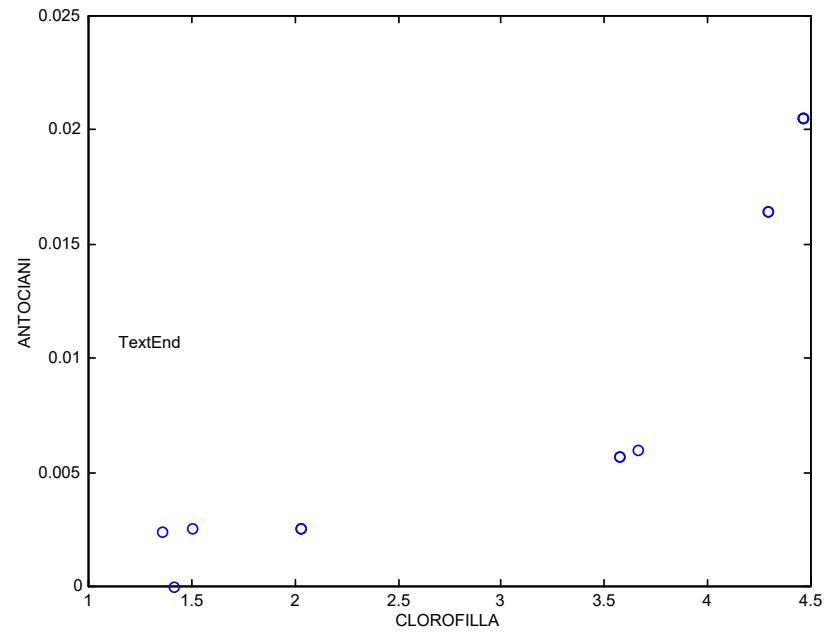
# Example

## Chlorophyll and anthocyanins in peaches using Vis-NIR

- $\text{sptectra}(\mathbf{Y})$

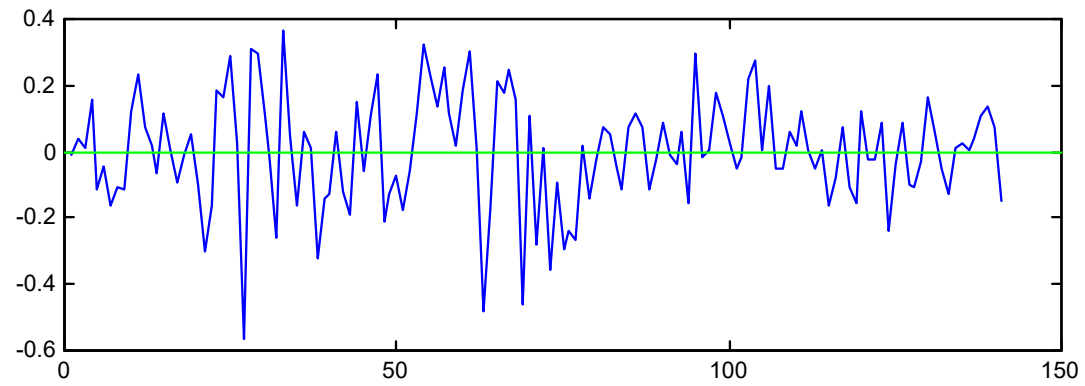
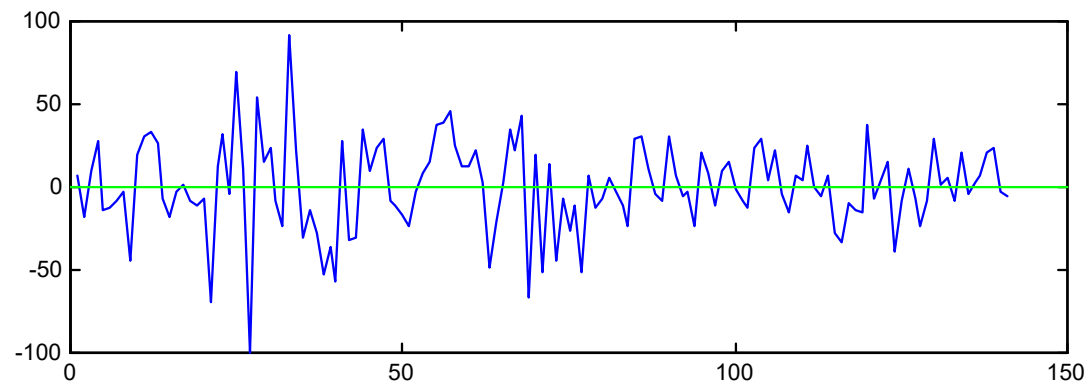


- Chlorophyll and anthocyanins ( $\mathbf{X}$ )



# results

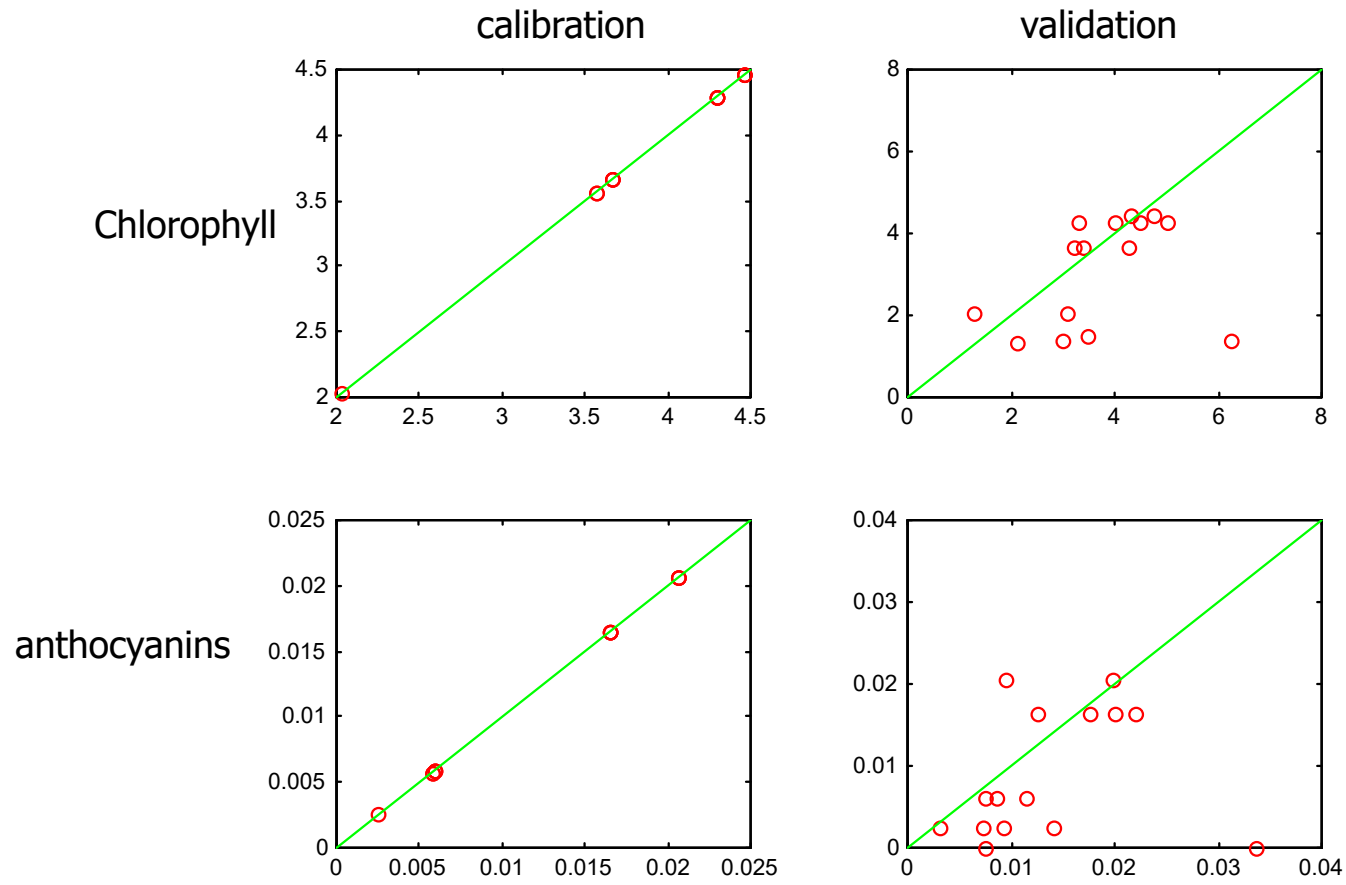
- Matrix coefficient **B**





# results

- $Y_{LS}$  and  $Y$  comparison
  - Scatter plot: x Axis: true value; y Axis : estimated value



# Principal components analysis (PCA)

Analysis of Variance

PCA and diagonalization of the covariance matrix

Scores and loadings

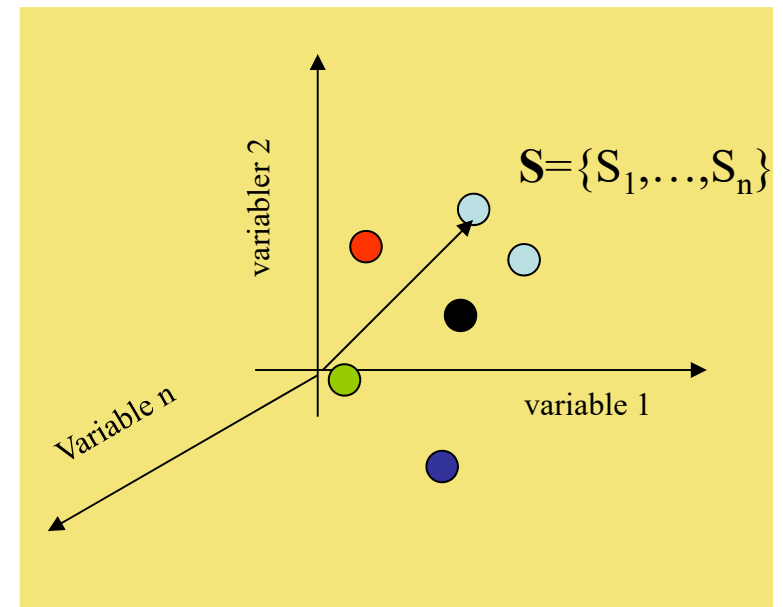
residual matrix

Applications to image analysis

Applying the multivariate regression: Principal Components Regression (PCR)

# Observations space

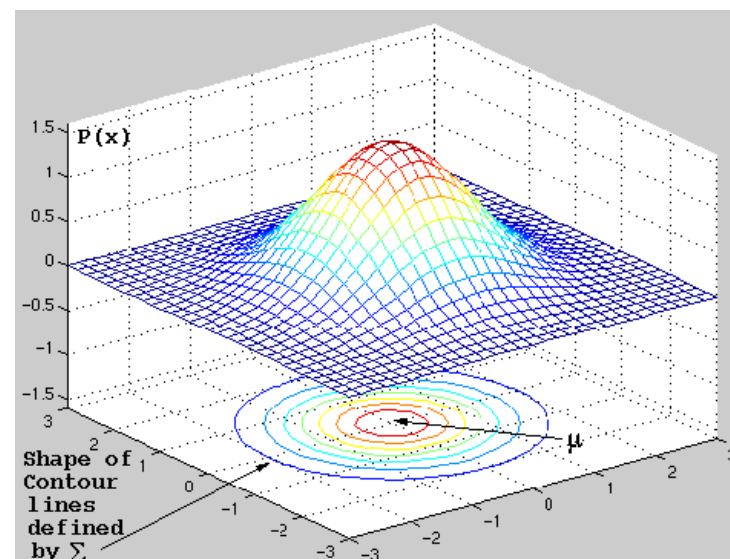
- Each multivariate measurement is represented by a vector in a space to N dimensions
- N is equal to the size of the vector that expresses the observation
- The statistical distribution of points (vectors) defines the properties of the entire data set.
- For each multivariate data we can define a PDF multivariate.
- Important: observations that describe similar samples are represented by closest points then mutual relation between distance and similarity between samples (Hypothesis of pattern recognition)



# Multivariate statistics

- the fundamental descriptors for Univariate distribution :
  - Average scalar  $\Rightarrow$  vector
  - Variance scalar  $\Rightarrow$  matrix (covariance matrix)
  - ....
- The normal distribution defined in univariate approach it keeps its importance in multivariate approach

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1d} \\ \vdots & \dots & \vdots \\ \sigma_{d1} & \dots & \sigma_{dd} \end{bmatrix}$$
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi} \sqrt{|\Sigma|}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu) \right]$$



# Covariance matrix

- In probability theory and statistics, a covariance matrix (also known as dispersion matrix or variance–covariance matrix) is a matrix whose element in the  $i, j$  position is the covariance between the  $i$ th and  $j$ th elements of a random vector. A random vector is a random variable with multiple dimensions. Each element of the vector is a scalar random variable. Each element has either a finite number of observed empirical values or a finite or infinite number of potential values. The potential values are specified by a theoretical joint probability distribution. Because the covariance of the  $i$ th random variable with itself is simply that random variable's variance, each element on the principal diagonal of the covariance matrix is just the variance of each of the elements in the vector. Every covariance matrix is symmetric. In addition, every covariance matrix is positive semi-definite.
- The covariance matrix can be done by :  $\text{COV}(xy) = x^T y$

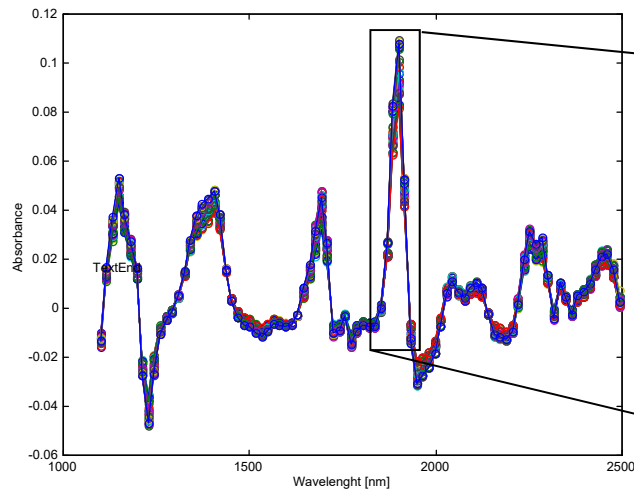
# Multicollinearity

- Multicollinearity (also co-linearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.
- In case of perfect multicollinearity the design matrix  $X$  has less than full rank, and therefore the moment matrix  $X^{(T)} * X$  cannot be inverted. Under these circumstances, for a general linear model  $Y = cX + E_r$ , the ordinary least-squares estimator does not exist.

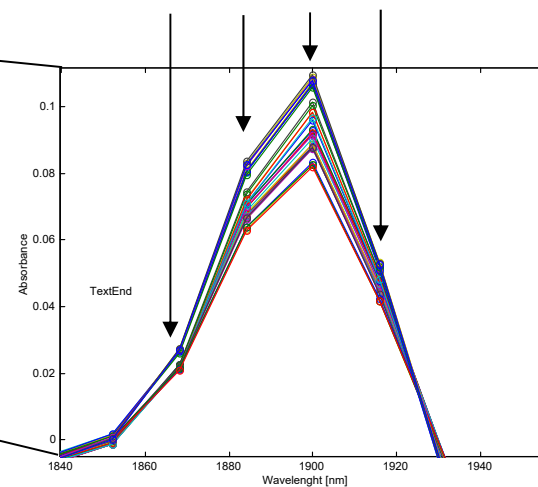
# Co-linearity example

- In an optical spectrum the spectral lines cover a range of wavelengths, this interval is generally covered by more spectral channels, so that more variables combine to form a spectral line.
- If the line is proportional to a characteristic of the sample (eg. Glucose concentration) all the spectral channels related to the line will be proportional to the sample characteristic, and then the relative variables (columns in the data matrix) will become collinear.
- co-linear variables depend quantitatively by the sample characteristics

NIR of fruits



*Collinear variables*



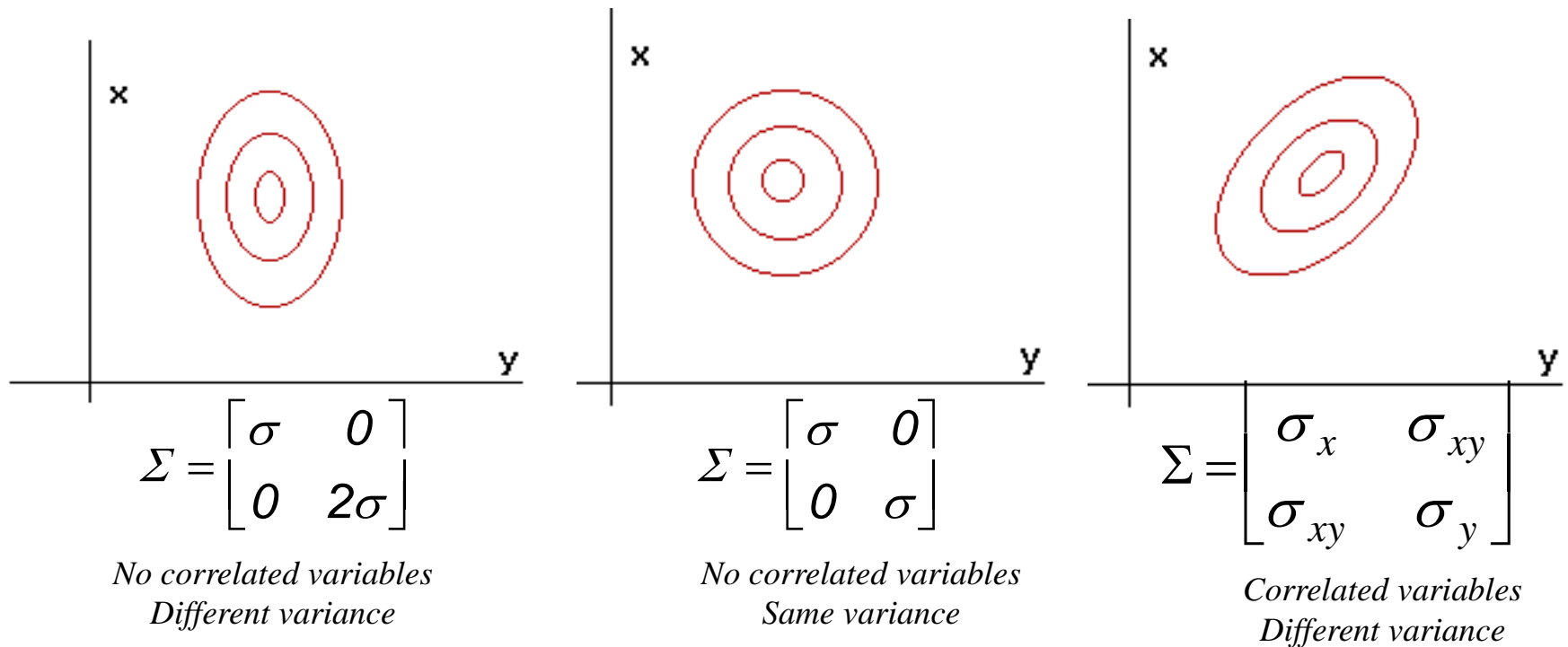
# Covariance matrix and co-linearity

- The co-linearity is expressed by the covariance matrix.
- In case of co-linearity the non-diagonal terms of the covariance matrix are nonzero.
- Remove the co-linearity it means manipulating the covariance matrix in diagonal form by introducing new latent variables.
- The principal component analysis technique allows, among other things, to obtain this result!!



# Example of covariance matrix and points probability

- Example of bivariate distribution



# Multivariate PDF and covariance matrix

- The multivariate distribution only makes sense if the covariance matrix describes the parameters correlated with each other, that is, if the matrix is not diagonal.
- In fact, for two quantities ( $x$  and  $y$ ) unrelated and independent the probability to observe simultaneously the value of  $x$  and  $y$  is simply the product of the two univariate distributions:

$$P(x, y) = P(x) \cdot P(y)$$

# The covariance matrix in canonical form

- The covariance matrix can be written in diagonal form with an appropriate change of the reference system.
- Such a reference system corresponds to the eigenvectors of the covariance matrix, ie the main ellipse constructed as quadratic form from the covariance matrix itself.
- This operation makes variables uncorrelated and the PDF as a product of the univariate PDF .
- On the other hand the new variables are no longer physical observables (object of measurement) but are linear combinations of these.
- The new variables are called Principal Components and the set of calculation procedures and interpretation of the main components is called principal component analysis (PCA)

$$a \cdot x^2 + 2b \cdot xy + c \cdot y^2 = \begin{bmatrix} x & y \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\Rightarrow \lambda_1 \cdot u^2 + \lambda_2 \cdot w^2 = \begin{bmatrix} u & w \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} u \\ w \end{bmatrix}$$

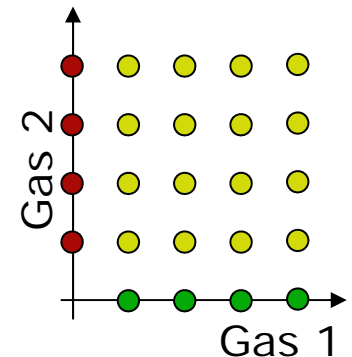
# Dimension of the data set

- If the variables of a multivariate phenomena have a certain degree of correlation then the representative vectors of the phenomenon will occupy only a portion of the observation space .
- So a variable of size N will lie in a space of smaller dimension

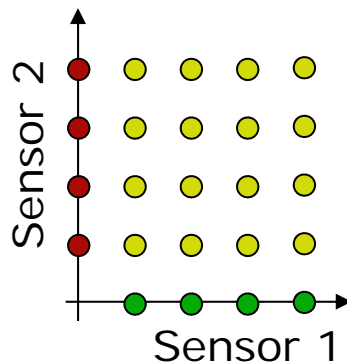
Example: linear sensors

$$\begin{cases} s_1 = k_{11} \cdot g_1 + k_{12} \cdot g_2 \\ s_2 = k_{21} \cdot g_1 + k_{22} \cdot g_2 \end{cases}$$

Independent variables

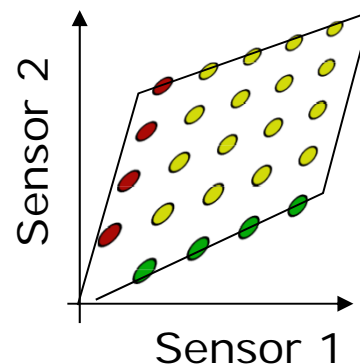


**Specific sensors**  
 $k_{12}=k_{21}=0$



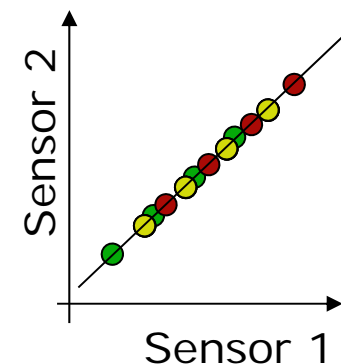
$C=0$  Dim=2

**No specific sensors but**  
 $k_{11}; k_{12}; k_{21}; k_{22}$  different



$C > 0$  and  $< 1$  Dim

**No specific but equal k**

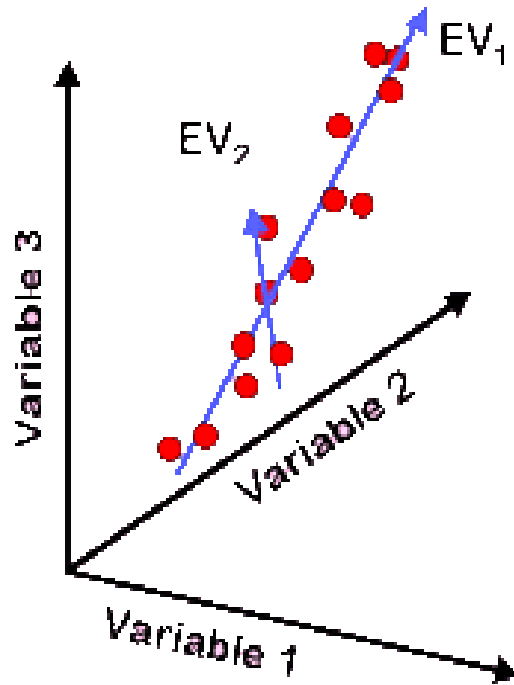


$C=1$  Dim=1

# Principal Component Analysis

- The purpose of the PCA is the representation of a data set having covariance matrix not diagonal and with a space of smaller dimension in which the same data are represented by a diagonal covariance matrix.
- The diagonalization is achieved with a coordinate rotation in the base of the eigenvectors (principal components)
- For each eigenvector it is associated an eigenvalue which corresponds to the variance of the associated component. If the original variables were partially correlated some eigenvalues have a negligible value.
- In practice the corresponding eigenvectors can be ignored by limiting the representation only to eigenvectors with the largest eigenvalues.
- Since the covariance matrix in the base of the main components is diagonal, the total variance is the sum of the variances of the individual components.

# PCA procedure



- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

# PCA

- PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

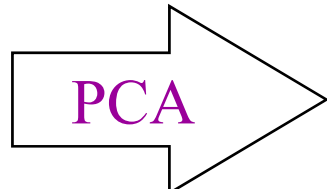
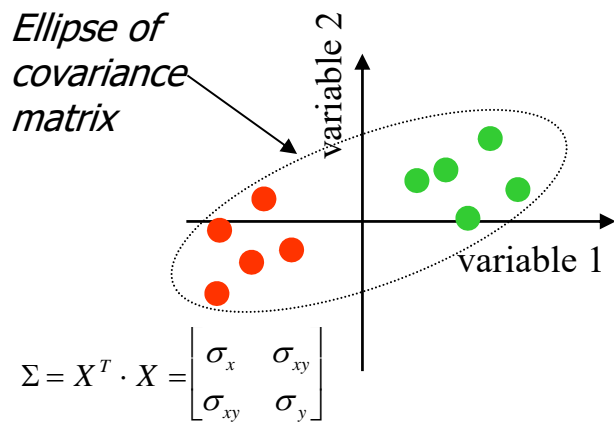
# PCA

- PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its (in some sense; see below) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.
- PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.
- PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.



# PCA

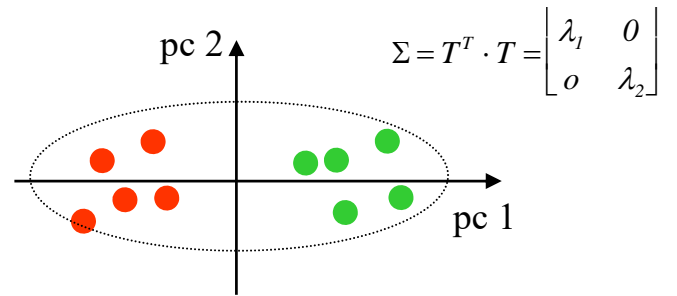
## Observation space



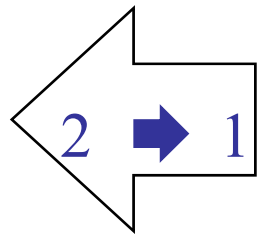
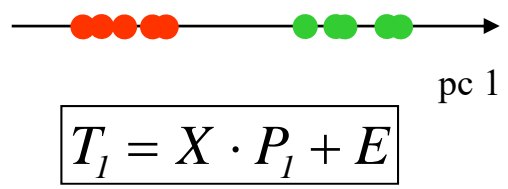
$$\Sigma = X^T \cdot X \Rightarrow \Lambda \cdot P^T$$

$$T = X \cdot P ; X = T \cdot P^T$$

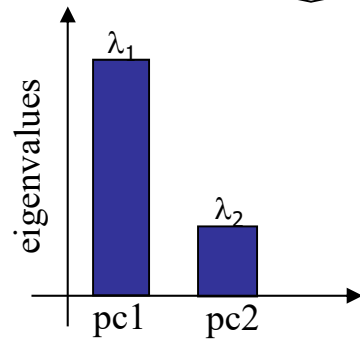
## PCA space



## Reduced space



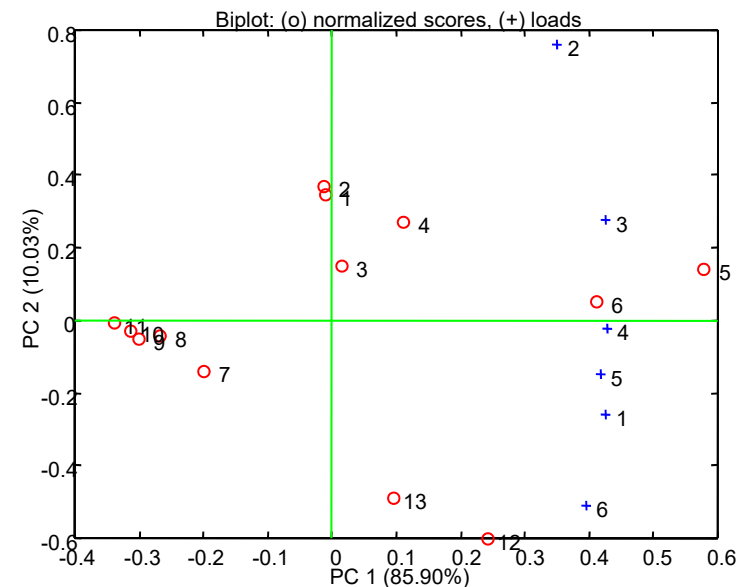
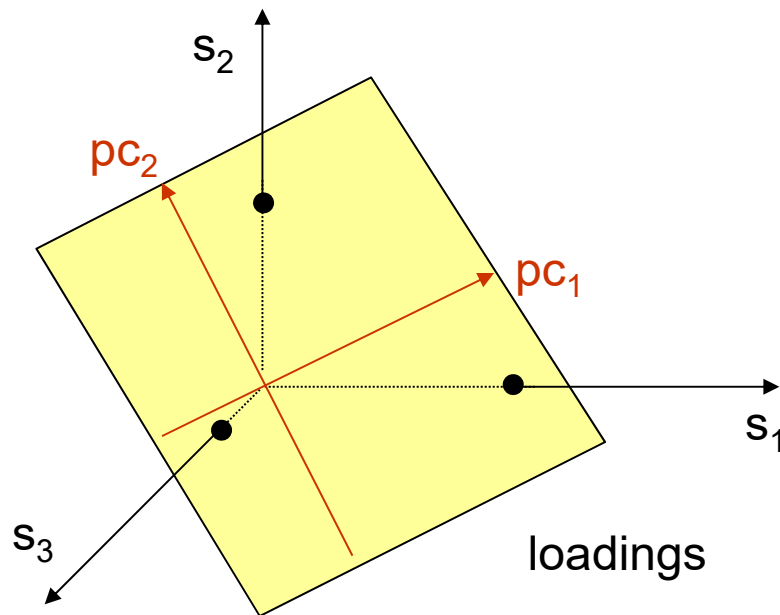
Dimensions reduction



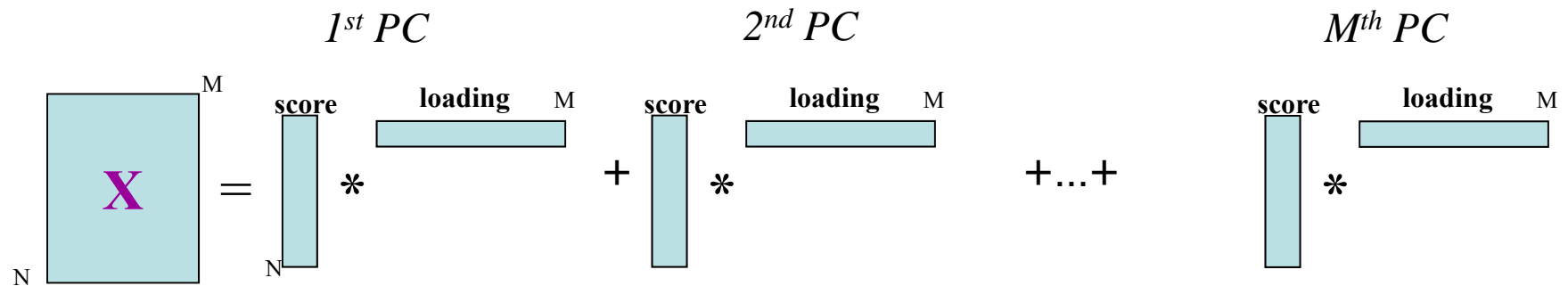
*eigenvalues: a PC has a greater information content than an other*

# PCA: scores e loadings

- The new coordinates of the vectors corresponding to the observations (the rows of the matrix  $x$ ) in the base of the principal components are called scores
- The coefficients of the linear combinations that define the principal components are called loadings
- The loading therefore provides a measure of the contribution of each observable to the principal components
- The loadings are also represented as scores as they are the projection of the original axes in the subspace identified the principal components, and scores and loadings can be plotted together



# PCA matrix Decomposition

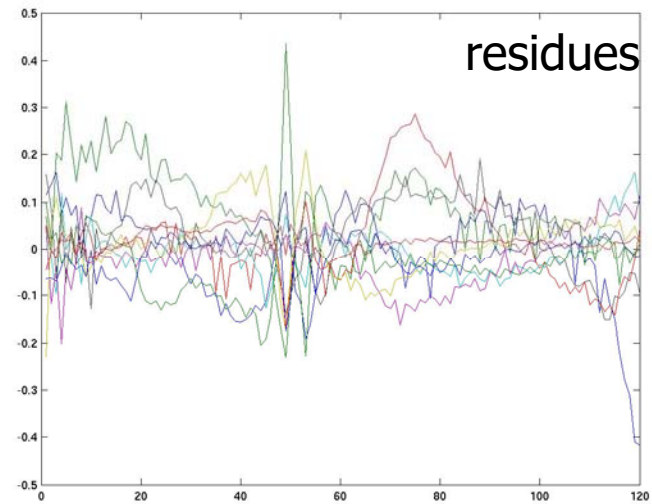
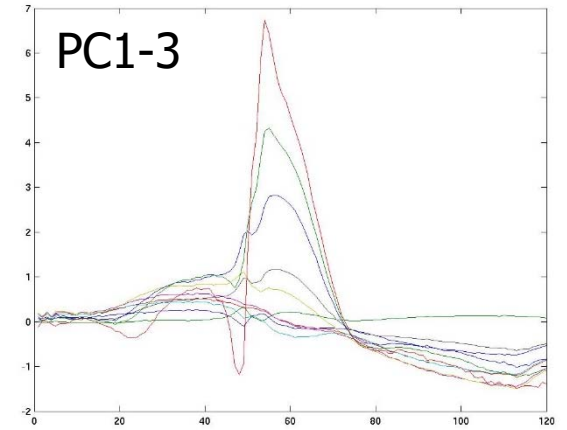
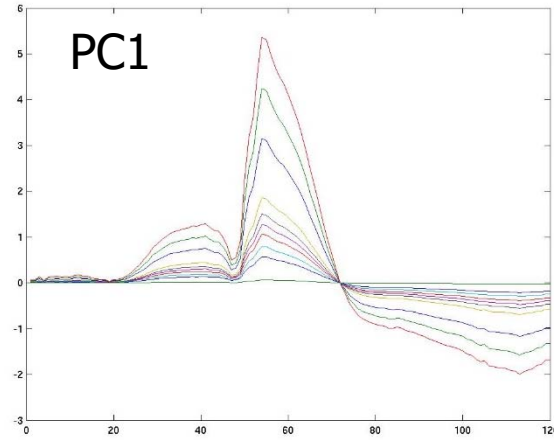
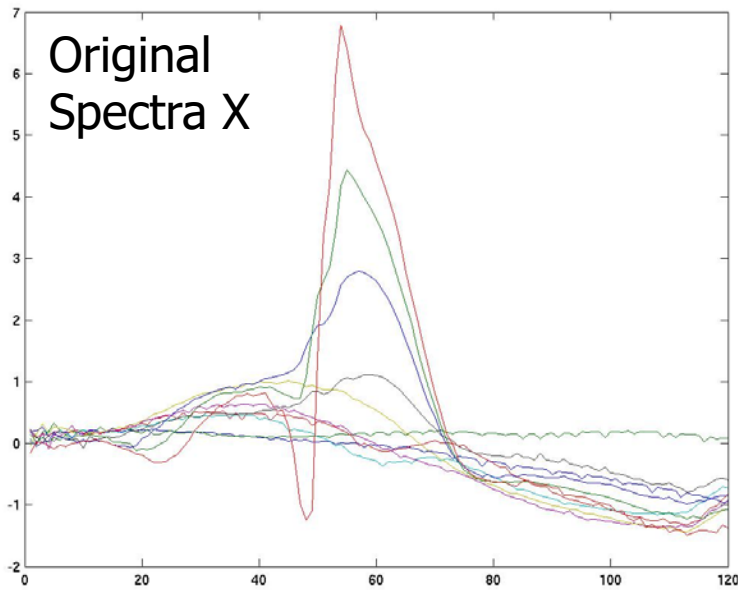


$$X_{nm} = S_{np} \cdot L_{pm}^T + \text{Residual}$$

# PCA, correlation and noise

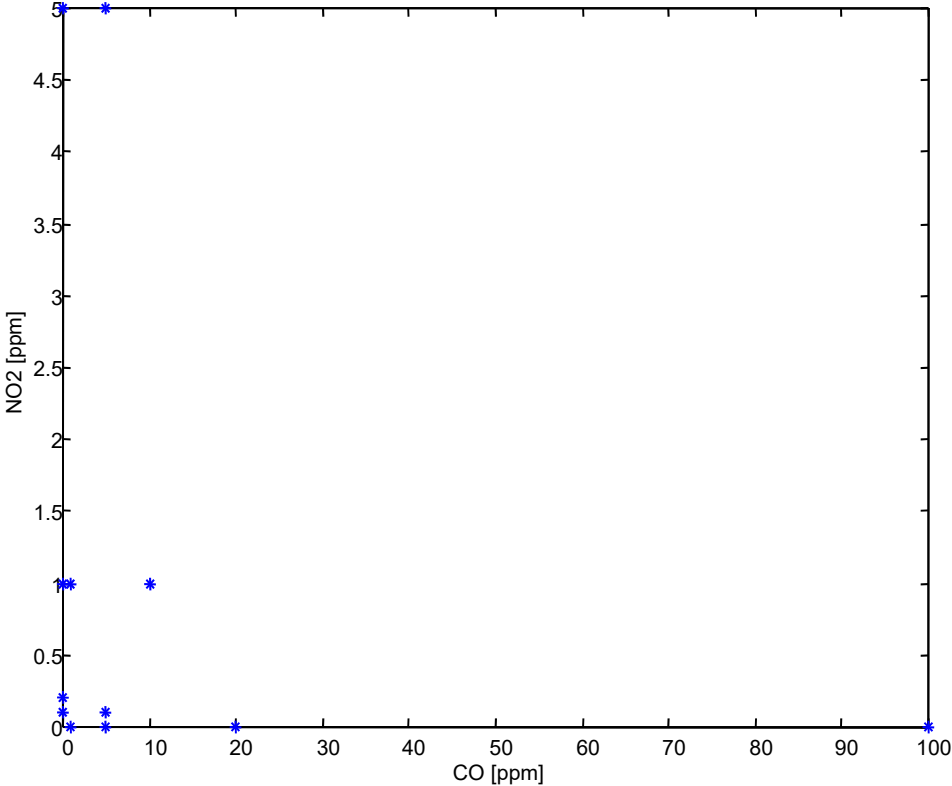
- Noise is an additional stochastic term that belongs to every observation.
- The noise is the term that makes the measurement a statistical operation.
- The principal components describe the directions of maximum correlation between the data, for which the higher-order PC are oriented towards the directions of maximum correlation and those of lower order towards the poor correlation directions
- Decomposing the major components of higher order means holding the maximum correlation directions and remove those that are no-correlated. In no-correlated directions where there is the noise
- The PCA therefore is a method for reducing the noise in a set of multivariate data.
- example: spectroscopy, GC, ...

# removing noise : Reflectance Anisotropy Spectroscopy

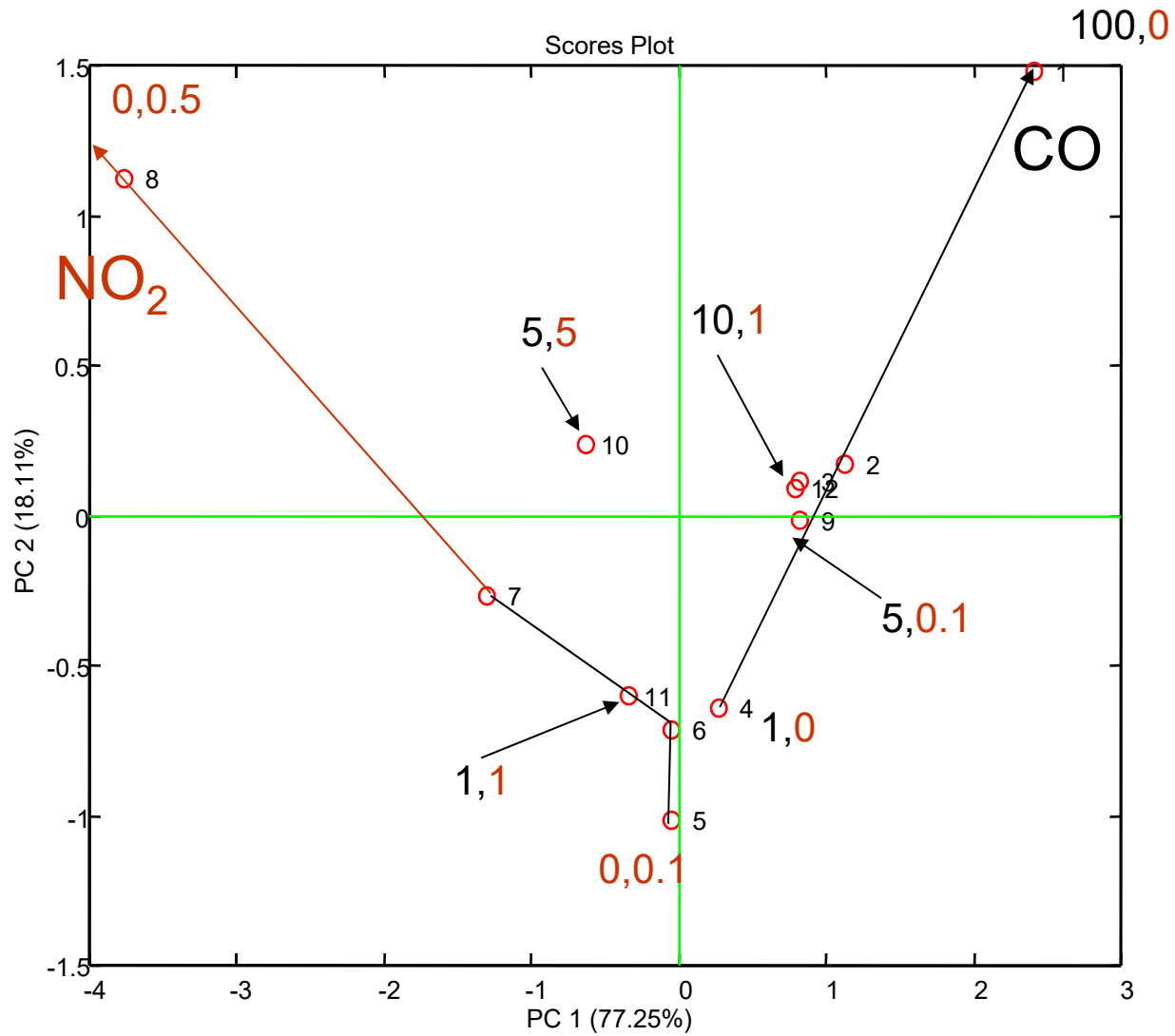


# 3 SnO2 sensors for 2 gas

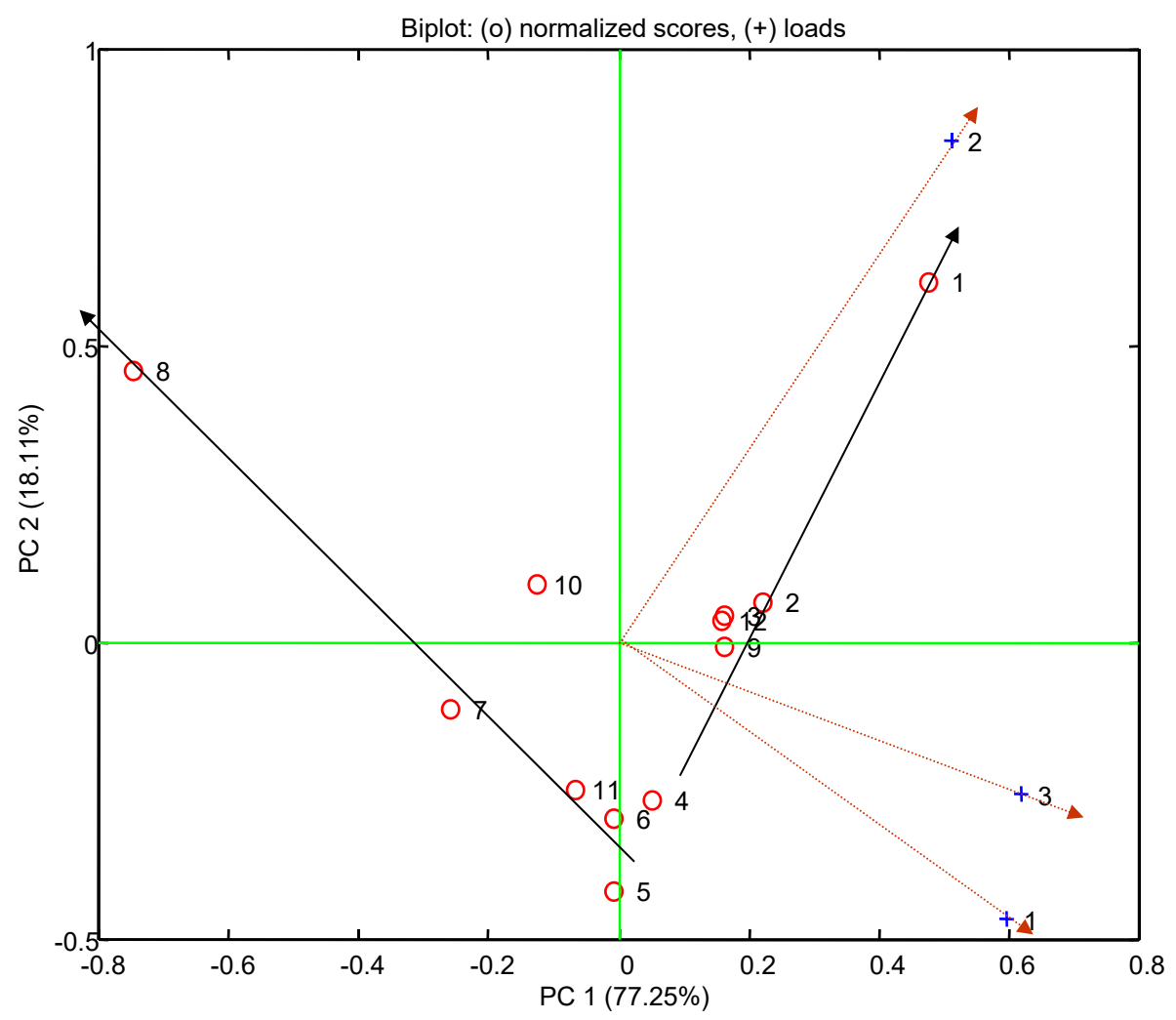
$G_r/G_i$			CO	NO <sub>2</sub>
0.25482	0.63354	0.77832	100.00	0.0000
0.093899	0.27108	0.39692	20.000	0.0000
0.043410	0.23361	0.079543	5.0000	0.0000
0.0097185	0.043353	-0.0021311	1.0000	0.0000
-0.018016	-0.053860	-0.073648	0.0000	0.10000
-0.028579	0.0023183	-0.36593	0.0000	0.20000
-0.25167	-0.028831	-2.4367	0.0000	1.0000
-1.6960	-0.075037	-3.8650	0.0000	5.0000
0.057521	0.21072	0.16777	5.0000	0.10000
-0.13089	0.13002	-2.1376	5.0000	5.0000
-0.068079	-0.0027190	-0.90852	1.0000	1.0000
0.050023	0.22771	0.020198	10.000	1.0000



# PCA score plot

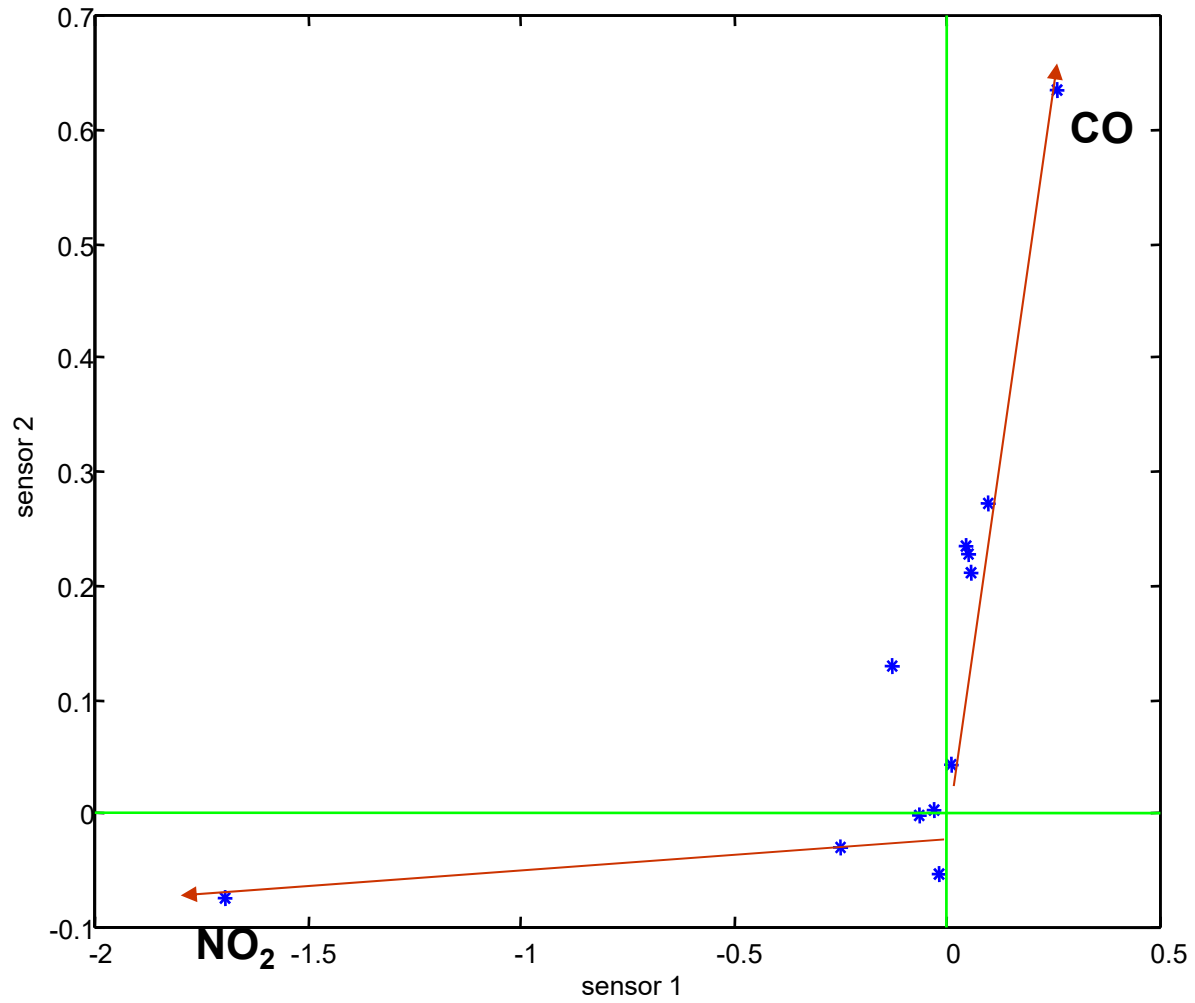


# PCA bi-plot





# Sensor 1 vs sensor 2

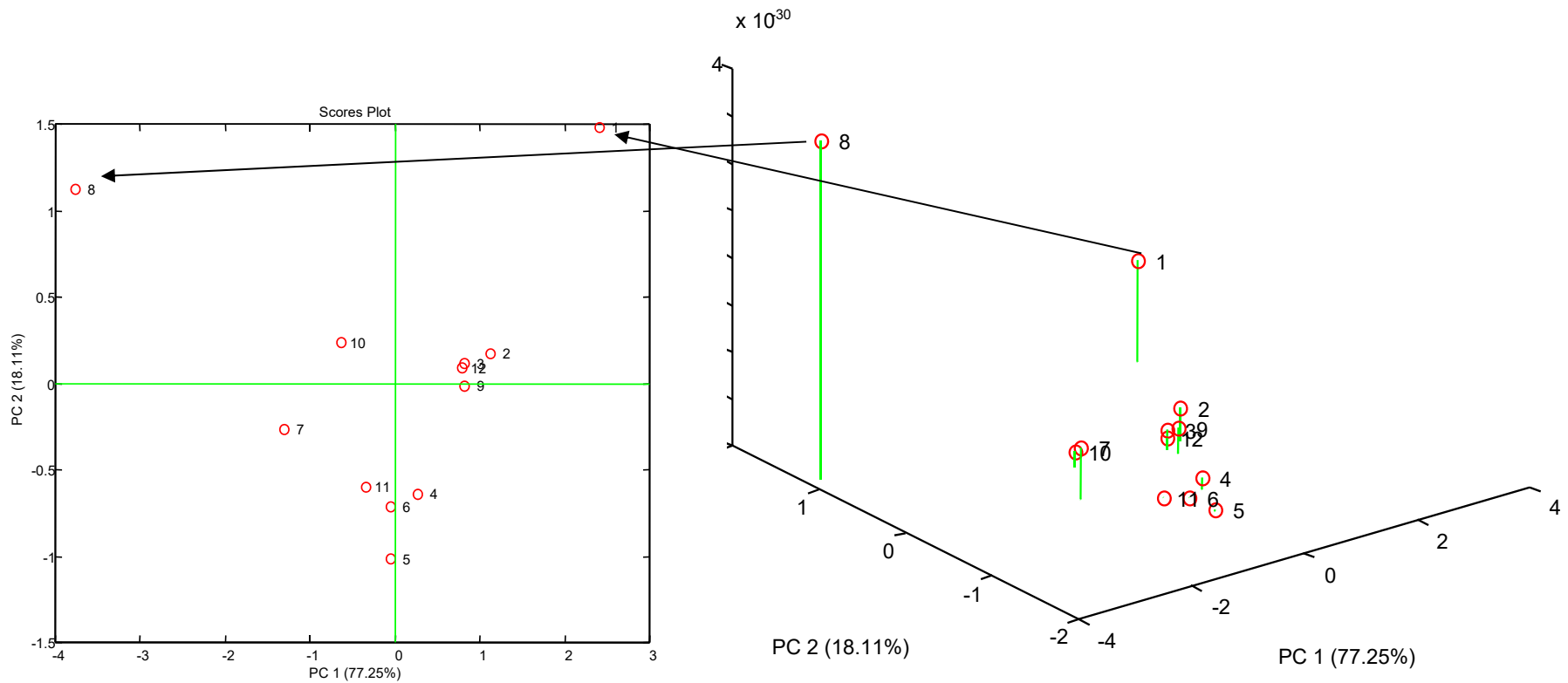


# Residual of PCA representation (leverage)

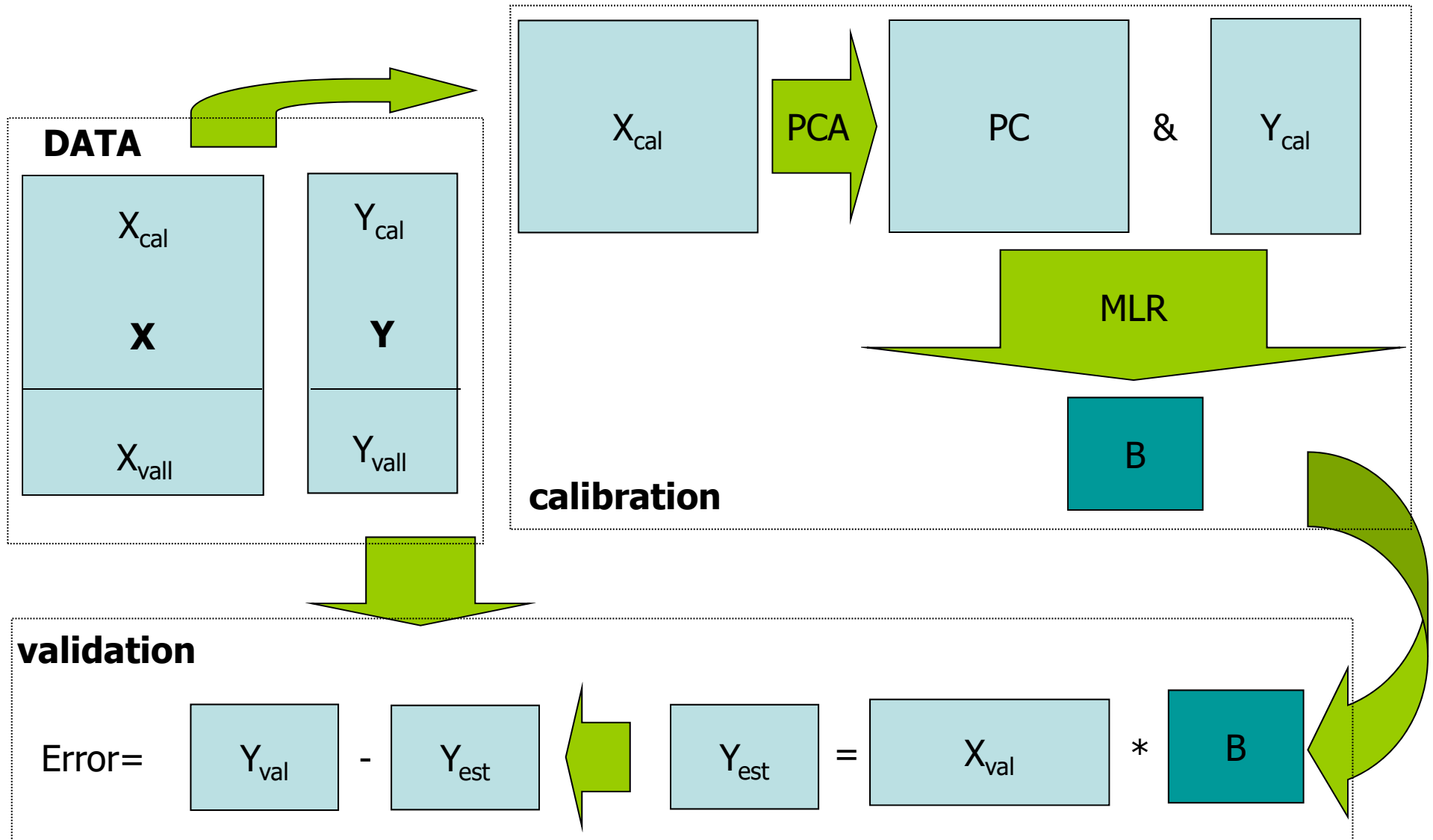
$$x_i = a \cdot s_1 + b \cdot s_2 + \dots + n \cdot s_n$$

$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$

Scores Plot



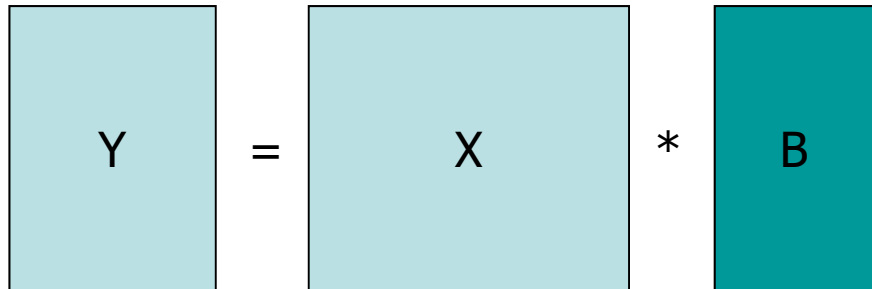
# PCR procedure



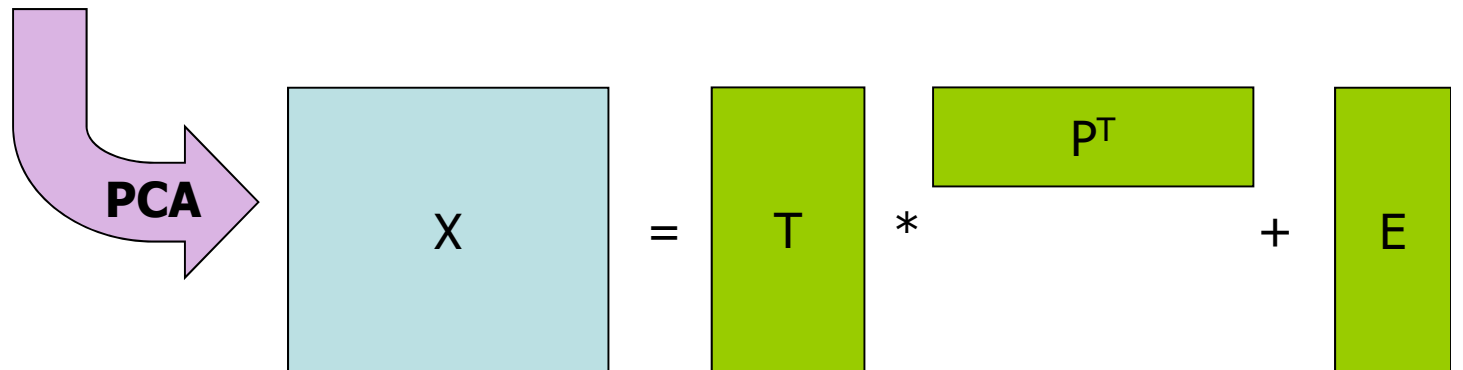
# PCR algorithm

$$Y = X * B$$

Original problem

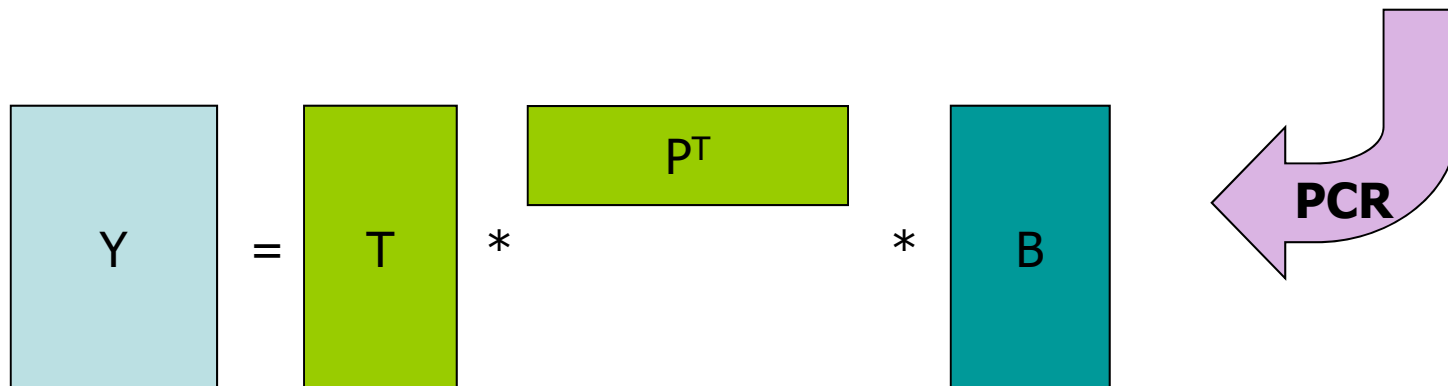


PCA

$$X = T * P^T + E$$


$$Y = T * P^T * B$$

PCR

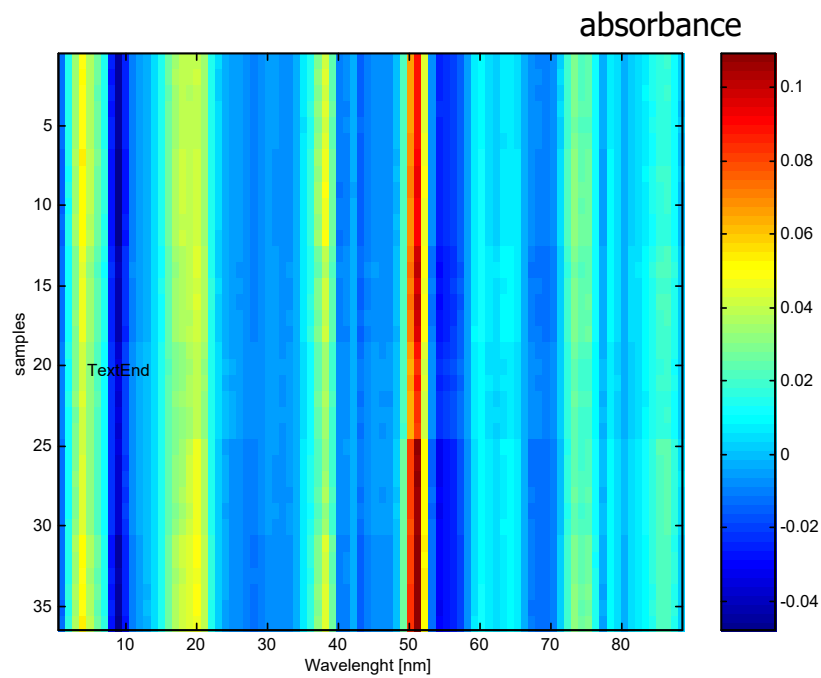


## Example: NMR fruits spectra

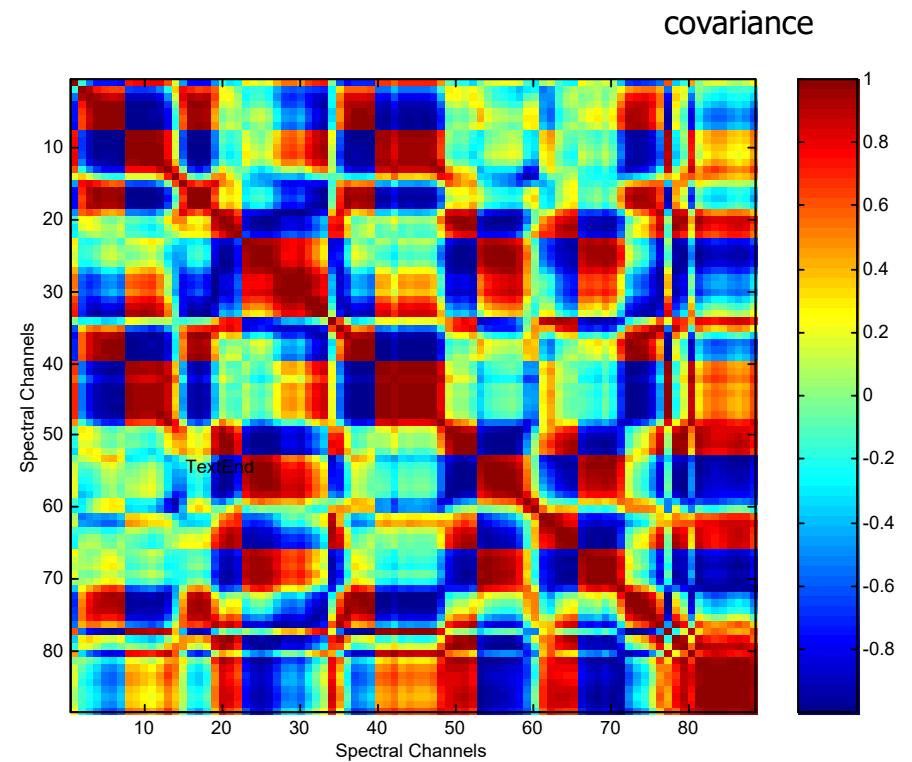
- We carried out 36 NIR spectra of fruits and we want to create a model for humidity and total acidity.
- Each spectrum is formed by 88 variables corresponding to the spectral channels in the range of 1.1-2.5 microns.
- For each fruit was measured humidity and acidity with other methods.
- We want the two parameters of Y from the spectrum X .Therefore is necessary to estimate the parameter K

$$Y_{1 \times 2} = X_{1 \times 88} \cdot K_{88 \times 2}$$

# X matrix and covariance matrix

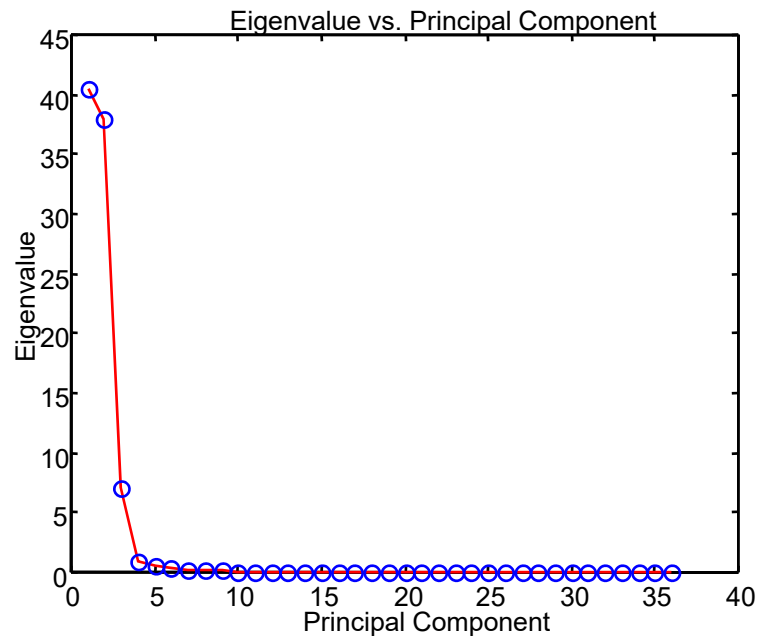


- high collinearity
- The high correlation of blocks (+ and -) correspond to the spectral lines represented by colored columns in absorbance matrix



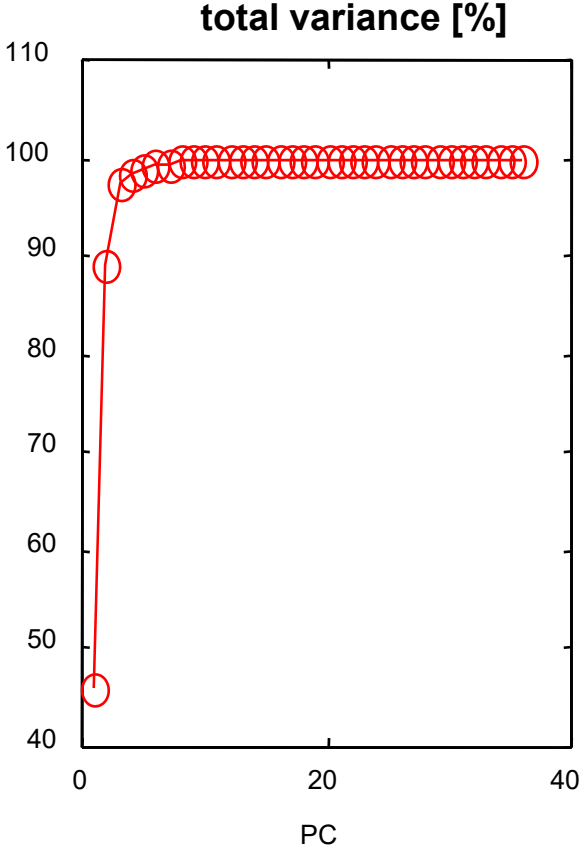
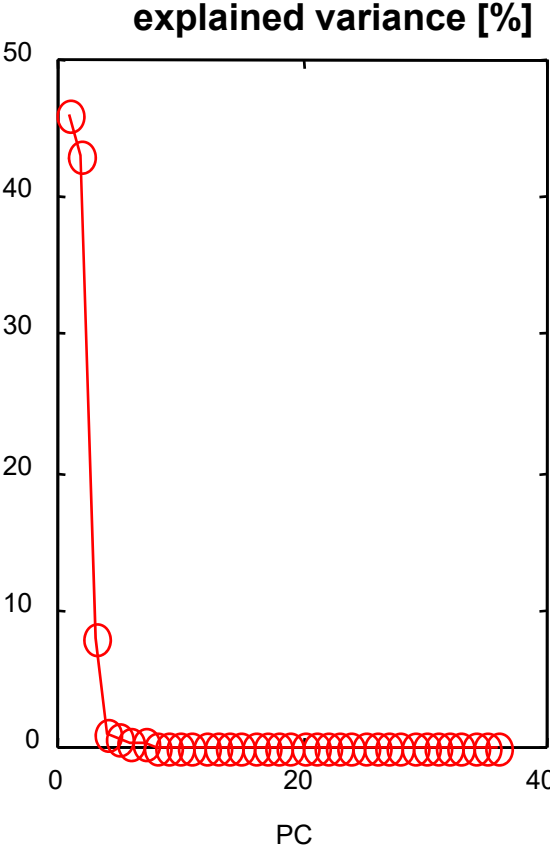
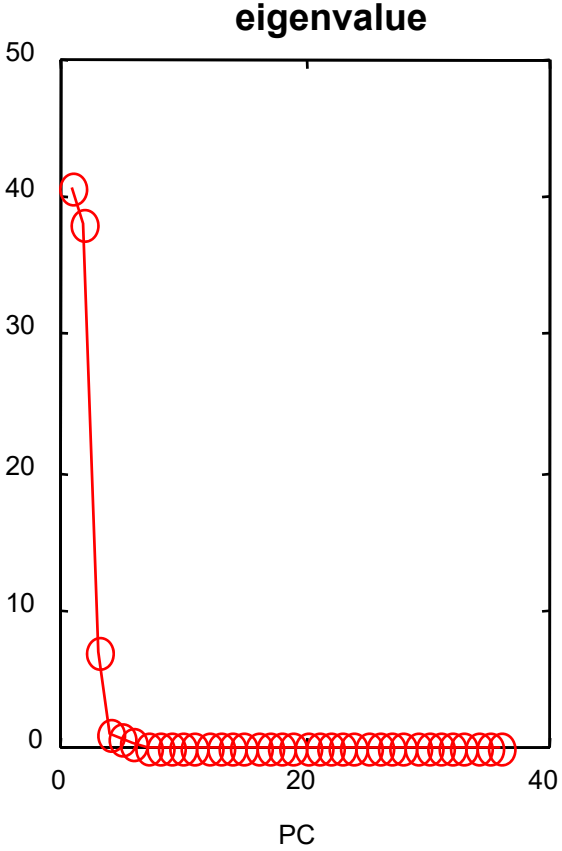
# PCA computation

- The spectra average is reduced to zero therefore if the normal distribution assumption is satisfied, the whole information is in the covariance matrix.
- Eigenvectors and eigenvalues calculation



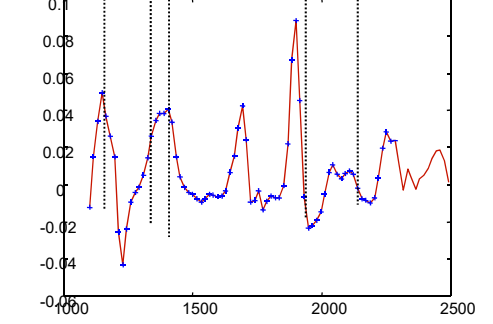
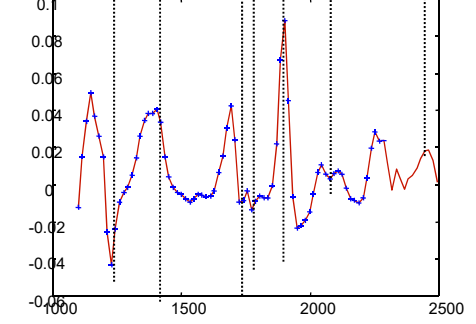
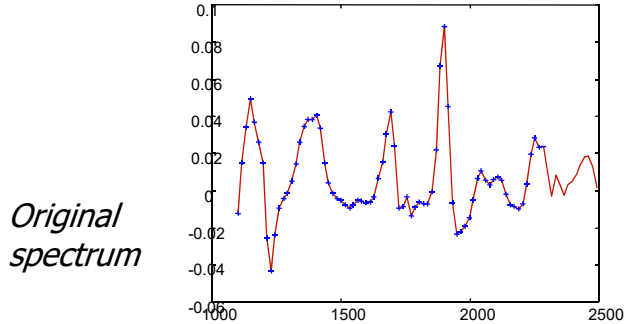
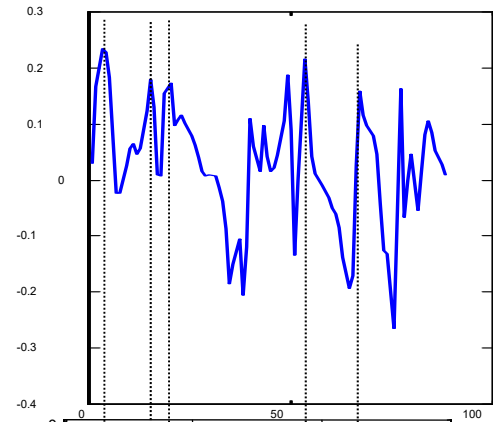
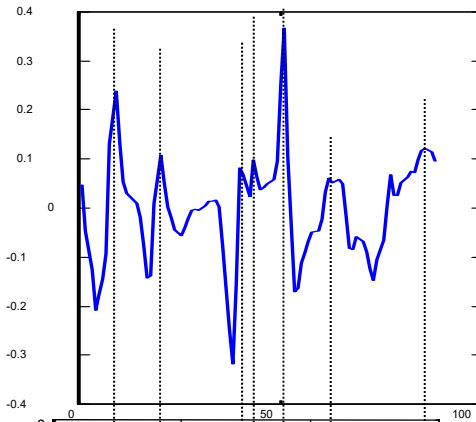
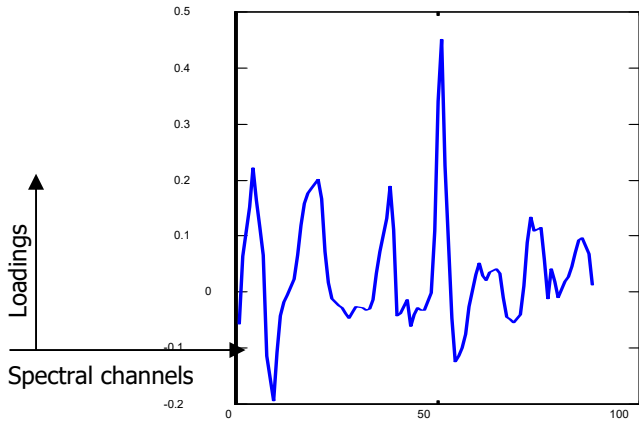
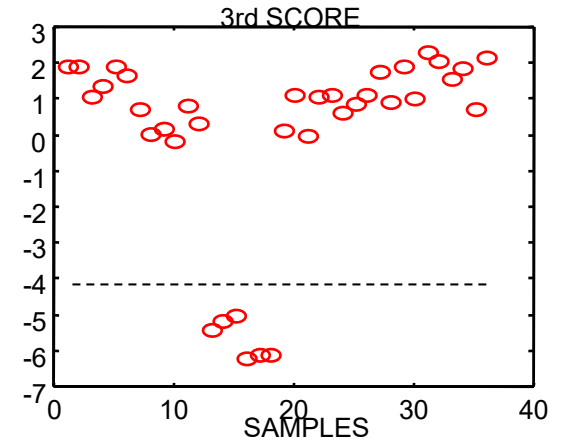
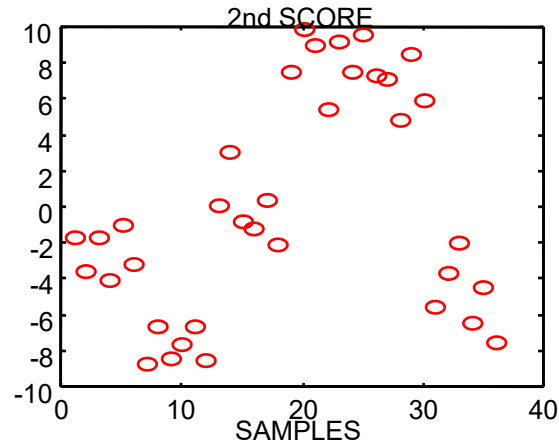
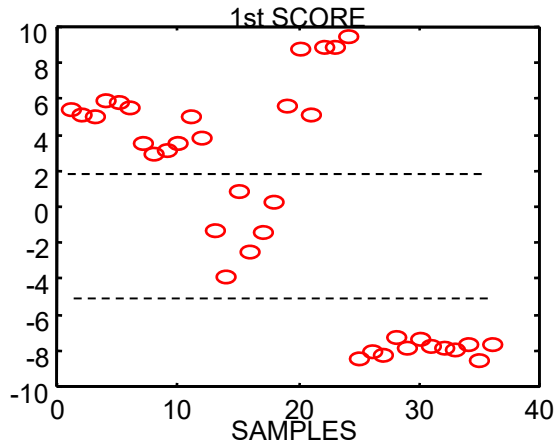
- The first 3 eigenvalues have values significantly different to zero.
- The 88 spectra, vectors in a dimensional space of 88, are largely limited to a subspace dimension of 3.

# Eigenvalue and variance

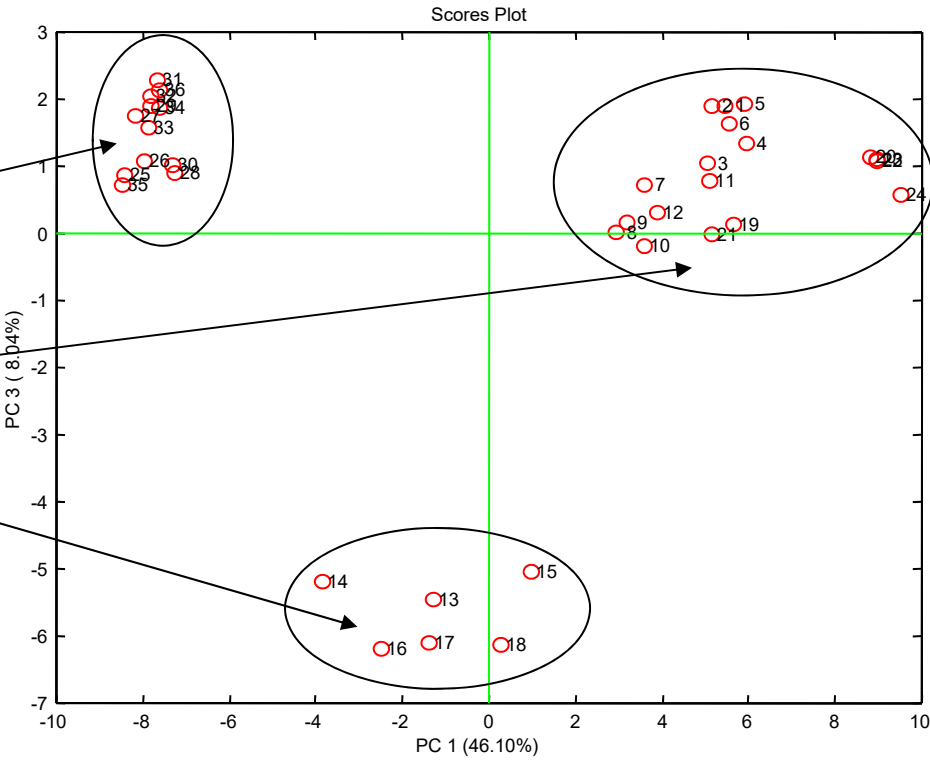
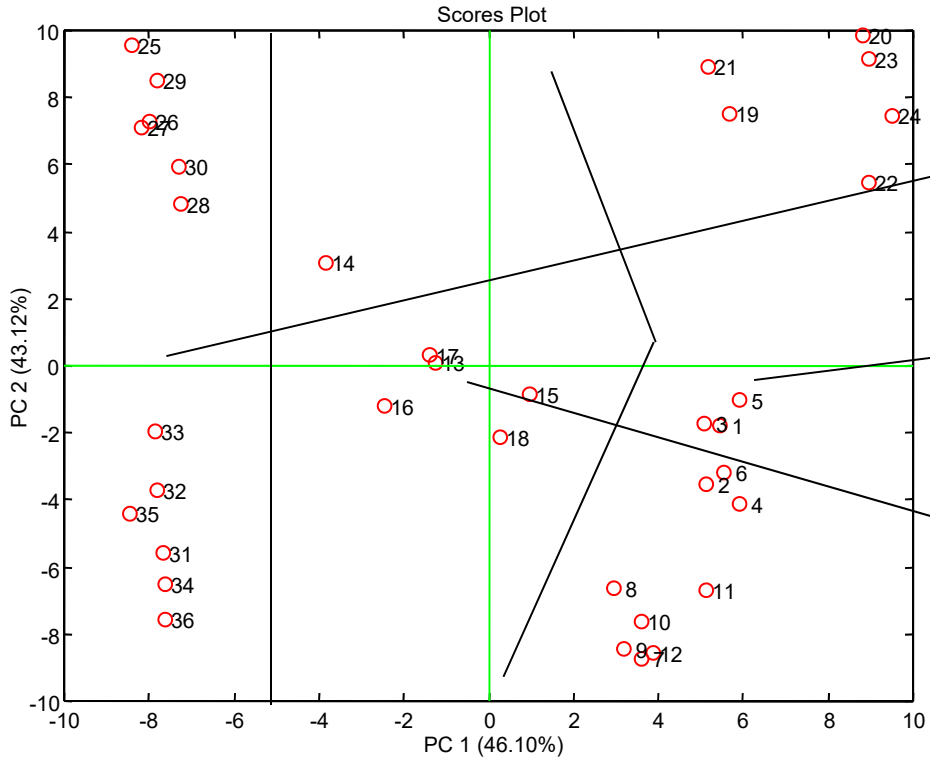




# Scores e loadings

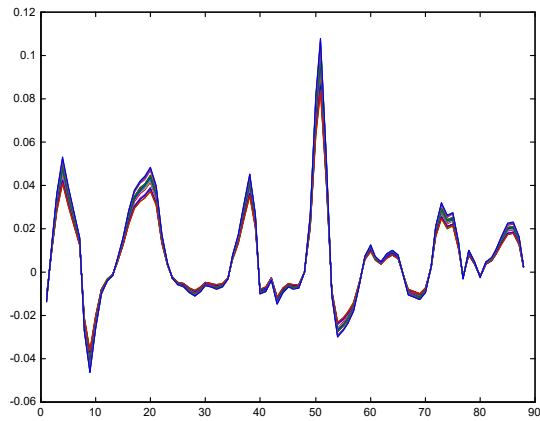


# Scores plot

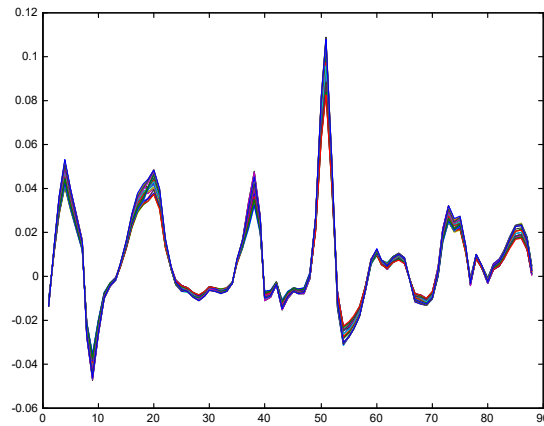


# Decomposition and residues

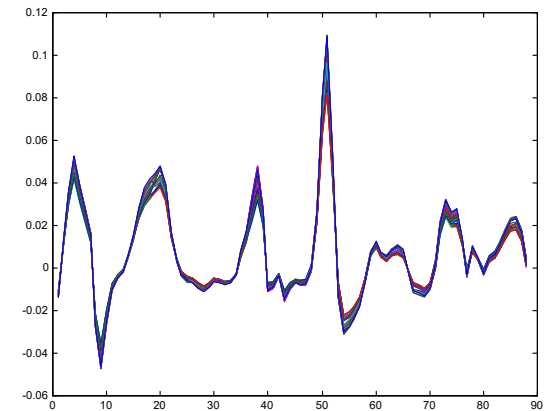
## First PC



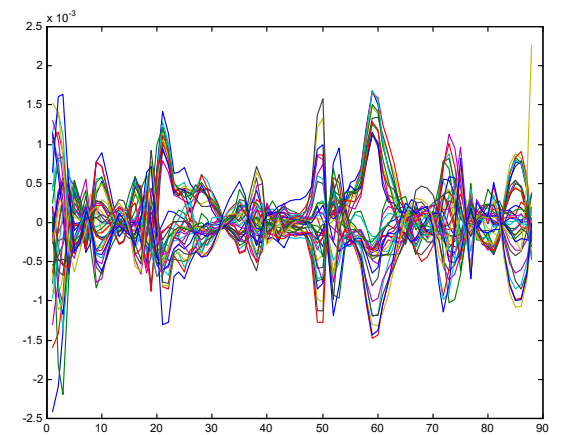
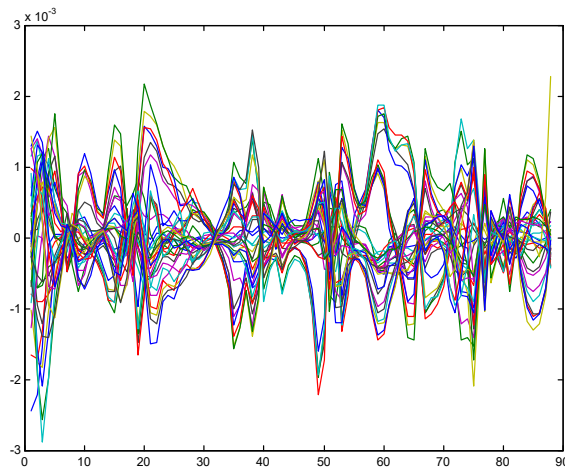
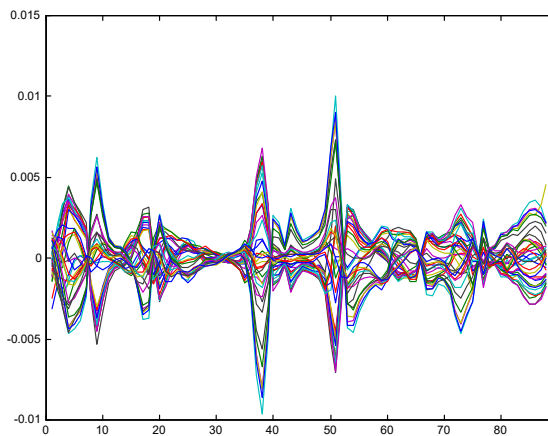
## Second PC



## Third PC



## Residues



# Principal Components Regression (PCR)

- We divide the dataset into two:
- 26 for the calculation of PCcal model,  $Y_{cal}$
- 10 for the error evaluation PCval,  $Y_{val}$
- The model calculates the regression matrix  $B_{pcr}$

$$Y_{cal} = X_{cal} \cdot B^T \Rightarrow B^T = P \cdot \Lambda^{-1} \cdot T^T \cdot Y_{cal}$$

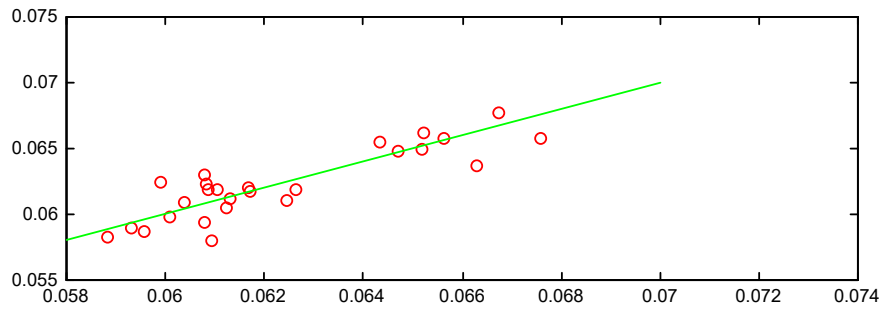
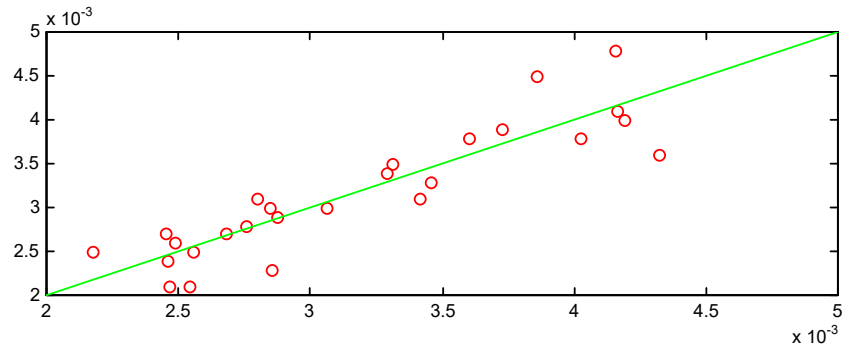
- We calculate an estimation of the validation set (and for comparison also of the calibration)
- RMSEC and RMSECV

$$stimaY_{cal} = X_{cal} \cdot B^T$$

$$stimaY_{val} = X_{val} \cdot B^T$$

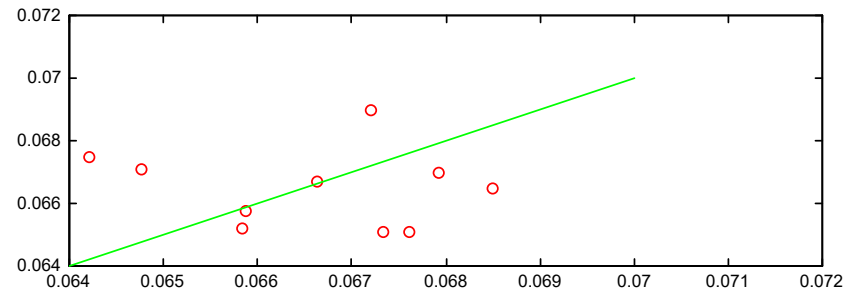
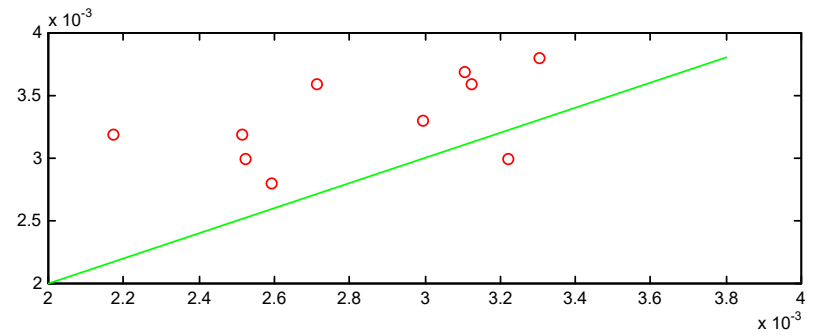
# results

## calibration



RMSEC<sub>acidity</sub> =  $3.1 \cdot 10^{-4}$   
RMSEC<sub>humidity</sub> = 0.0013

## test



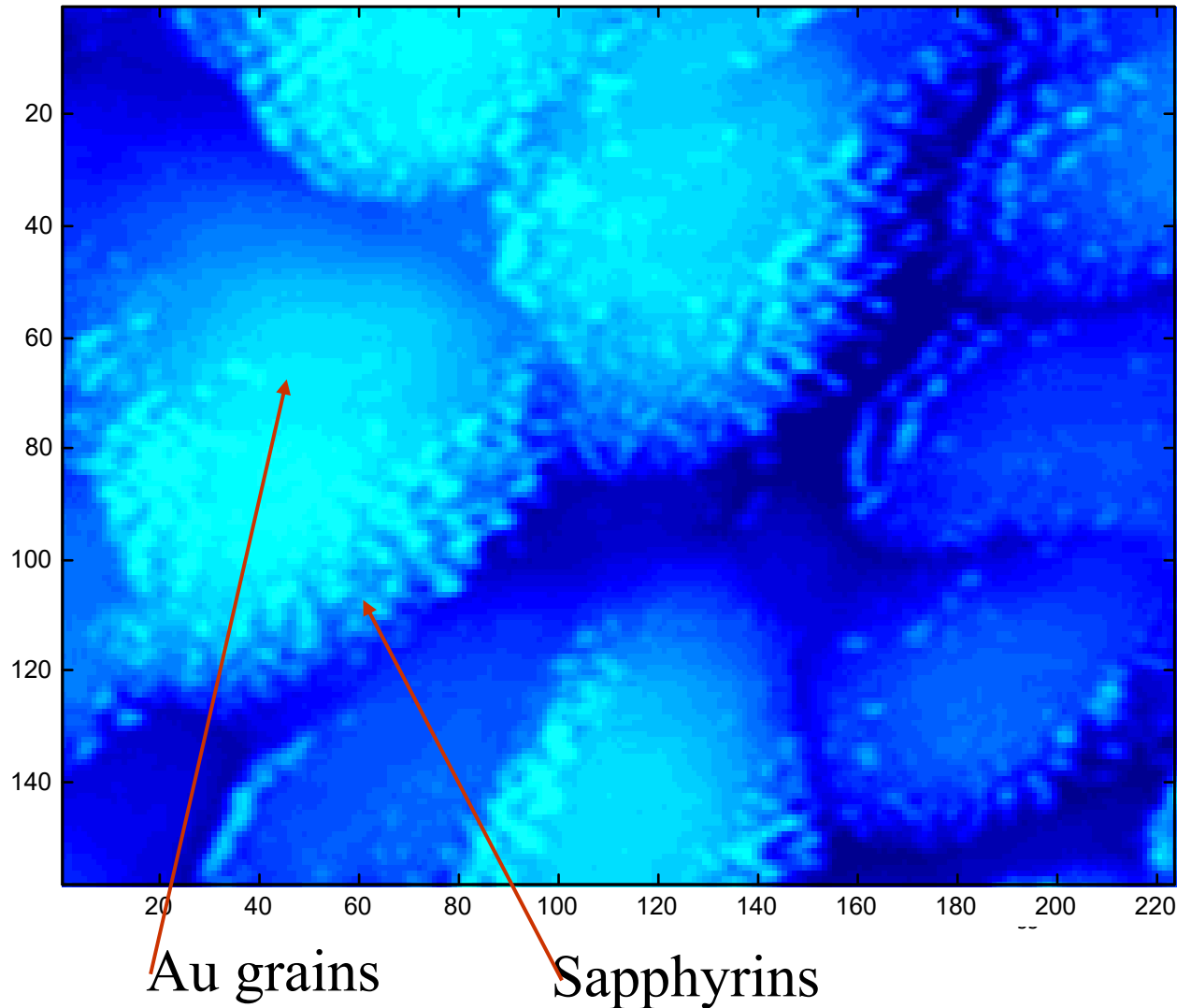
RMSECV<sub>acidity</sub> =  $5.9 \cdot 10^{-4}$   
RMSECV<sub>humidity</sub> = 0.0019

# Application to the analysis of the images

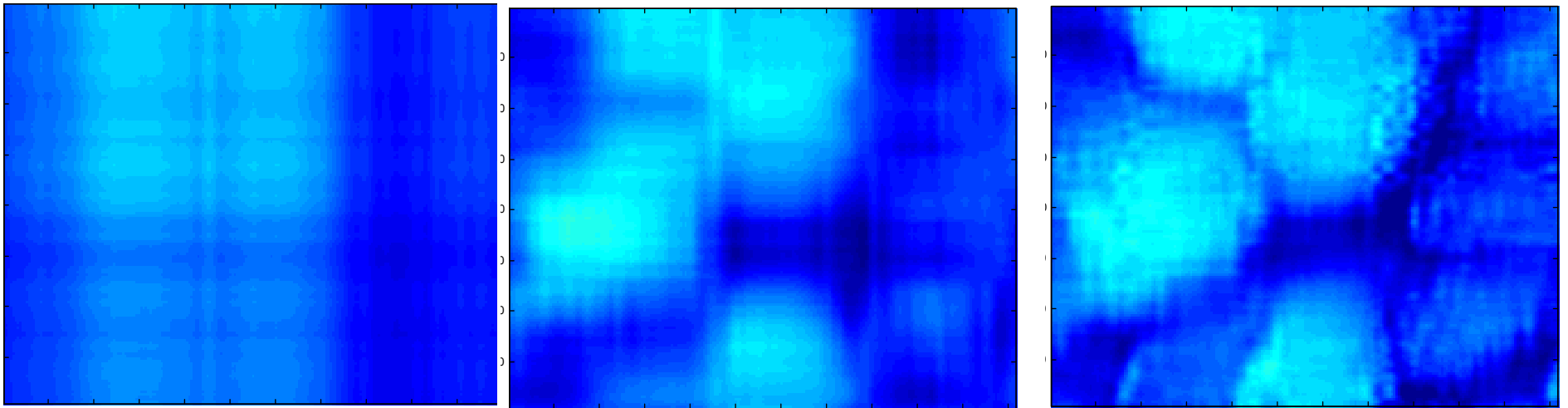
- A scanned image can be seen as an  $N \times M$  matrix in the case of gray scale (black to white scale image) or  $N \times M \times 3$  (in the case of color image)
- A picture can be considered as a matrix and we can apply the PCA
- The PCA decomposition may bring out some peculiar structures of the image allowing to study the characteristics of the image.

# PCA: Application to Image Analysis (example 1: I)

- STM image of Sapphyrin molecules growth as a Langmuir-Blodgett film onto a gold substrate.



# PCA: Application to Image Analysis (example 1: II)



$$X = S_1^T \cdot L_1$$

$$X = S_{1:10}^T \cdot L_{1:10}$$

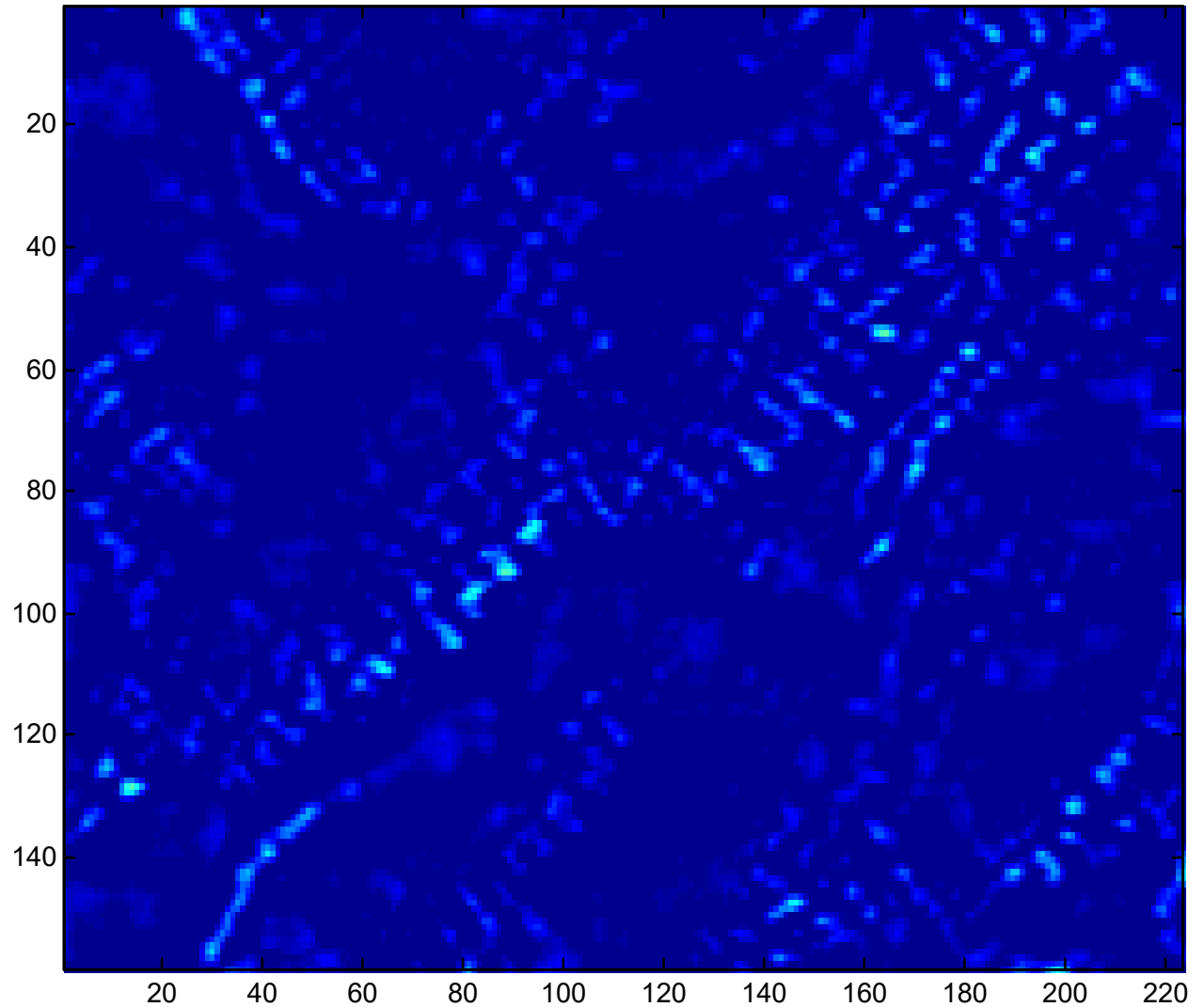
$$X = S_{1:15}^T \cdot L_{1:15}$$



# PCA: Application to Image Analysis (example 1: III)

- The residuals of the expansion at the tenth PC put in evidence the sapphyrine film only.

$$X - S_{1:10}^T \cdot L_{1:10}$$

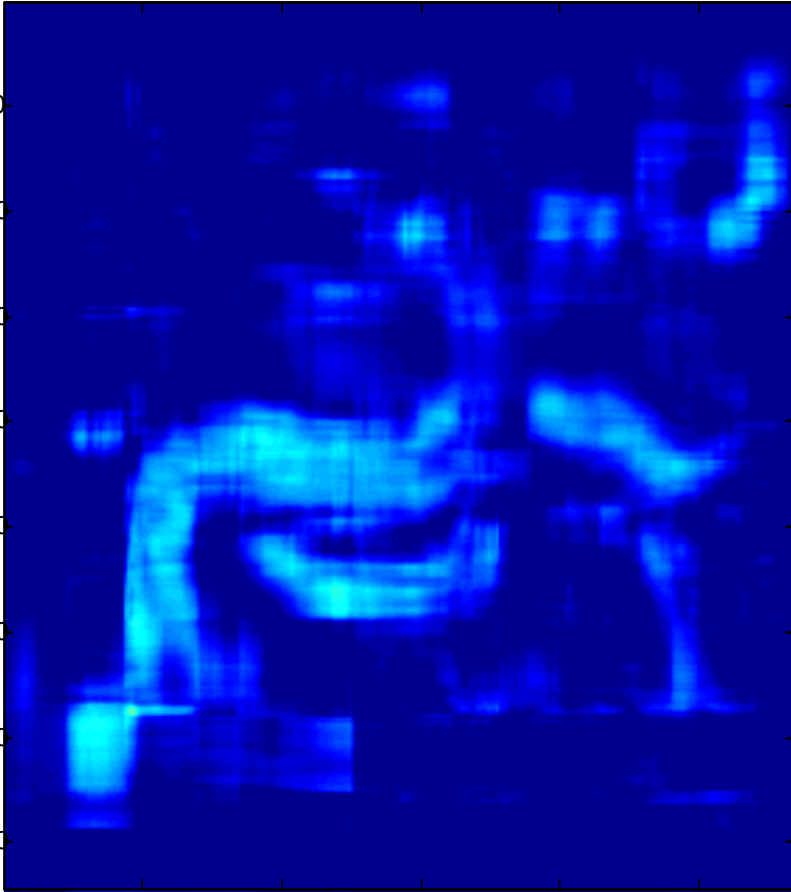


# PCA: Application to Image Analysis (example 2: I)

- Caravaggio Deposition



# PCA: Application to Image Analysis (example 2: II)



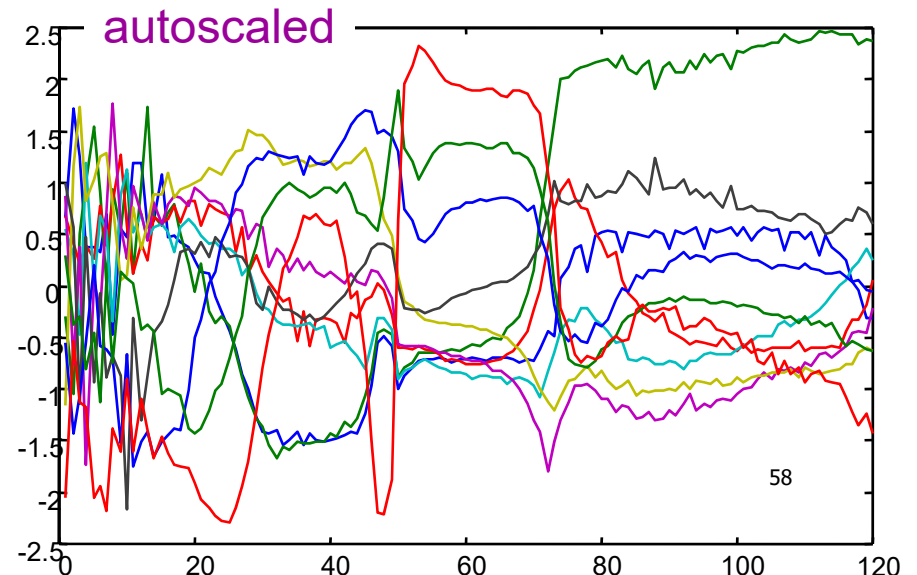
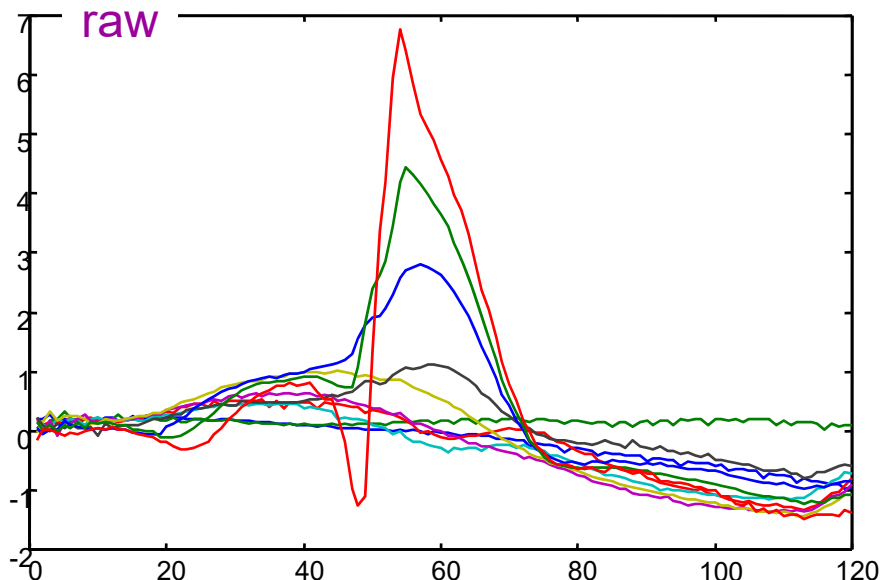
$$X = S_{1:10}^T \cdot L_{1:10}$$



$$X - S_{1:10}^T \cdot L_{1:10}$$

# The normalization problem

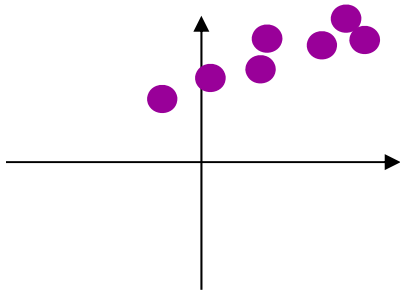
- The normalization is an operation that reduces the matrix columns (features) to zero average (zero average and variance equal to one).
- The autoscaling gives the same weight to every feature, this procedure is good if we are sure that every feature has the same importance in the problem.
- The autoscaling becomes dangerous when one or more features are noisy or when the numerical relationships between features are important
- Typical case is the spectroscopy where autoscaling completely destroys the information



# Normalization and Pattern Recognition

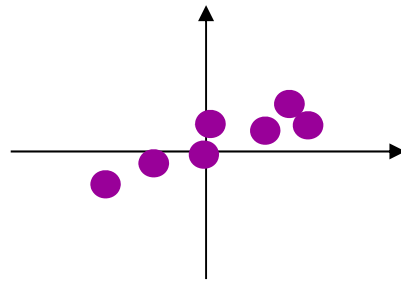
*raw*

**$X$**



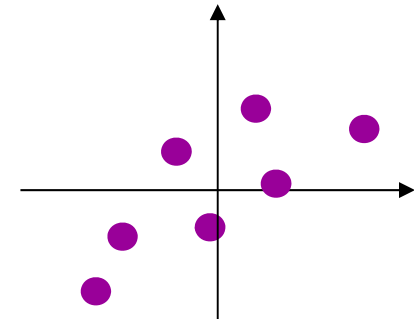
*centered*

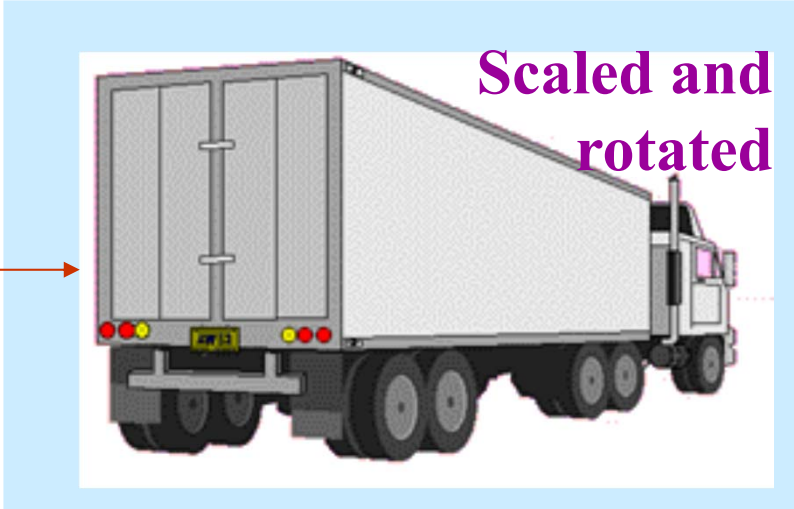
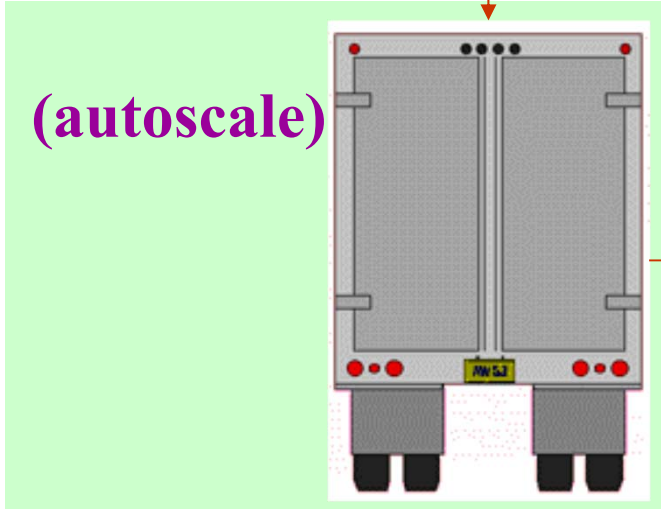
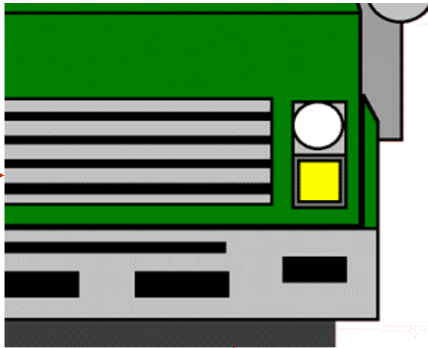
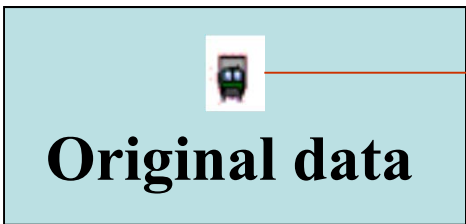
**$G = X - \mu$**



*autoscaled*

**$Z = \frac{X - \mu}{\sigma}$**





# PCA and pattern recognition

- The principal component analysis is a method that allow:
- To define features of a new set (linear combination of the original) that are uncorrelated between them
- To decompose the variance of the data in the sum of the variance of the new axes (principal components)
- To reduce the representation of the pattern to a subspace identified by the main components of greatest variance
- To study the contribution of the original features to the core components by identifying the most significant higher contribution features.

# Example: Fruits parameters

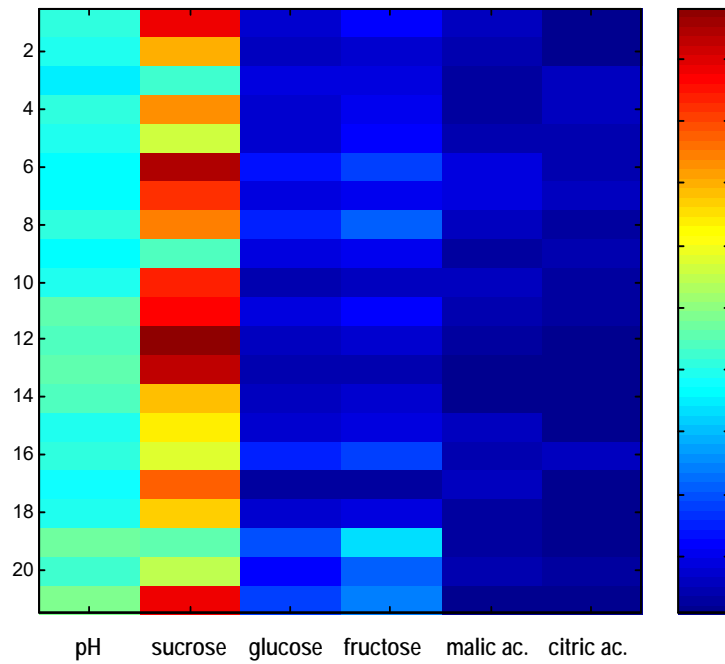
- Suppose we have measured the following quantities in peaches: pH, sucrose, glucose, fructose, malic acid and citric acid, and we want to study the classification and the relationship using these parameters.

	<i>pH</i>	<i>sucrose</i>	<i>glucose</i>	<i>fructose</i>	<i>malic acid</i>	<i>citric acid</i>
<i>baby gold</i>	4.10	8.80	0.80	1.20	0.60	0.20
<i>grezzano</i>	4.0	7.0	0.60	0.80	0.50	0.10
<i>iris rosso</i>	3.50	4.30	0.90	1.0	0.40	0.60
<i>maria aurelia</i>	4.10	7.30	0.80	1.10	0.40	0.60
<i>snow queen</i>	3.90	5.70	0.80	1.30	0.50	0.50
<i>spring star</i>	3.60	9.40	1.40	1.90	1.0	0.50
<i>super crimson</i>	3.70	8.20	1.0	1.10	0.90	0.60
<i>venus</i>	4.10	7.40	1.60	2.20	0.70	0.40
<i>argento roma</i>	3.60	4.40	0.90	1.10	0.40	0.50
<i>beauty lady</i>	3.90	8.30	0.50	0.70	0.60	0.30
<i>big top</i>	4.50	8.60	0.90	1.30	0.50	0.40
<i>doucer</i>	4.40	9.80	0.70	0.80	0.40	0.10
<i>felicia</i>	4.60	9.30	0.50	0.50	0.20	0.20
<i>kurakata</i>	4.40	6.90	0.60	0.80	0.20	0.20
<i>lucie</i>	3.90	6.40	0.80	1.0	0.70	0.20
<i>morsinai</i>	4.10	5.80	1.60	1.90	0.50	0.60
<i>oro</i>	3.80	7.70	0.40	0.40	0.60	0.20
<i>royal glory</i>	4.0	6.70	0.80	0.90	0.40	0.10
<i>sensation</i>	4.70	4.60	2.0	3.40	0.30	0.20
<i>sweet lady</i>	4.20	5.50	1.30	2.10	0.50	0.40
<i>youyeong</i>	4.90	8.80	1.80	2.50	0.20	0.10

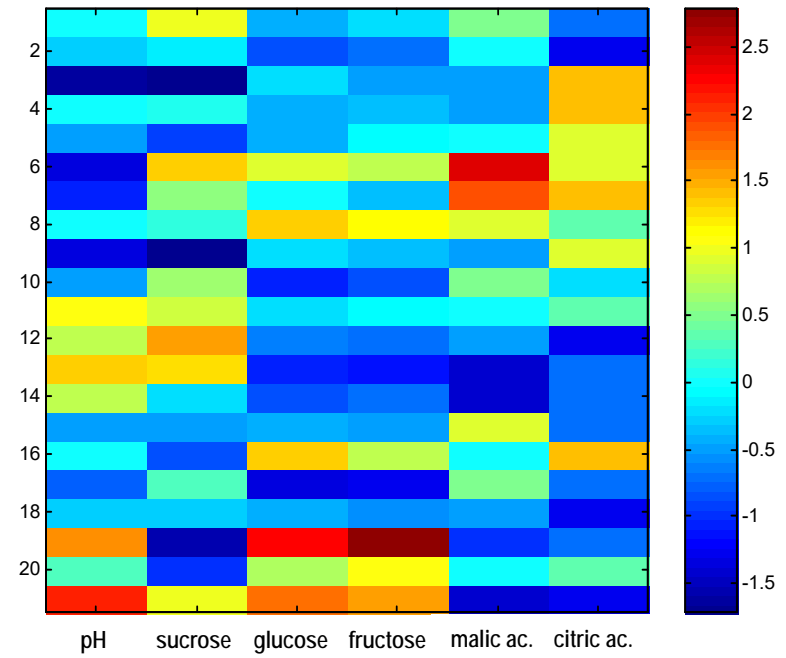


# Color map

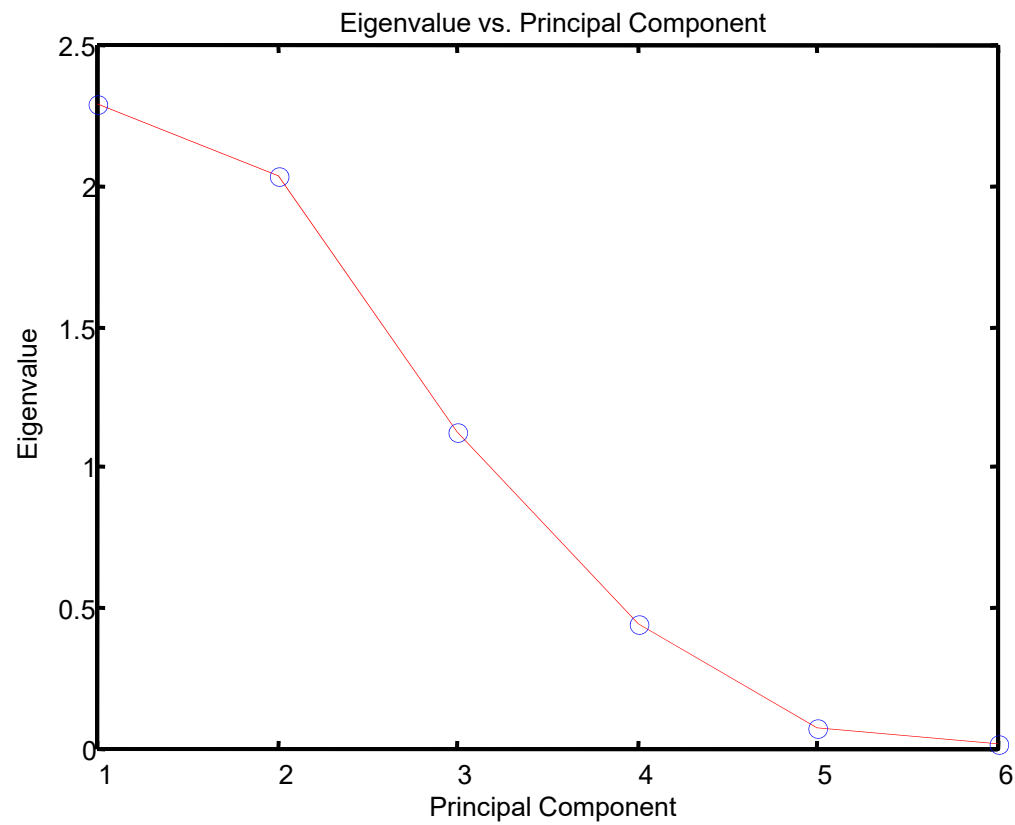
*raw data*



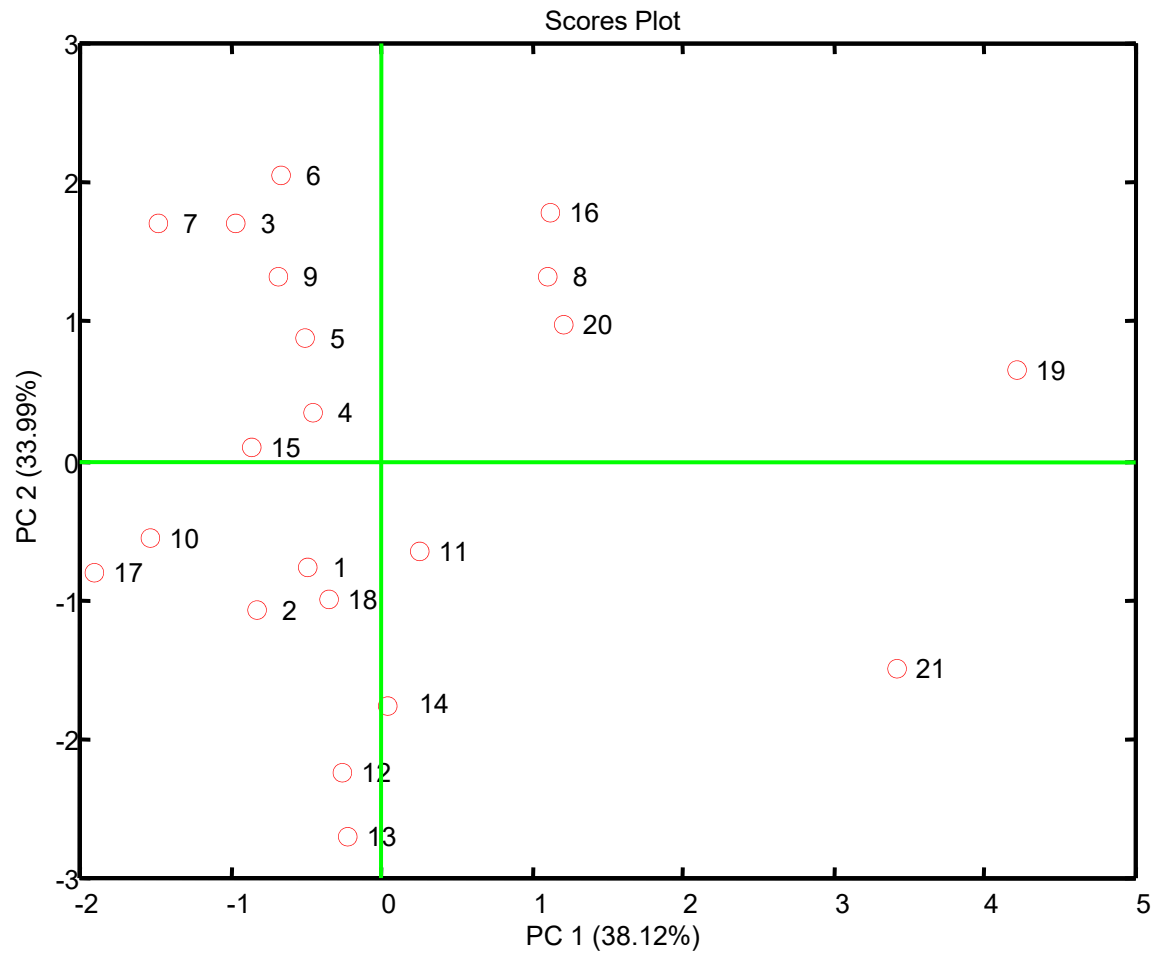
*Autoscaled data*



# PCA: peaches data eigenvalues vs. PC



# PCA: scores plot

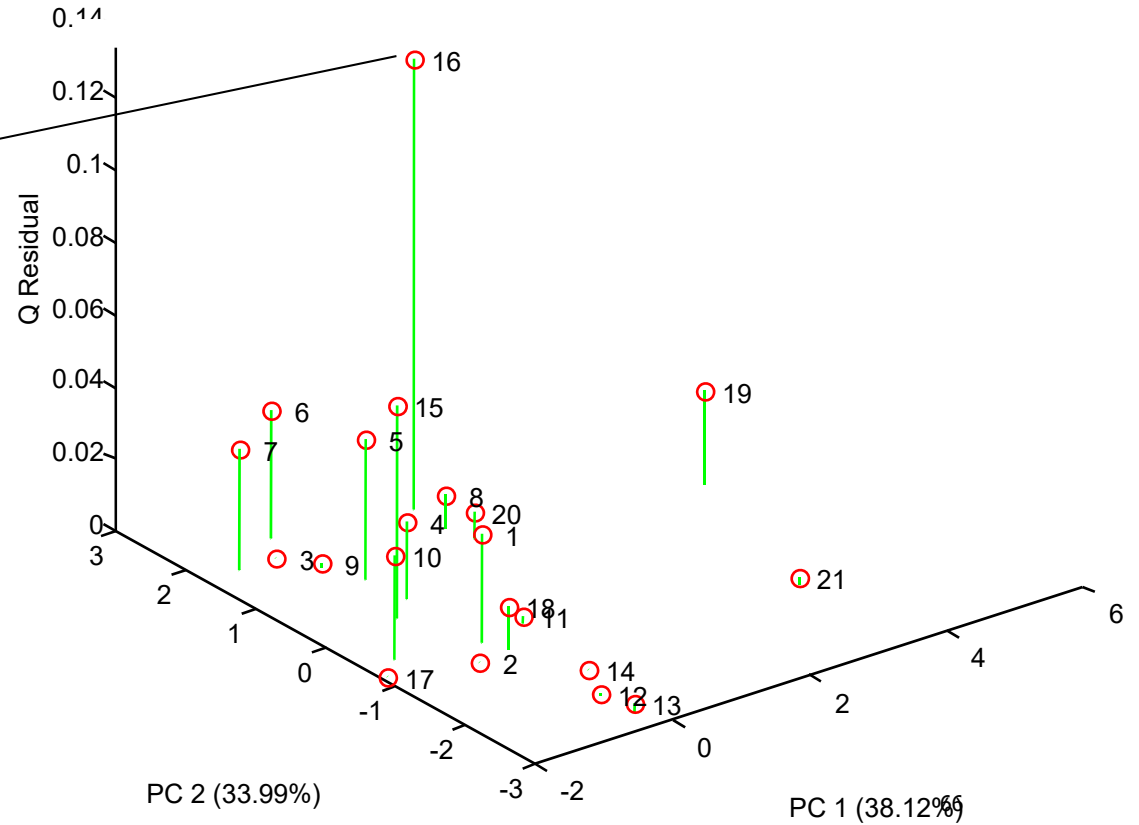
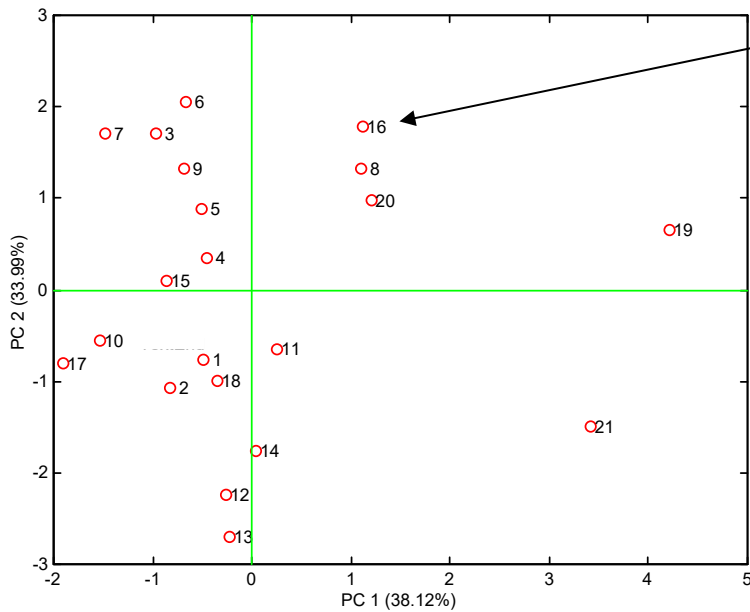


# Residuals representation

$$x_i = a \cdot s_1 + b \cdot s_2 + \dots + n \cdot s_n$$

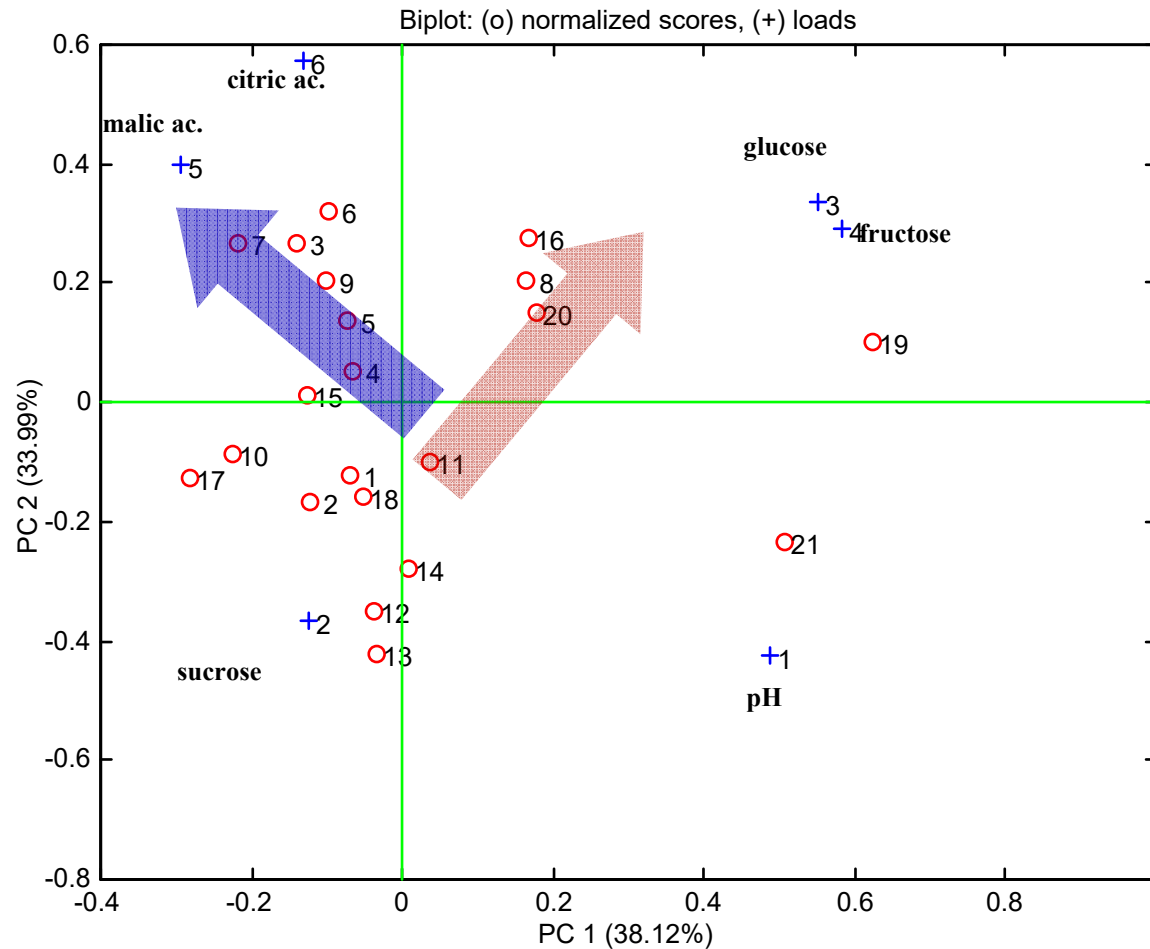
$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$

Scores Plot



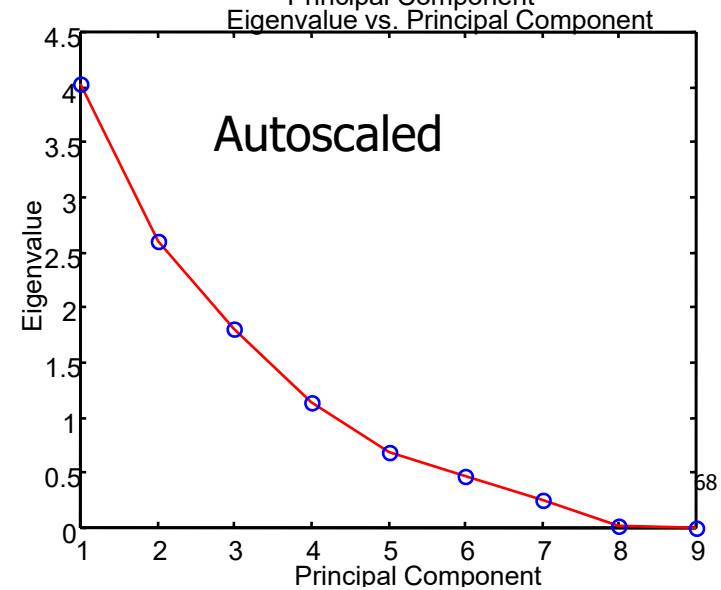
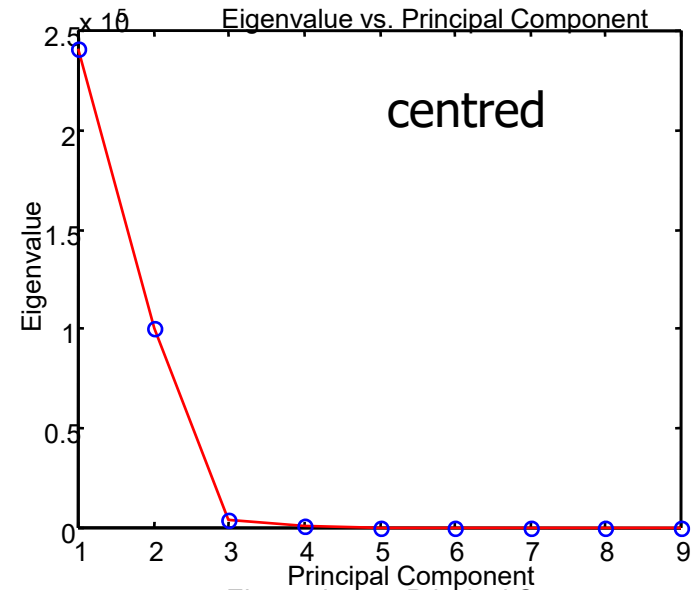
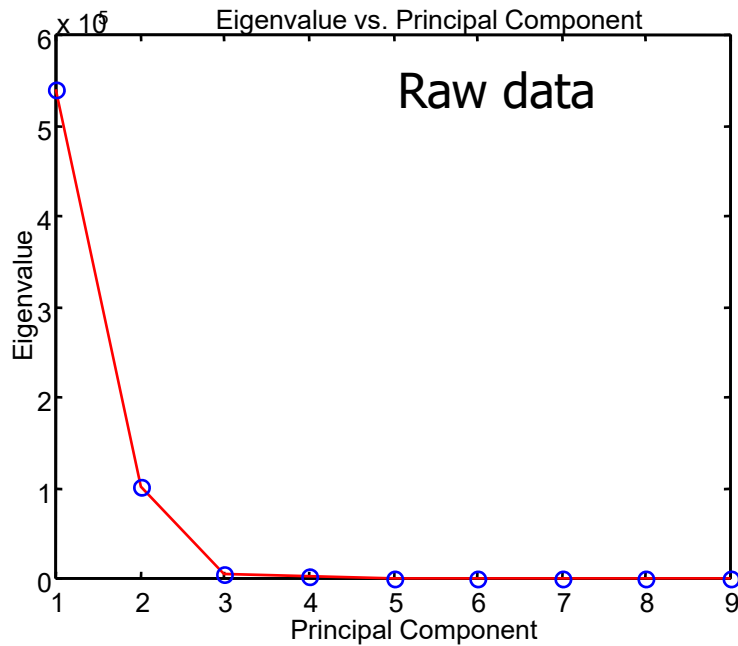
# PCA example: peaches data

## bi-plot: scores+loadings



- The sugars are orthogonal to acids
- We identify the direction of the acidity and the sweetness
- The sucrose is anticorrelated to glucose and fructose
- The pH is obviously anticorrelated to acids

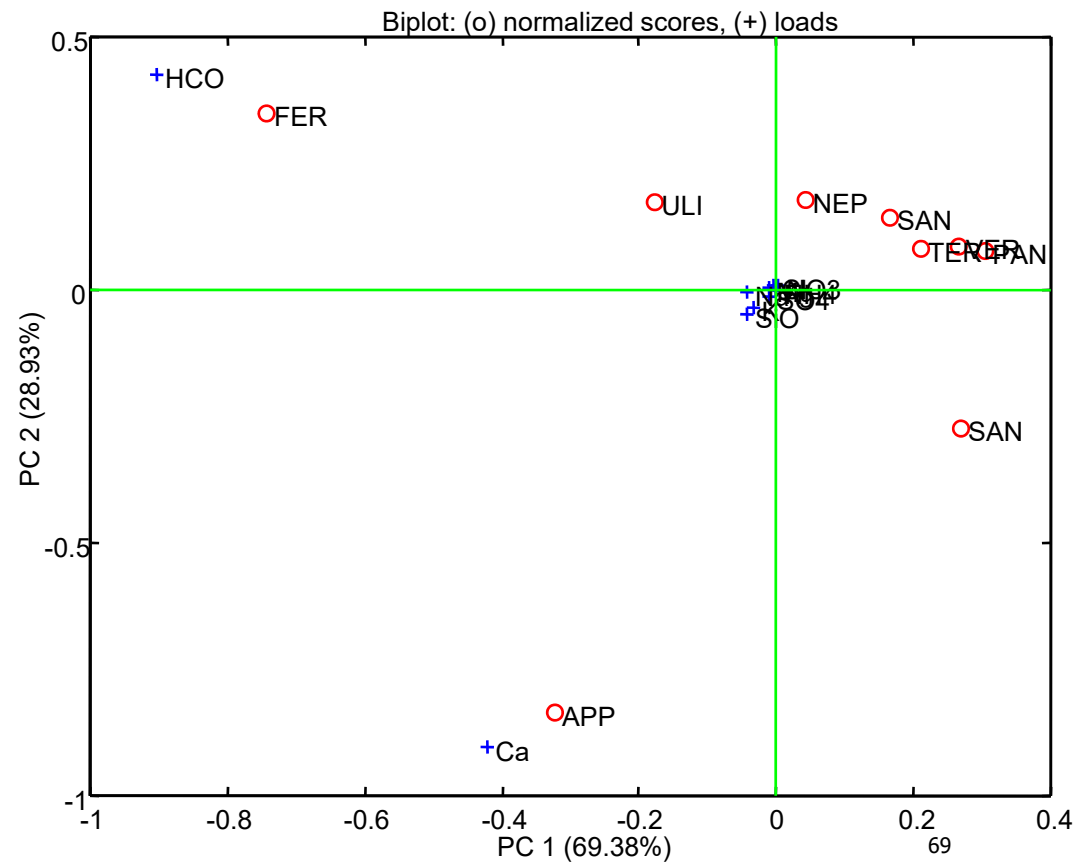
# PCA example: mineral waters



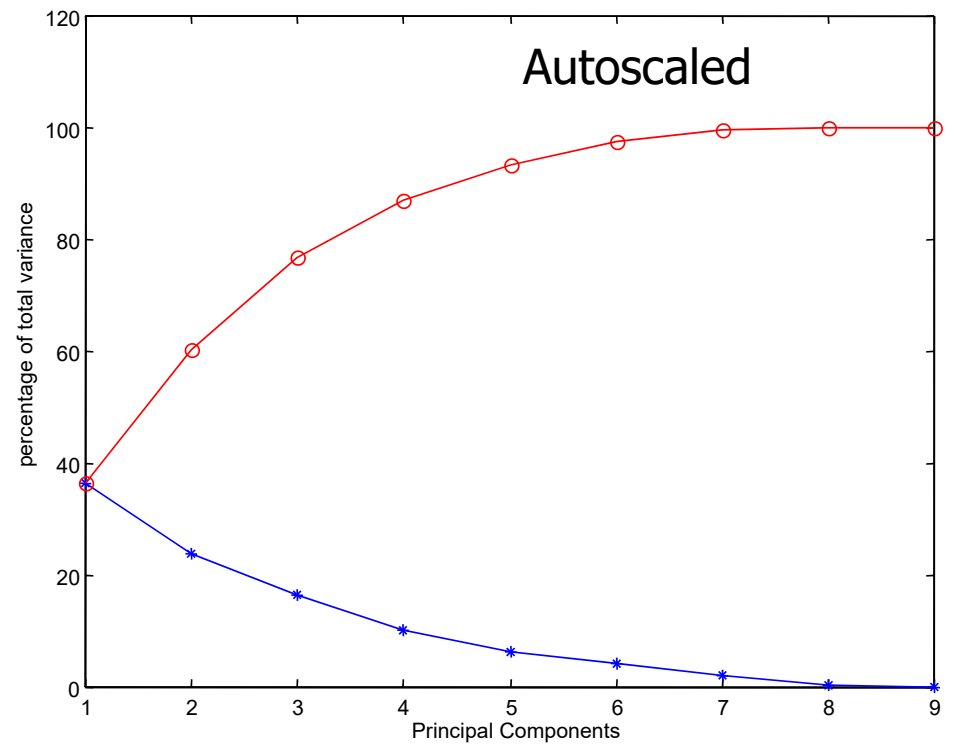
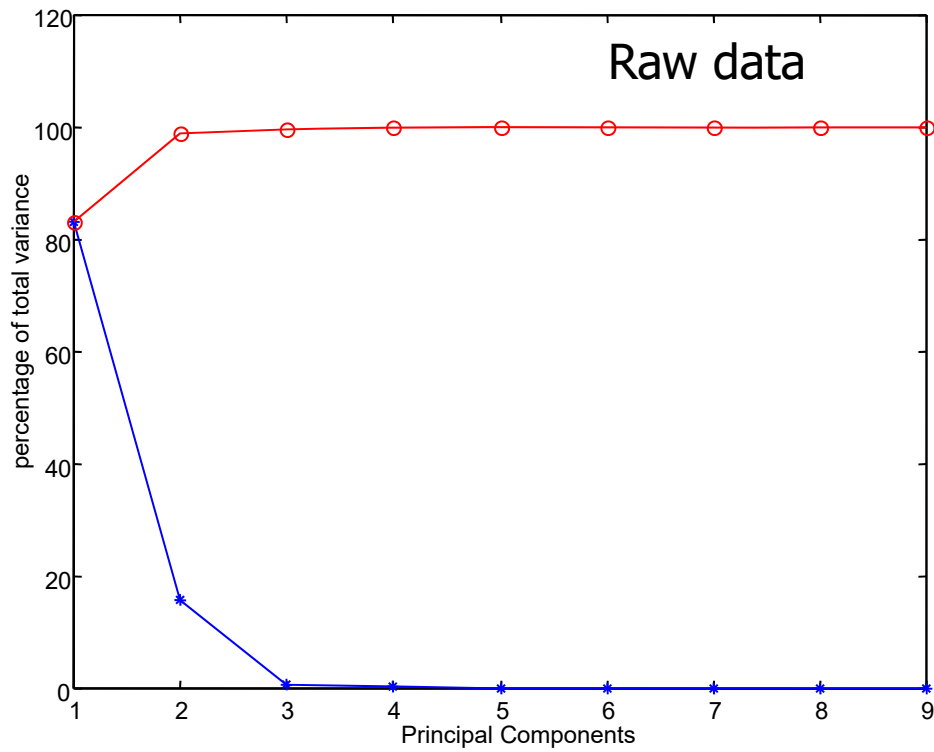
The autoscaling makes homogeneous the features by increasing the number of important dimensions

# Mineral waters: PCA biplot raw data

- Only features numerically significant are important (HCO and Ca)
- The other features are around the origin and don't contribute to the classification
- HCO and Ca are orthogonal
- Orthogonal means uncorrelated
- Only RES and APP are different from others
- in this plot has 98% of the variance



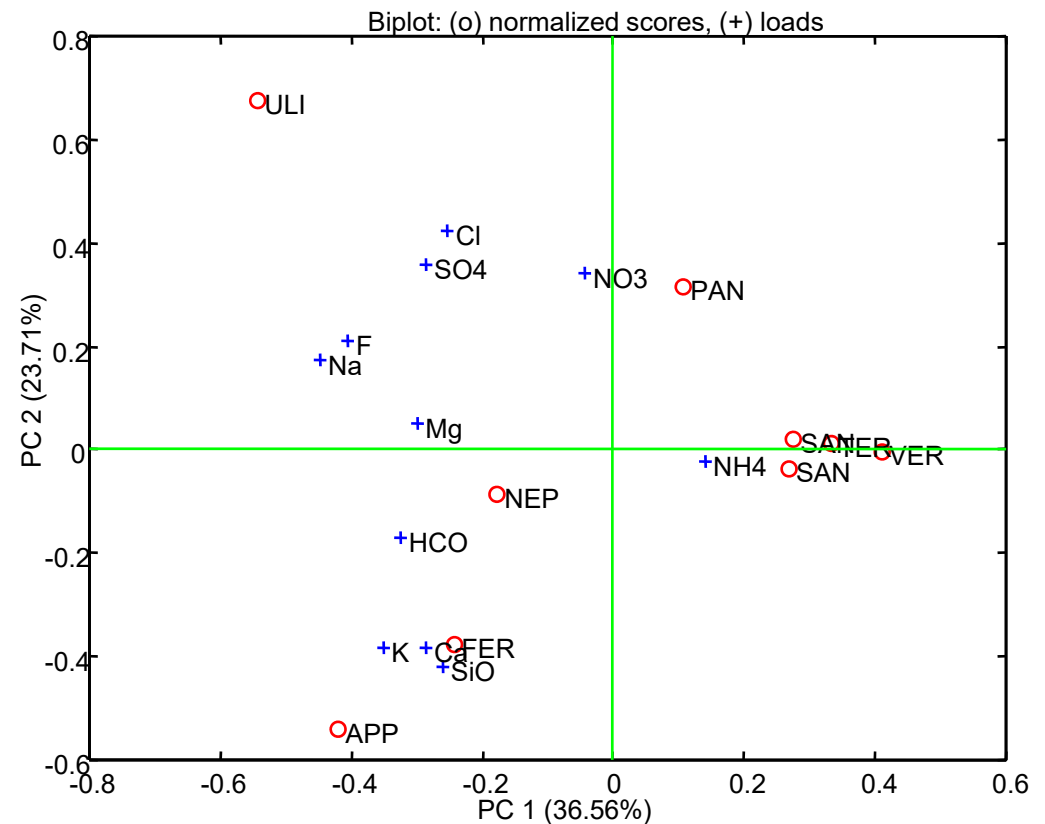
# Scree plot variance



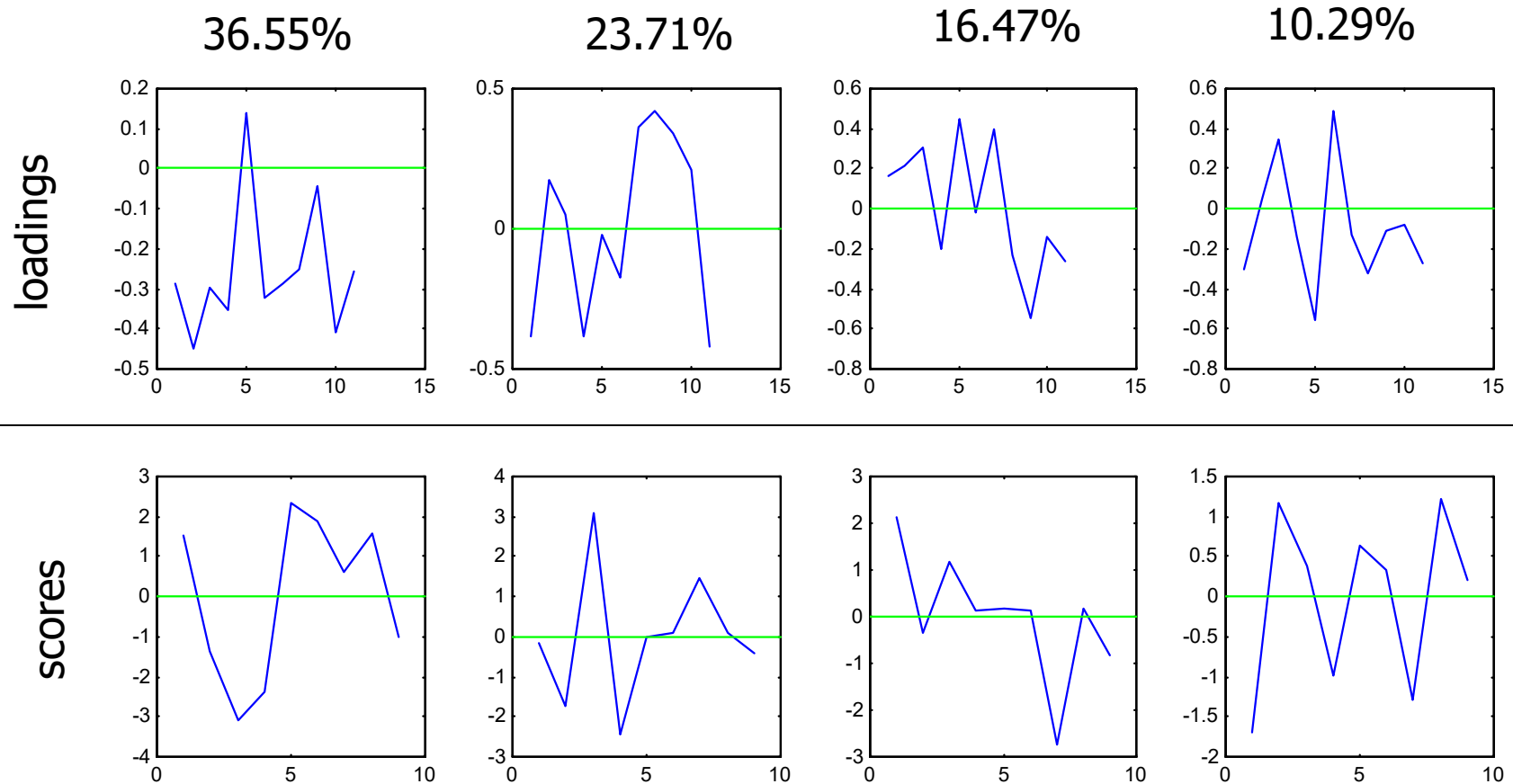


# Mineral waters: Autoscaled PCA biplot

- The features contribute homogeneously
- The following are groups of waters:
  - SAN, TER, VER, SAB
  - minerals oligo
  - PAN
  - Oligo but with increase of NO<sub>3</sub>
  - NEP, FER, APP
  - Intensification of Mg, HCO, Ca, K
  - ULI
  - Increasing in Cl, SO<sub>4</sub>
  - For ULI, NEP, FER, APP
  - Common increasing of F, Na
  - 60% of the variance in this plot
  - And the other 40%?



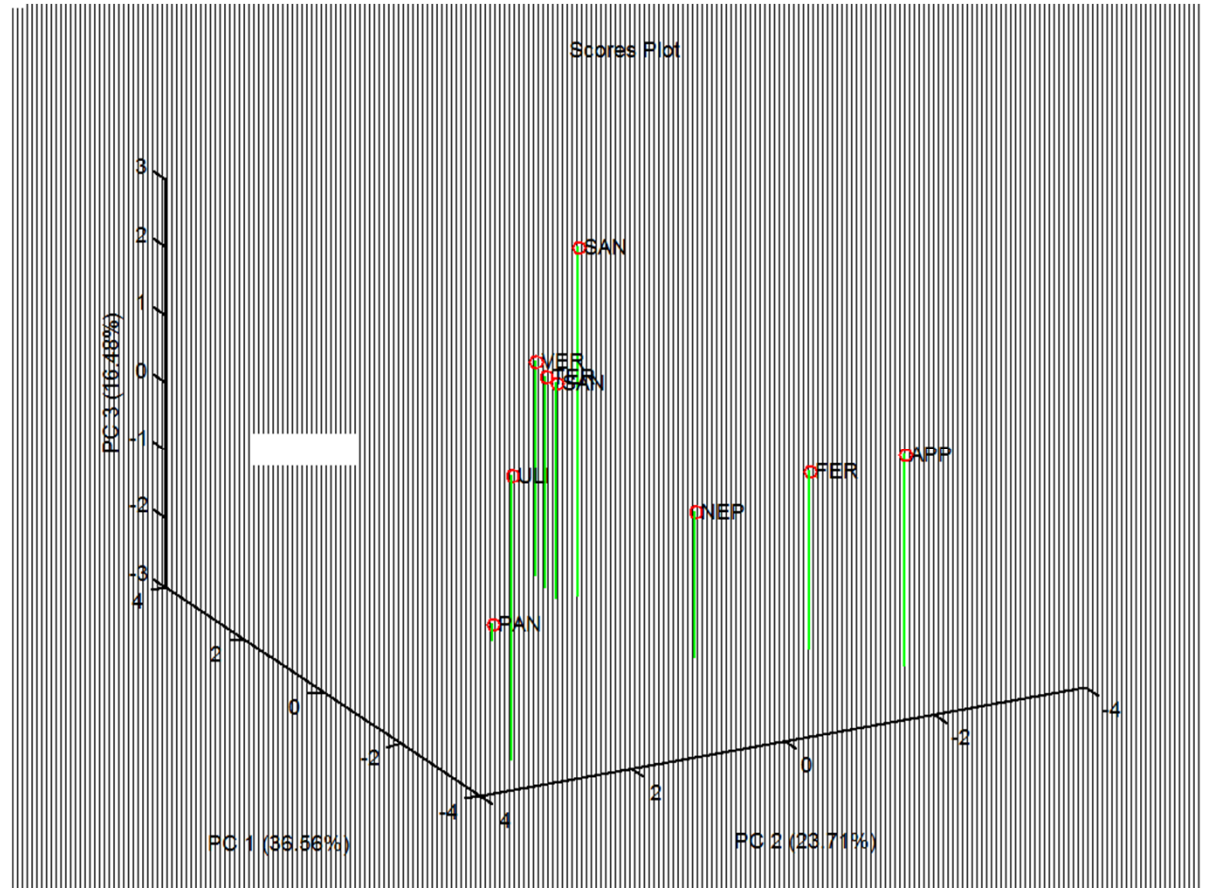
# Mineral waters: loadings and scores analysis



# PCA: mineral waters

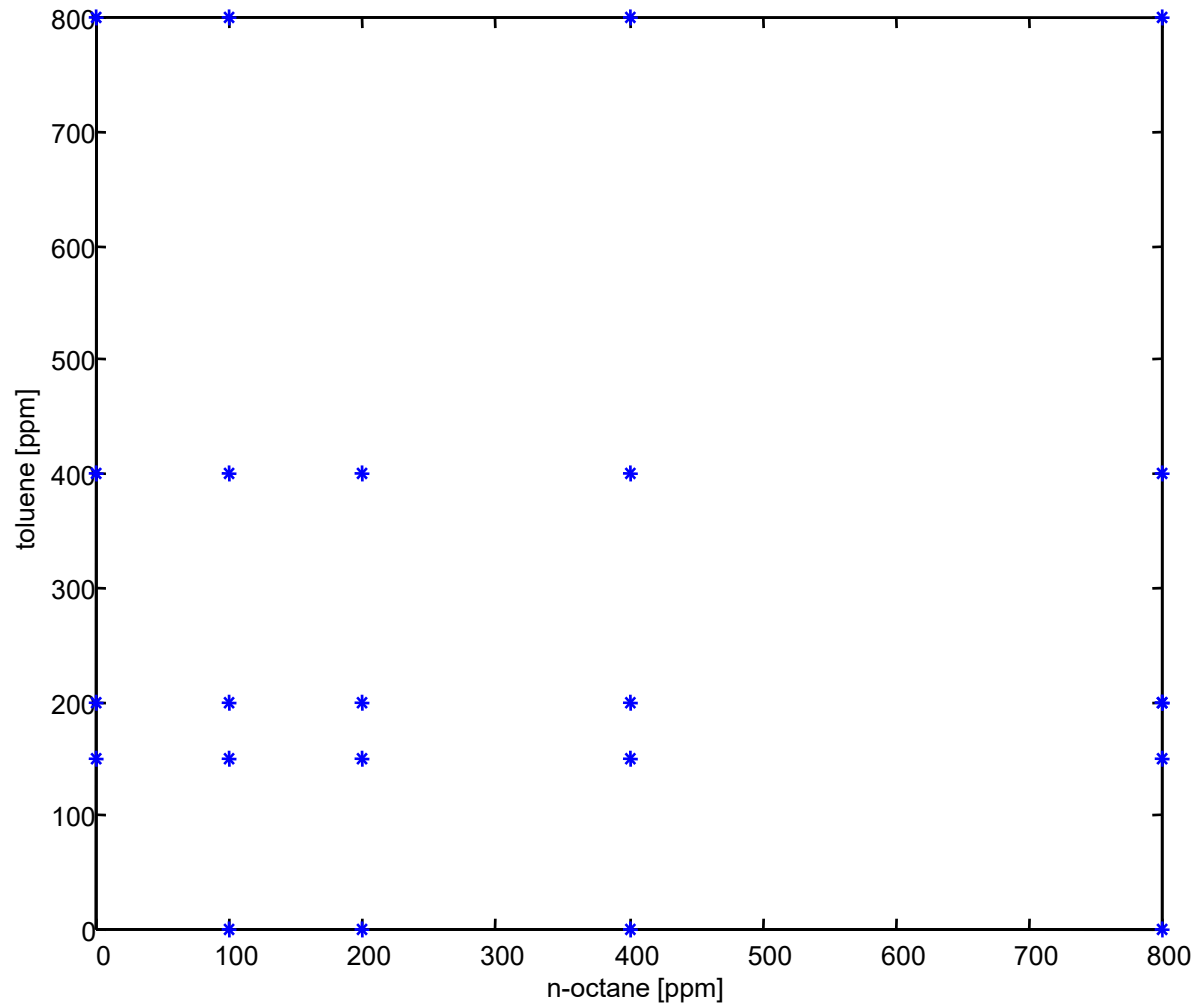
- The 2D representation is not sufficient because the distribution of eigenvalues. 2D representations capture only different aspects of the problem.

Score plot 3D  
76% di variance  
SAN is separated showing  
specific characteristics

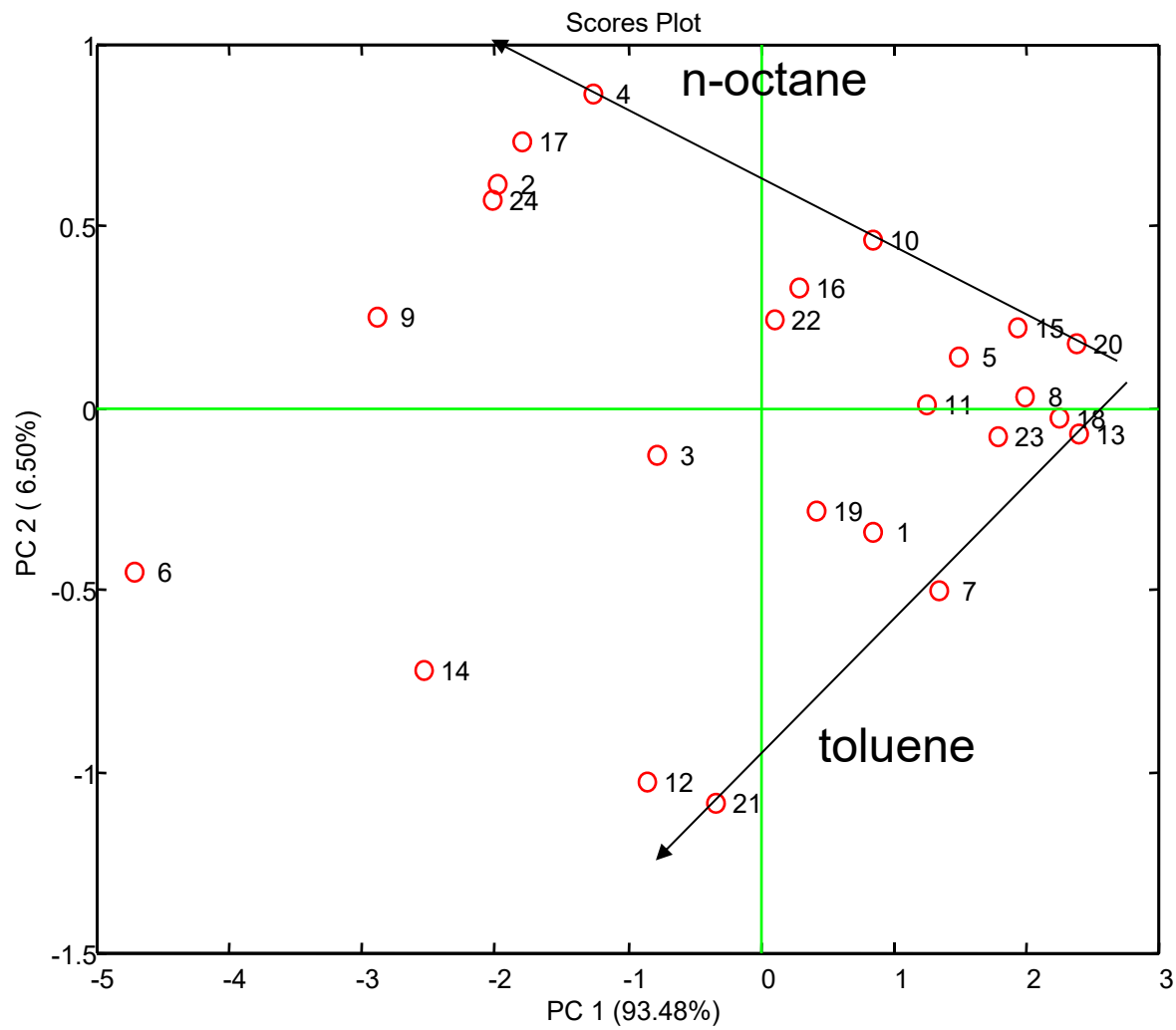


# 4 sensors for 2 gas

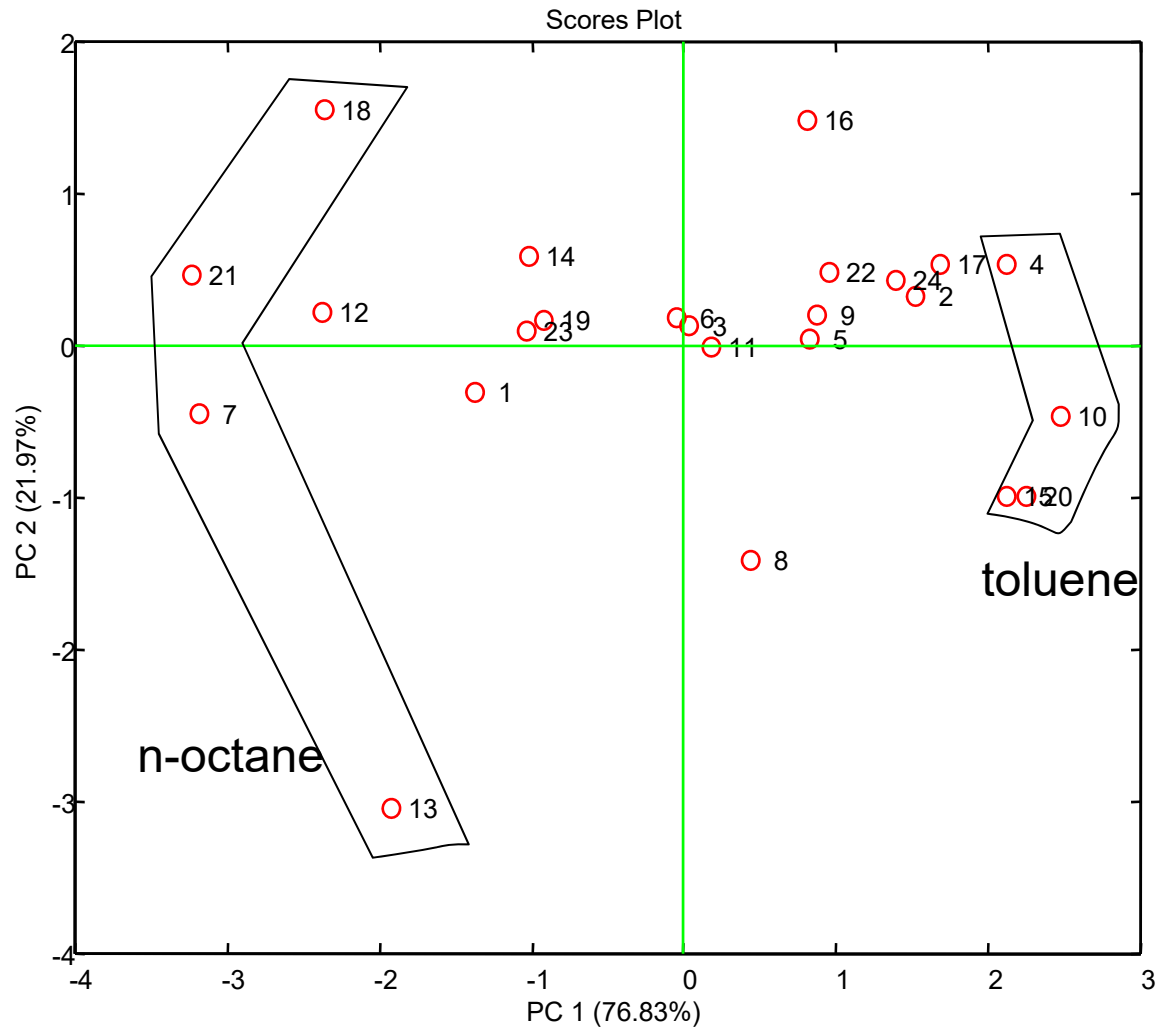
Measurements Plot



# PCA scores

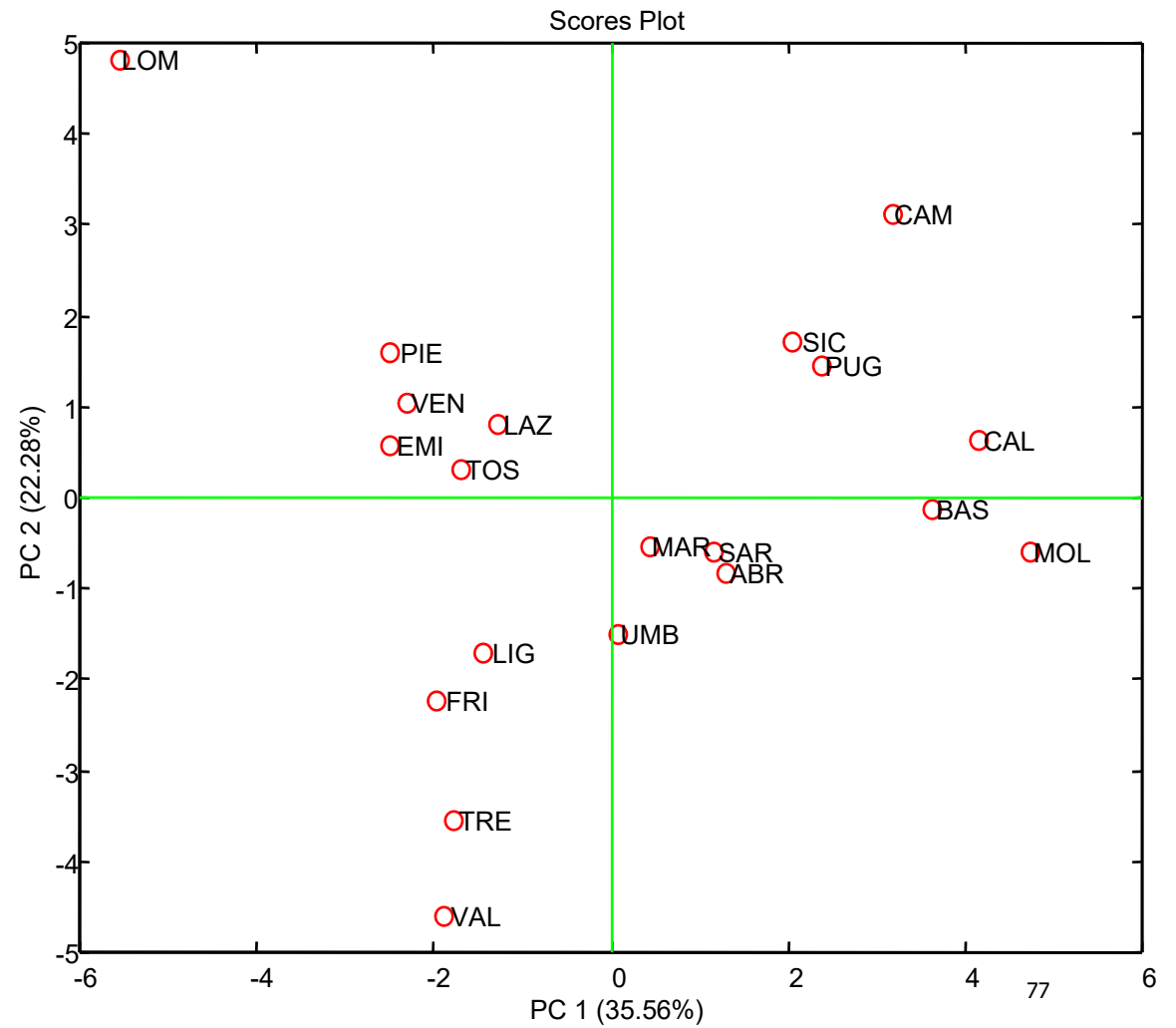


# PCA scores normalization

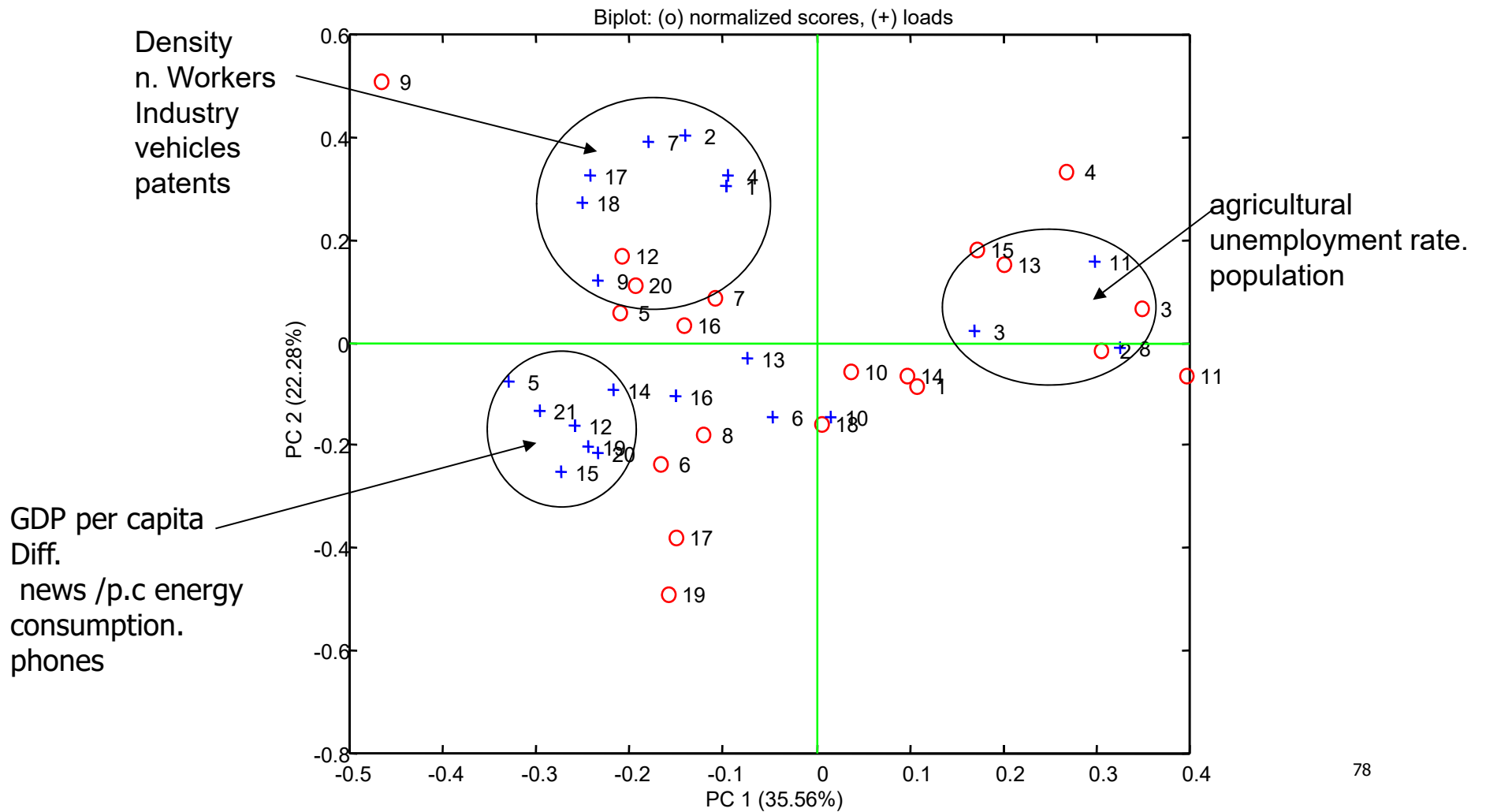


# Welfare of Italian regions

Based on 21 geographic and economics data

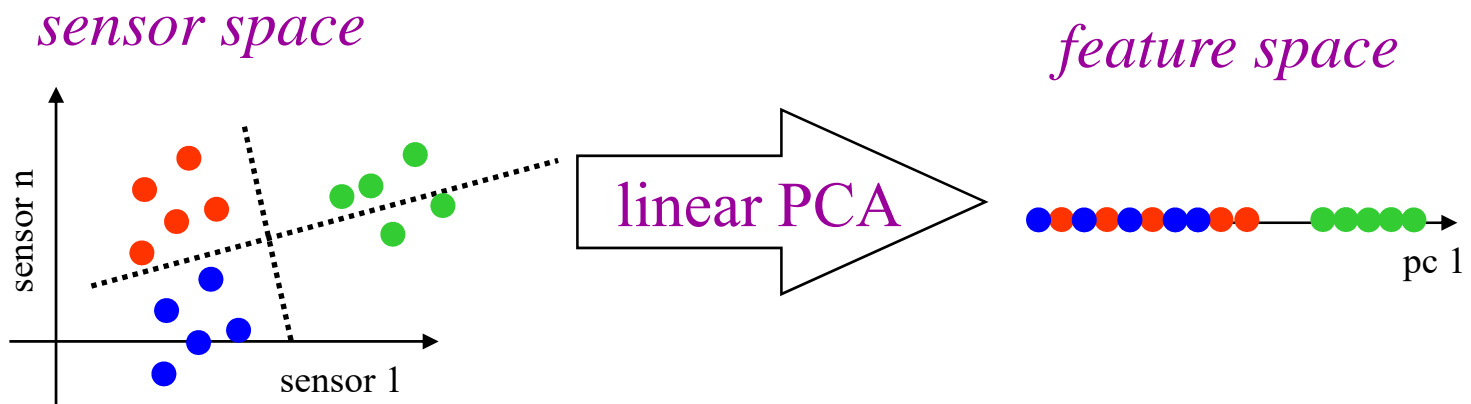


# PCA bi-plot





# PCA



# PCA limitations

- The PCA representation is driven by the data characteristics in covariance matrix
- If the data are not normally distributed the covariance matrix does not satisfy the statistics of data, so the PCA representation is formally incorrect
- The score plot of the PCA is a linear projection from one to N dimension space to one dimension in the space 2 or 3. We can have false projection effects involving classification errors

# Partial Least Squares (PLS)

Partial Least Squares  
PLS toolbox di MATLAB

# From PCR to PLS geometric approach

- The PCR solution is through the decomposition of the data matrix in the matrix of the principal components
- The principal components are the directions, in the space of the variables  $X$ , maximizing the variance and generate a base in which the  $X$  data are not correlated
- PCR in the principal components has new variables (not correlated) so becomes more easily solved.
  
- In PLS algorithm also the  $Y$  matrix is decomposed into principal components and principal components of  $X$  are rotated in the direction of maximum correlation to the principal components of  $Y$
- PLS has latent variables, similar to the principal components maximizing the variance of both  $Y$  and  $X$

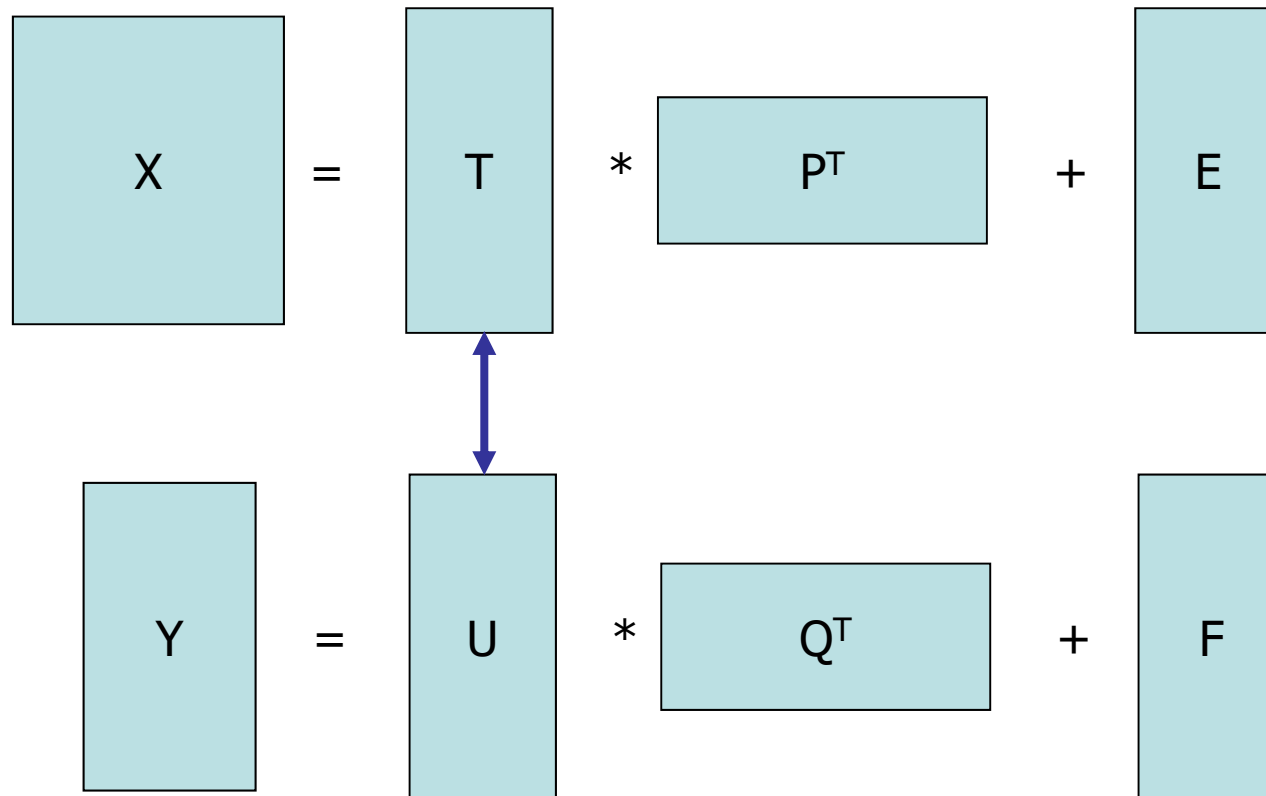
# PLS importance

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical.

PLS is used to find the fundamental relations between two matrices (X and Y), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. By contrast, standard regression will fail in these cases (unless it is regularized).

The PLS algorithm is employed in partial least squares path modeling, a method of modeling a "causal" network of latent variables (causes cannot be determined without experimental or quasi-experimental methods, but one typically bases a latent variable model on the prior theoretical assumption that latent variables cause manifestations in their measured indicators).

# PLS latent variables computation



The principal components are calculated maximising the correlation between  $T$  and  $U$  and their variance

$$\max [corr^2(U, T), var(U) \cdot var(T)]$$

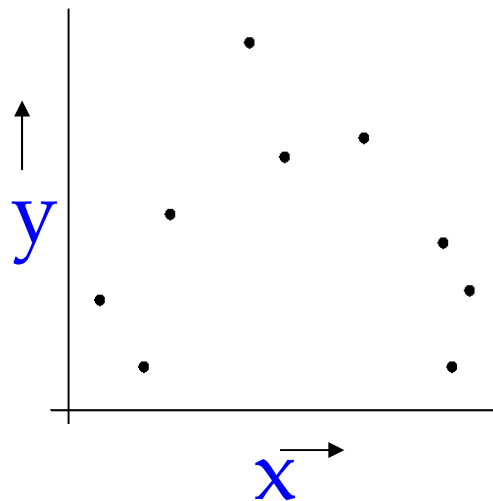
# PLS Overfitting

- PLS has overfitting
- The number of latent variables must be optimized in a cross-validation process
- The overfitting is given by the fact that the latent variable  $k$  is obtained by fitting the subspace of dimension  $k + 1$ . The latent variables are not **orthogonal** to each other, there is no limit to the possibility of fitting the data in calibration.
- The cross-validation sets the number latent variables accuracy estimated on the validation set. Normally this value is larger than the error obtained from the model on the calibration data
- Such errors are quantified by variables:
  - RMSEC    Root Mean Square Error in Calibration
  - RMSECV    Root Mean Square Error of Calibration in Validation

# Linear or no-linear model

## The validation problem

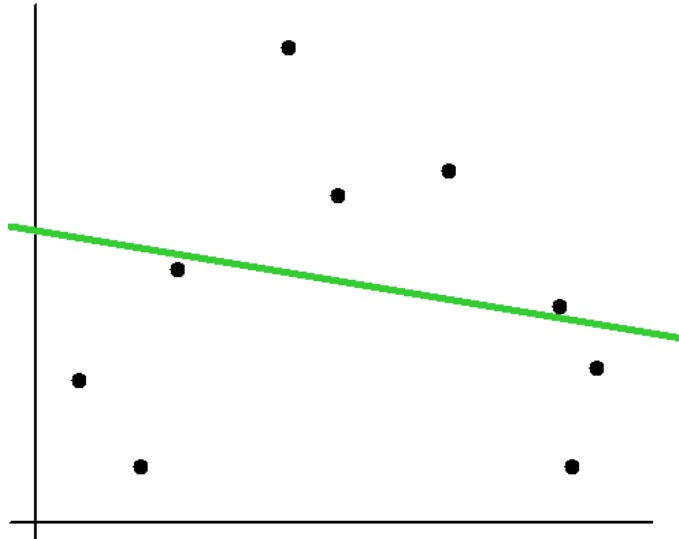
- Which is the best function that describes the experimental data?
- The One that allows you to predict with minimal error variables that have not been used to build the model.
- The operation that allows us to estimate this error is called cross-validation.
- Example: Consider the following information:  $y=f(x)+e$ 
  - What is the best function that describes the relationship between x and y?



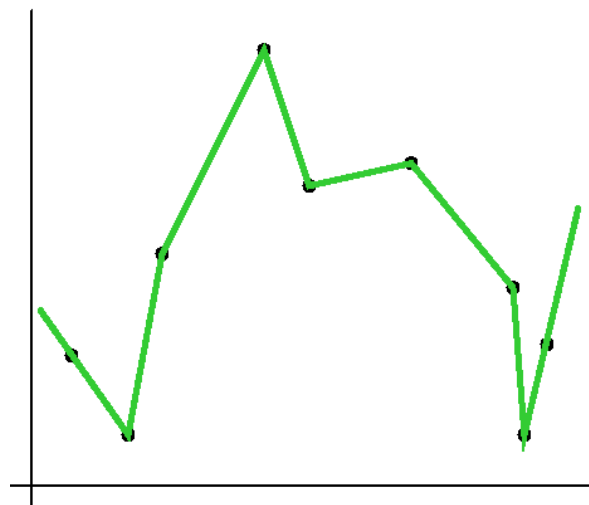
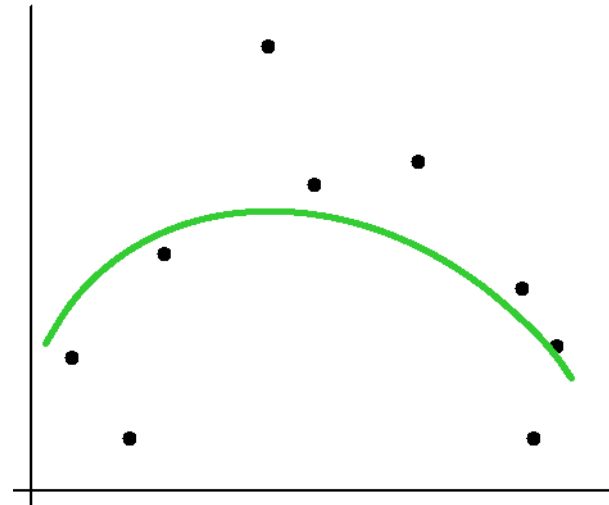


# solutions

linear



No-linear



No-linear

# test method

- The data set is divided into two
- The model is determined on a subset of the data (calibration with training set)
- The error is evaluated on the subset (test set)
- The prediction of the test set gives significance to the model. The data were not used for calibration. So the model can be used in the real world to estimate unknown data.

# Regression predictors

- PRESS- Predicted Sum of Squares

$$PRESS = \sum_i (y_i^{LS} - y_i)^2$$

- RMSEC - Root Mean Square error of calibration

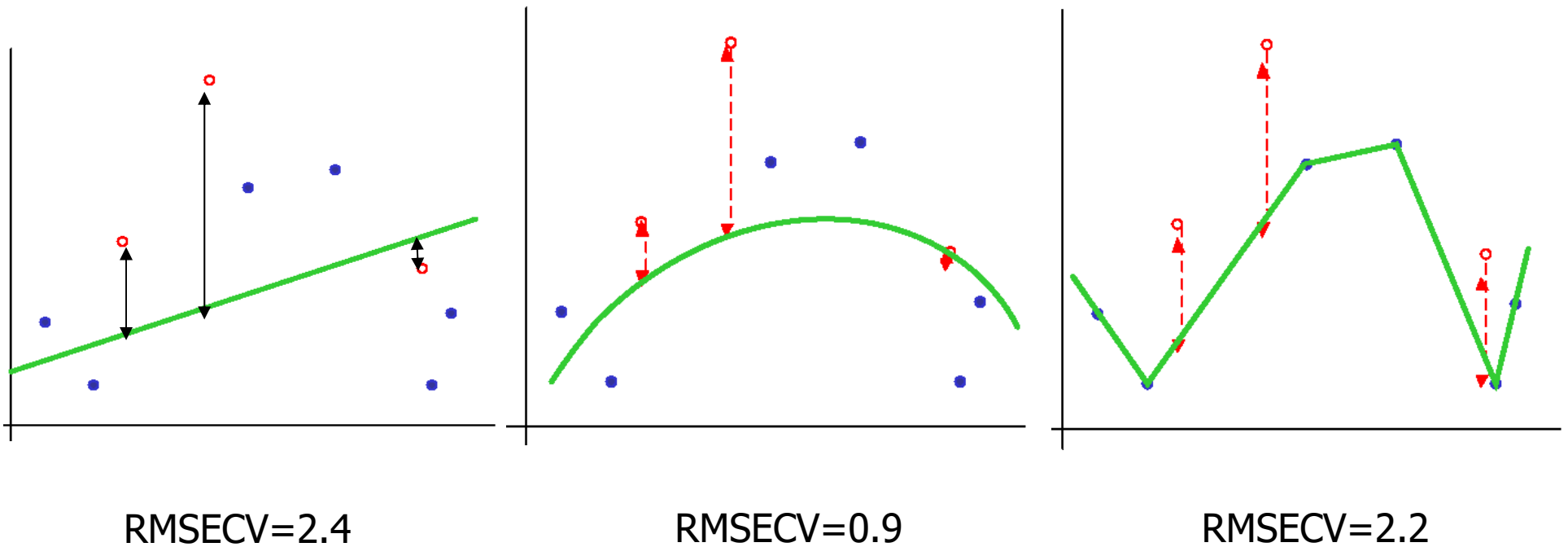
$$RMSEC = \sqrt{\frac{PRESS}{N}}$$

- RMSECV - Root Mean Square error of Cross-Validation

$$RMSECV_k = \sqrt{\frac{PRESS_k}{N}}$$

# Test methods application

The data marked in red are the test set. The model is calculated on the remaining data (blue dots).  
The error on the test data is evaluated as RMSECV

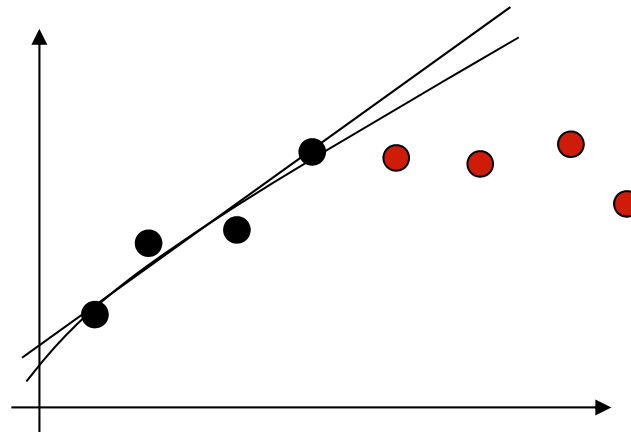


# Discussion

- The best method is moderately non-linear (quadratic)
- The linear method has mistakes both in calibration and testing
- The highly non-linear method has a calibration error null but a high testing error. Such a model is "too specialized" in describing the calibration data and is not able to generalize.
- This effect is called overfitting and is typical in the case of highly non-linear models.

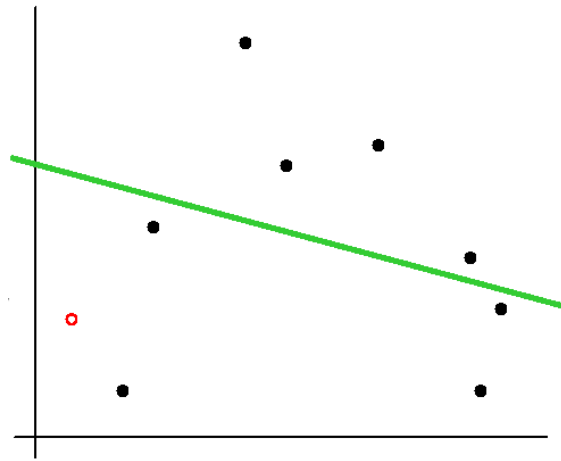
## Some consideration on the test-set

- The method is very simple but requires several sets of data.
- The selection of the data is not easy in general should be done randomly but you have to avoid the two sets unbalancing
- You should check that the two sets have the same variance and the same average
- If the two sets are uncorrelated there may be apparent overfitting phenomena
- Apart from simple cases, usually the models fail in the prediction of measurements outside the range for calibration.

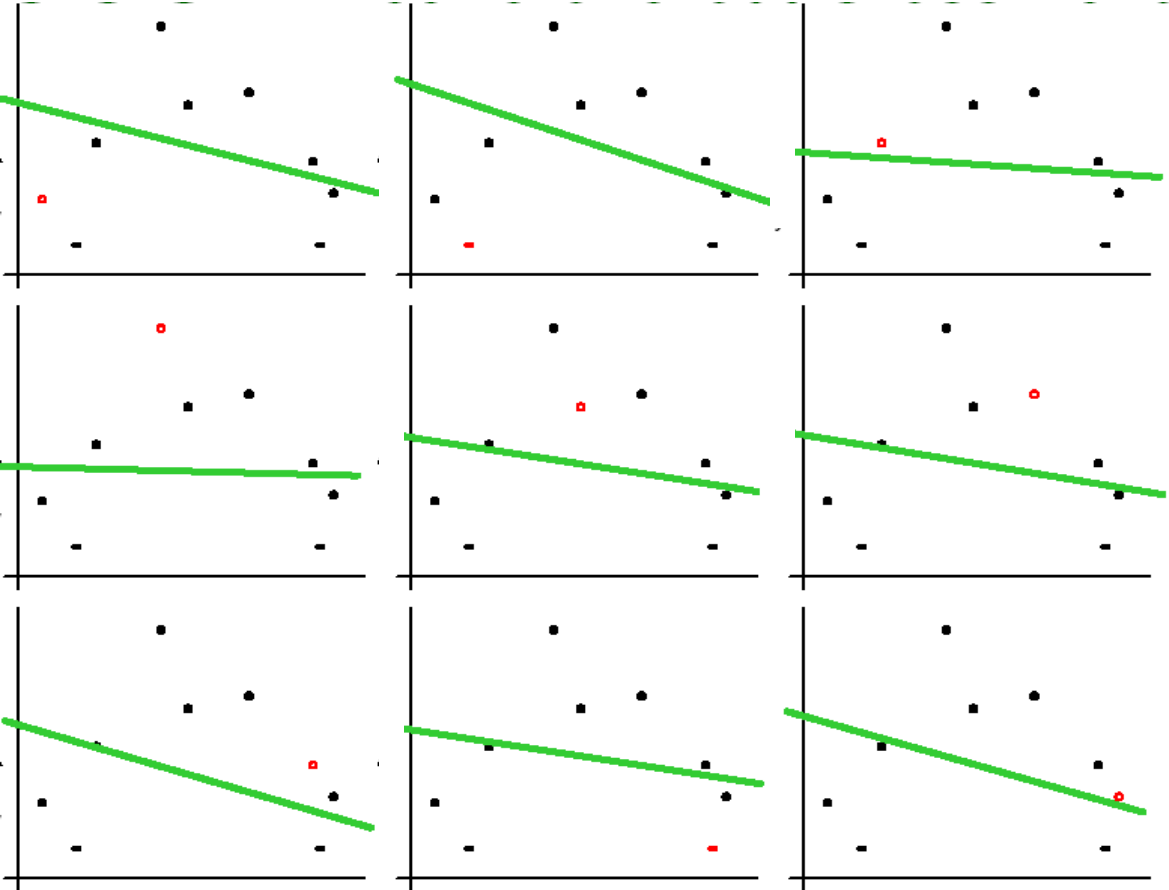


# Leave-One-Out cross-validation

- When the number of data becomes small it is necessary to use other strategies for the selection of the feature and the error estimation.
- The most used method is the leave-one-out
- Leave-one-out cross-validation (LOOCV) is a particular case of leave-p-out cross-validation with  $p = 1$ . The process looks similar, however with cross-validation you compute a statistic on the left-out sample(s), while in the other case you compute a statistic from the kept samples only.
- LOO cross-validation does not have the problem of excessive compute time as general LpO cross-validation



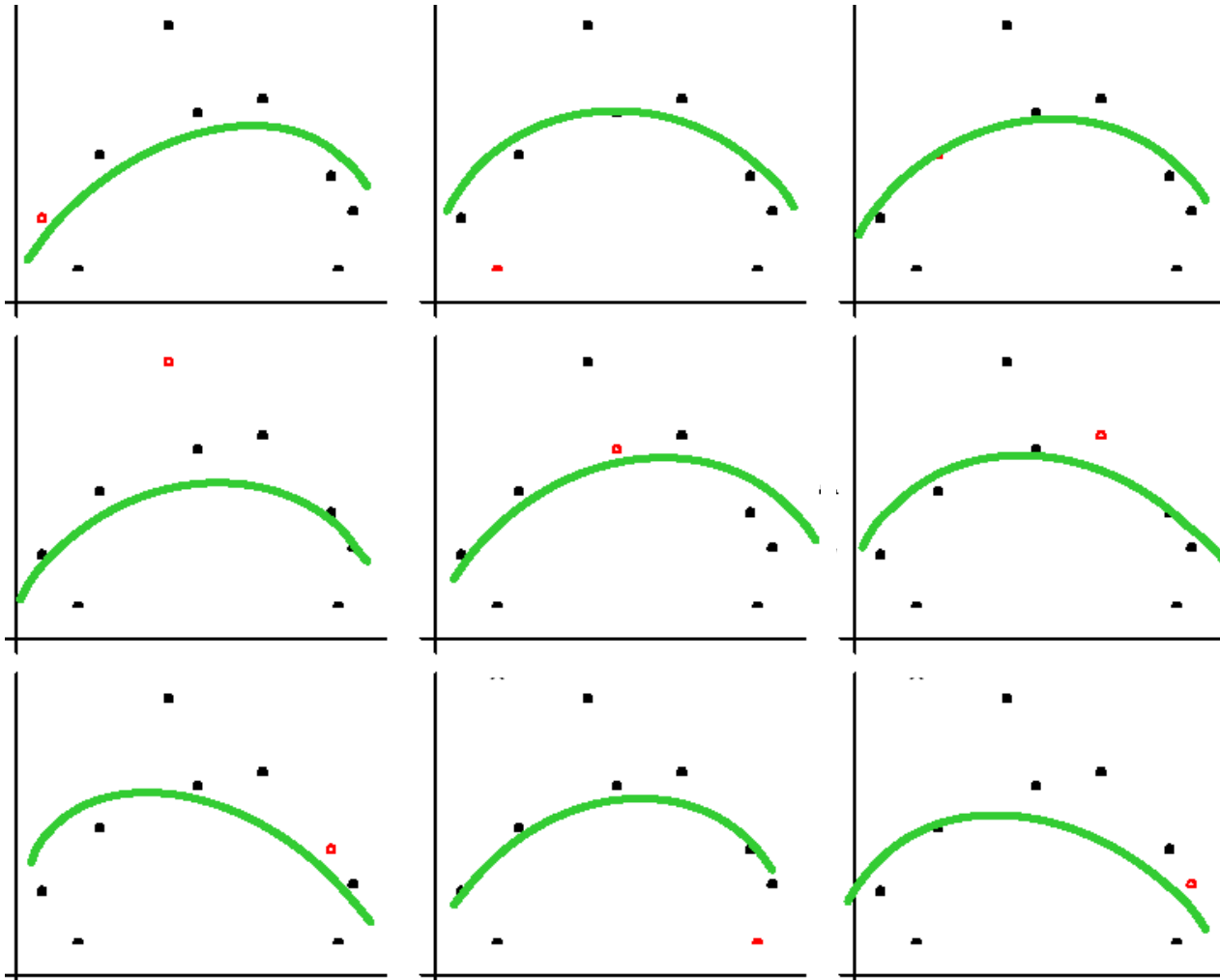
# LOO linear model



RMSECV=2.12

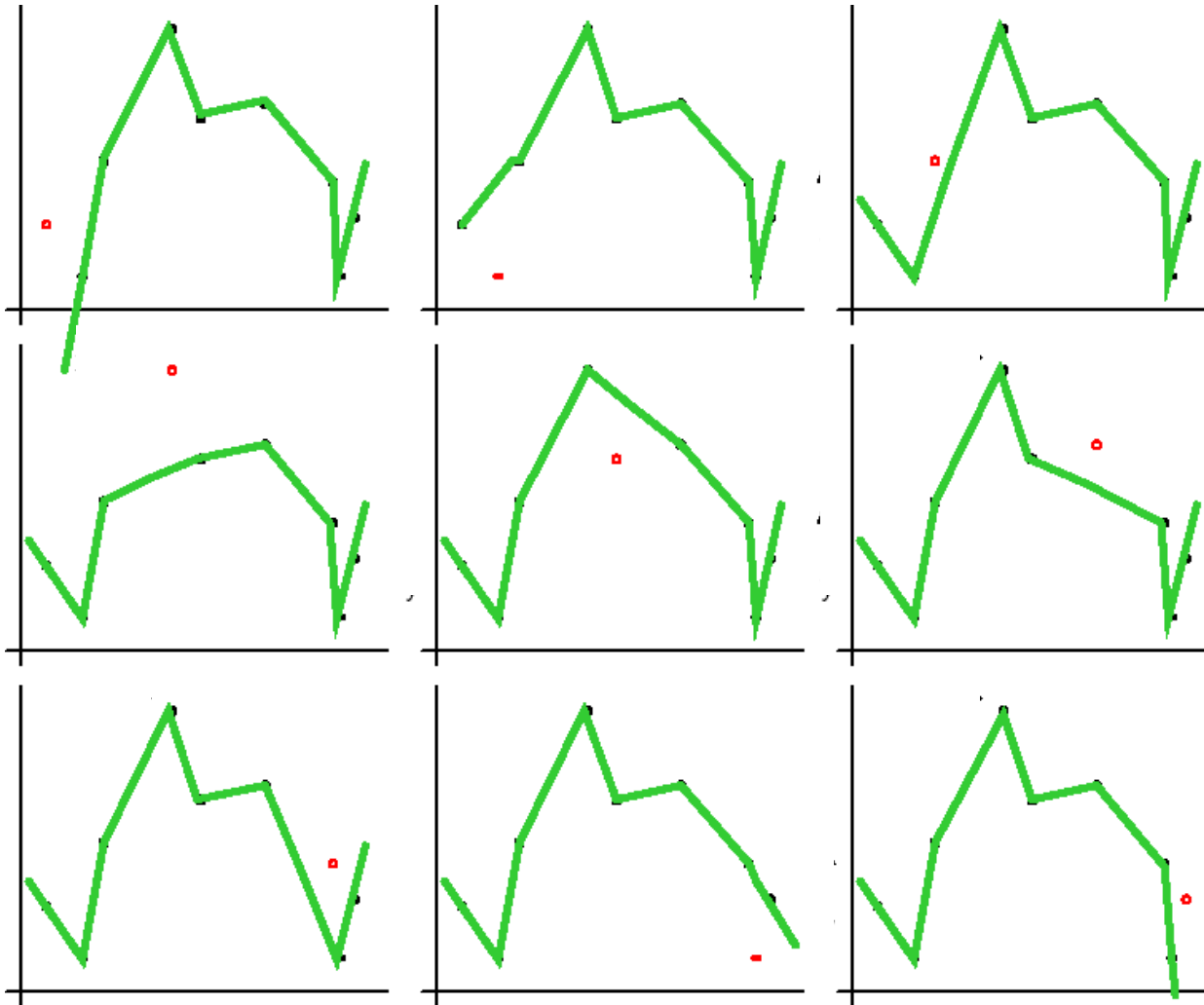


# LOO no linear model



RMSECV=0.96

# LOO highly no linear model



RMSECV=3.33

## test – LOO Comparison

- LOO provides a better estimation of the prediction error than the test set whose error estimate is unreliable.
- LOO takes full advantage of the entire data set.
- Obviously LOO is the method with minimum validation.
- For sets of large dimensions LOO is expensive from the points of view of the calculation.
- It can be "softened" considering more than  $k$  data sets.

# Matlab PLS toolbox *modlgui*

- data: 4 sensors TSMR for the measurement of octane and toluene

The image displays the Matlab PLS toolbox *modlgui* interface. On the left is a vertical menu with options: MODL\_File, Load Data, Load Model, Load Scale, Load Labels, Save Data, Save Test, Save Model, Print Info, Preferences, Clear Data, Clear Model, and Exit MODL. The main window is titled 'Linear Regression' and contains the following information:

Var: s0,nt0  
 Data: modeled (calibration set)  
 Size: 24 by 4, 24 by 2  
 Samp Lbls:  
 Var Lbls:

Model: calibrated on loaded data  
 Method: SIMPLS  
 LV(s): 3  
 Data: 24 by 4, 24 by 2  
 Scaling: autoscaled

No. LVs: 3  show parameters

Latent Variable	Percent Variance X-Block Captured by Model		Percent Variance Y-Block Captured by Model	
	This LV	Cum	This LV	Cum
1	93.48	93.48	47.70	47.70
2	6.50	99.97	51.69	99.38
3	0.02	100.00	0.14	99.53
4	0.00	100.00	0.02	99.55

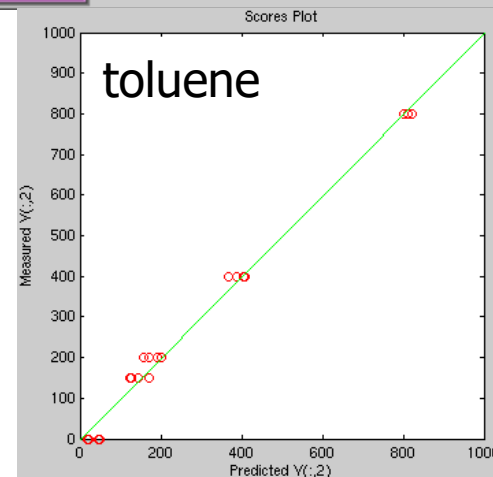
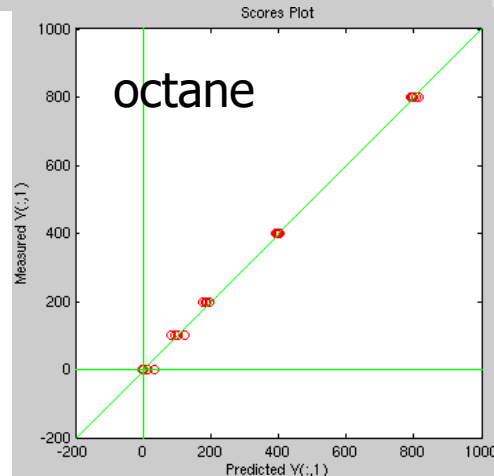
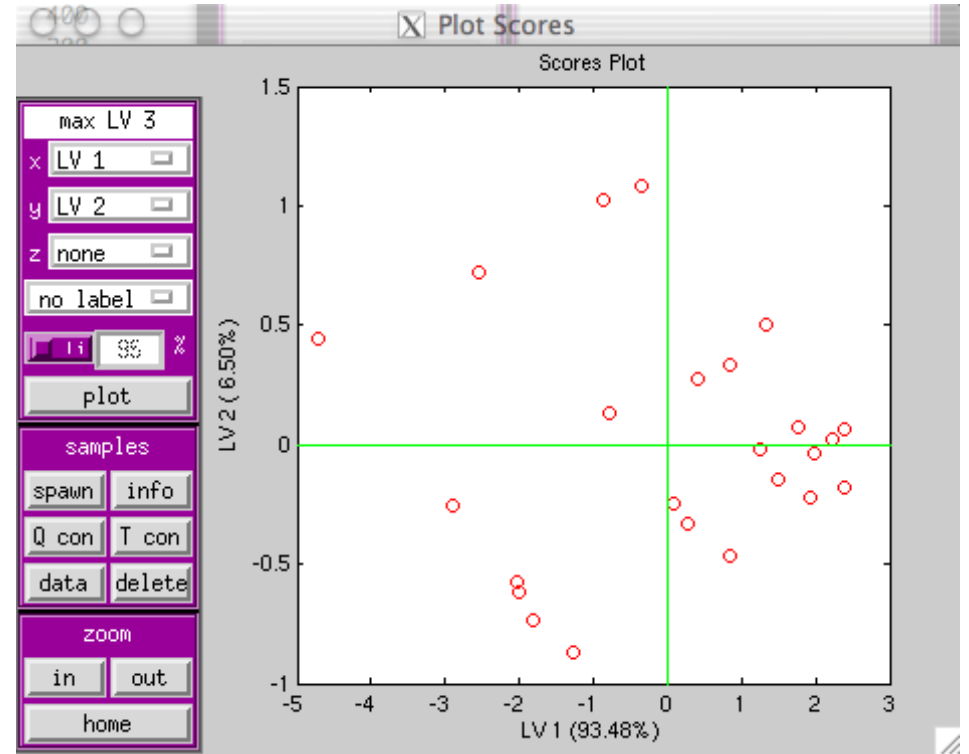
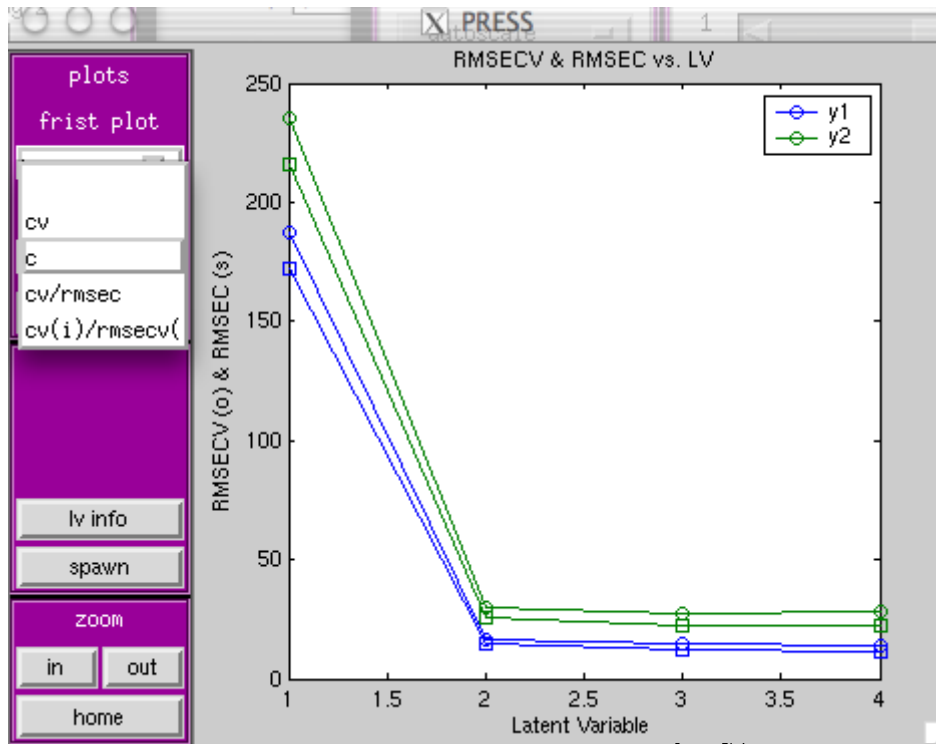
Buttons on the right: calc, apply, plots, press, scores, loads, biplot, data.

The 'Regression Parameters' window is also shown, with the following settings:

- Scaling: autoscale
- Regression: SIMPLS
- Cross Validation: leave one out
- Max LVs: 4

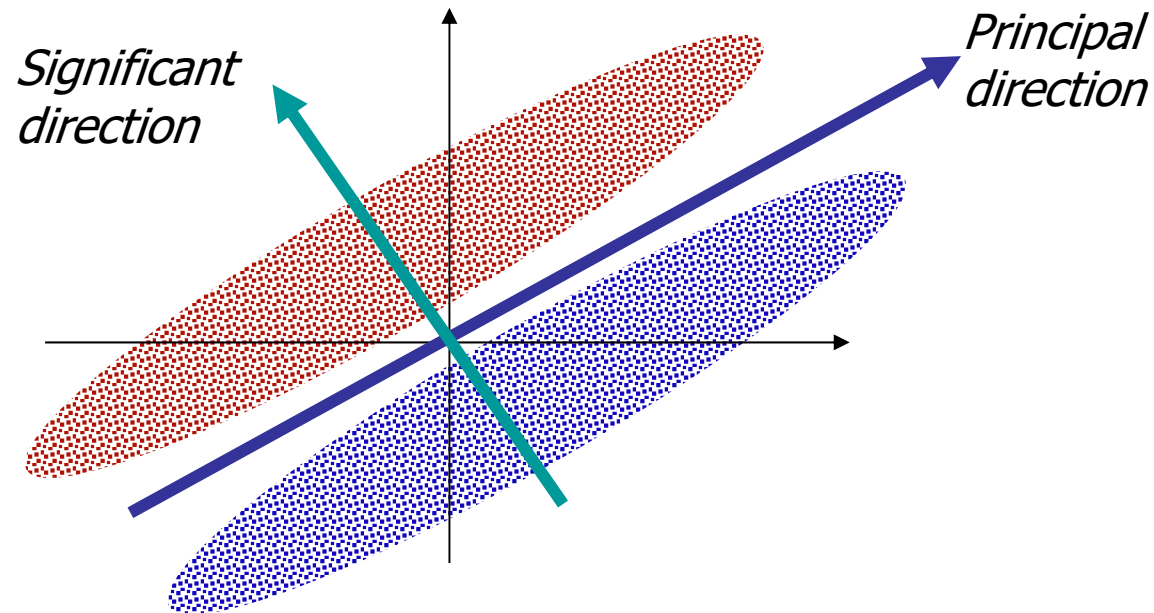
Additional options for Cross Validation: venetian blinds, contiguous block, random subsets.

# modlgui



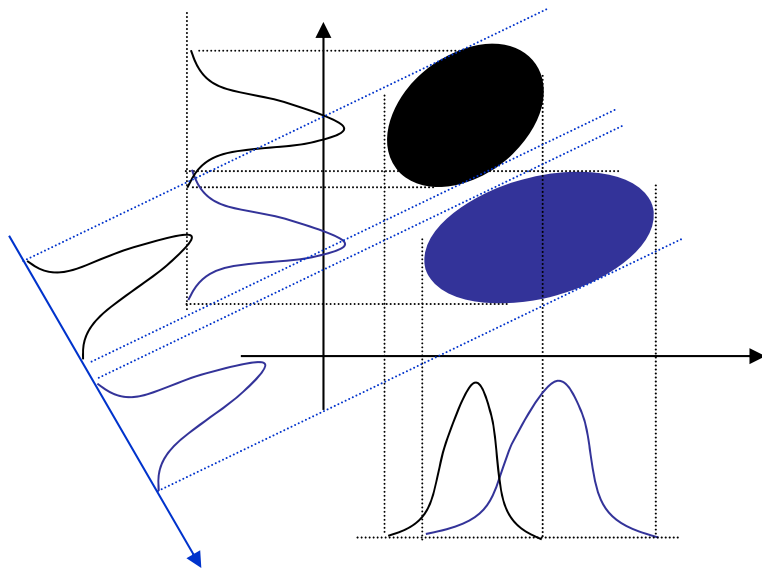
# Principal components and significant directions

- The principal components are the principal axes of the ellipsoid on the covariance matrix, nothing assures the fact that these directions are important for the problem under consideration
- The "important" direction can be found using a "supervised" view ie highlighting some properties of the data set



# Linear discriminant analysis (LDA)

- a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification



- There is a class of basic vectors (other than the PC) where the separation between classes is maximum
- If there are more classes you can introduce more directions
- Discriminatory directions are linear combination of real variables, you can study the contribution of each variable to the discriminant direction.

# PLS-Discriminant Analysis (PLS-DA)

- PLS is the ideal tool for the solution of linear classification problems.
- Minimizing the classification error, through the score and loading plots you can study what are the patterns of the variables that mostly contribute to the classification.



# PLS-DA

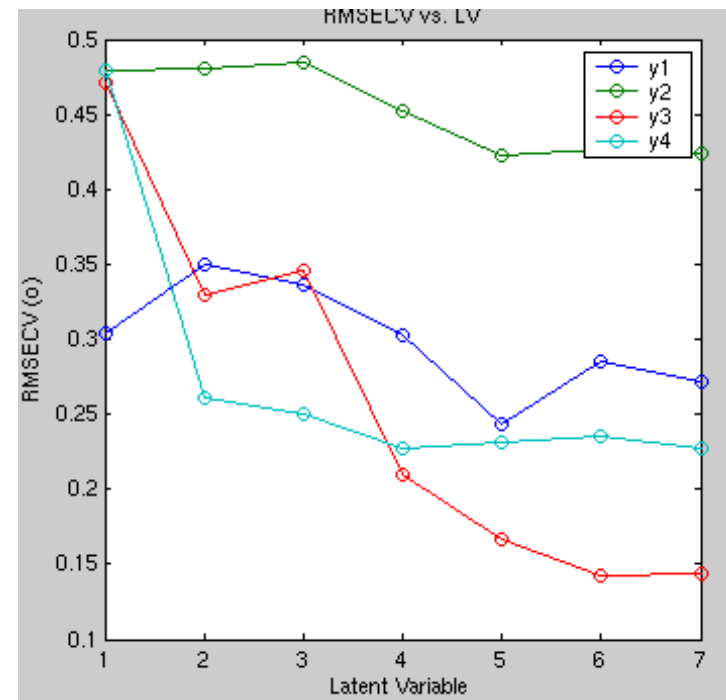
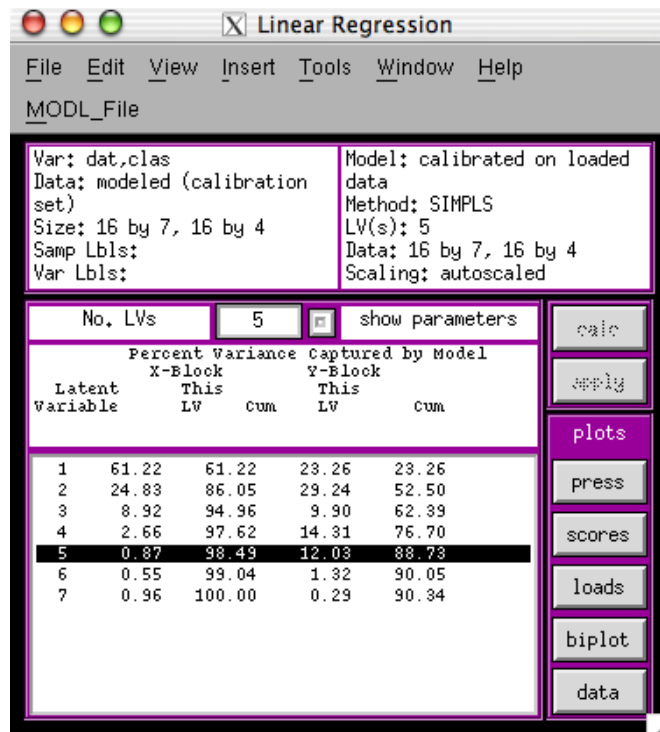
## fertilizers methods for apples

- Three fertilizer methods for apples
- 1-Urea, 2-calcium nitrate and potassium, 3- ammonium sulphates 4- One control
- four classes
- Each apple is characterized by a pattern of seven features:
- Total nitrogen, seed nitrogen, phosphorus, potassium, calcium, magnesium, weight

<b>Y</b>					<b>X</b>						
control	urea	potassium nitrates	ammonium and sulfate		total nitrogen	pit nitrogen	phosphorous	potassium	calcium	magnesium	weight
1	0	0	0	0	3240	1663	836	8747	218	388	97.3
1	0	0	0	0	3077	1663	891	8460	249	372	75.8
1	0	0	0	0	3205	1770	831	8575	261	376	78.5
1	0	0	0	0	3330	1755	889	8330	209	367	77.5
0	1	0	0	0	3755	1915	842	10375	145	408	108.2
0	1	0	0	0	5037	2180	930	10047	172	420	103.2
0	1	0	0	0	4753	2137	945	10447	160	421	95.3
0	1	0	0	0	4453	1967	850	9677	206	396	93.5
0	0	1	0	0	4200	2063	869	11190	184	398	111.8
0	0	1	0	0	5915	2050	1016	12060	184	461	109.7
0	0	1	0	0	5193	2210	958	11733	179	428	117.3
0	0	1	0	0	5347	2167	919	11910	191	420	99.6
0	0	0	1	0	5157	2357	1062	10210	136	401	86.7
0	0	0	1	1	7440	2975	1261	10820	122	446	85.5
0	0	0	1	1	6950	2527	1137	9710	191	408	70.8
0	0	0	1	1	4445	2075	846	8820	161	334	81.1

# PLS-DA

## fertilizers methods for apples



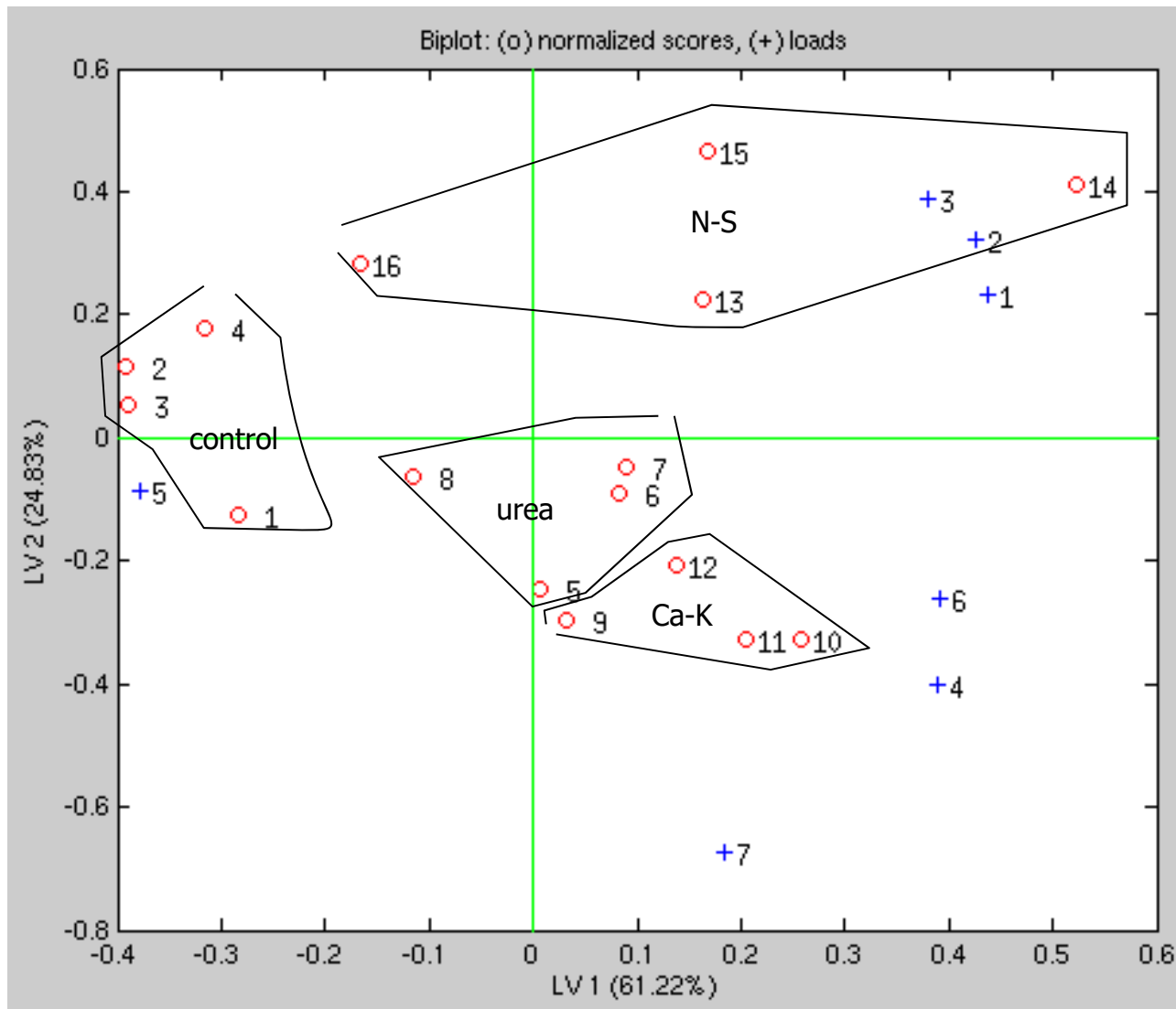
# PLS-DA

## fertilizers methods for apples

Y true				Y estimated			
1	0	0	0	<b>0.7839</b>	0.4456	-0.0168	-0.2128
1	0	0	0	<b>1.1728</b>	-0.3482	0.1035	0.0718
1	0	0	0	<b>0.8882</b>	0.1623	0.0387	-0.0892
1	0	0	0	<b>0.8729</b>	0.0584	-0.2114	0.2801
0	1	0	0	0.0515	<b>0.9332</b>	0.0552	-0.0398
0	1	0	0	0.0116	<b>0.9322</b>	-0.0086	0.0648
0	1	0	0	0.0748	<b>0.7485</b>	0.0089	0.1678
0	1	0	0	0.1801	<b>0.7226</b>	0.0919	0.0053
0	0	1	0	0.0482	-0.1203	<b>0.9887</b>	0.0835
0	0	1	0	0.0820	0.2404	<b>0.8671</b>	-0.1895
0	0	1	0	-0.0390	-0.0669	<b>1.0746</b>	0.0313
0	0	1	0	-0.1771	0.0942	<b>0.9599</b>	0.1230
0	0	0	1	0.1673	-0.0846	0.1036	<b>0.8137</b>
0	0	0	1	-0.2136	0.1390	-0.0245	<b>1.0990</b>
0	0	0	1	0.1372	-0.0736	0.0300	<b>0.9064</b>
0	0	0	1	-0.0410	0.2174	-0.0608	<b>0.8844</b>

# PLS-DA

## fertilizers methods for apples



The Scores show:

- The separation between the four groups
- From the control we have two directions: N-S and urea//Ca-K

The loadings show:

- The N-S treatment increases the total nitrogen and phosphorus in the seed
- The treatments with urea and Ca-K increase potassium, magnesium, and the weight of the fruit
- The greater amount of calcium is found in the control apples

## **Using Principal Component Analysis Modeling to Monitor Temperature Sensors in a Nuclear Research Reactor**

Rosani M. L. Penha  
Centro de Energia Nuclear  
Instituto de Pesquisas Energéticas e Nucleares - Ipen  
São Paulo, SP 05508-900 Brasil  
Email: rmpenha@net.ipen.br  
Phone: (11) 3817-7421

J. Wesley Hines  
Nuclear Engineering Department  
The University of Tennessee  
Knoxville, TN 37996-0750 USA  
Email: hines@utkux.utk.edu  
Phone: (865) 974-6561

### **Abstract**

Principal Component Analysis (PCA) is a data-driven modeling technique that transforms a set of correlated variables into a smaller set of new variables (principal components) that are uncorrelated and retain most of the original information. Thus, a reduced dimension PC model can be used to detect and diagnose abnormalities in the original system in a robust way. This paper presents an application of using the PCA modeling technique to detect abnormalities in temperature sensors used to monitor the primary loop of a 2MW research pool reactor. The PCA model maps the temperature variables into a lower dimensional space and tracks their behavior using  $T^2$  and Q statistics. The Hotelling's  $T^2$  statistic measures the variation within the PCA model while the Q statistic measures the variation outside of the PCA model. Five temperature sensors are considered in the model. Three sensors are located inside the pool and two sensors are located at the primary loop piping. The reduced dimension PCA model has well behaved  $T^2$  and Q statistics. To test the functionality of the model, a drift was imposed sequentially to each temperature sensor, and the PCA model was able to detect and identify the faulty sensors at very low thresholds.

### **Introduction**

Many researchers have addressed the use of Principal Component Analysis (PCA) modeling in the monitoring and fault detection of process sensors [1, 2, 3]. The objective of this work is to construct a PCA model that best maps a research reactor's primary loop temperature sensors into a lower dimensional space, in order to characterize the behavior of these variables through the use of  $T^2$  and Q statistics.

Five sensors have been chosen to be monitored. Three sensors are located inside the pool: T1 is located near the surface of the pool, T2 is located at the midway down into the pool and T3 is located just above of the reactor core. Temperature sensor T4 is located in the outlet pipe that takes the water from the core to the decay tank. At the outlet of the decay tank we have the sensor T6. The schematic diagram of the pool research reactor is shown in the Appendix.

### **Principal Component Analysis**

PCA is a method used to transform a set of correlated variables into a smaller set of new variables that are uncorrelated and retain most of the original information, where the variation in the signals is considered

to be the information. PCA takes advantage of redundant information existent in highly correlated variables to reduce the dimensionality. After developing a model using good (training) data, the reduced dimension PC model can be used to detect and diagnose process abnormalities in a robust way [4]. For a basic reference book on PCA see Jolliffe [5].

PCA decomposes the data matrix  $\mathbf{X}$  ( $m$  samples,  $n$  variables) as the sum of the outer product of vectors  $\mathbf{t}_i$  and  $\mathbf{p}_i$  plus a residual matrix  $\mathbf{E}$  [1]:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \dots + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E} \quad (1)$$

The vectors  $\mathbf{p}_i$  are orthonormal, and the vectors  $\mathbf{t}_i$  are orthogonal, that is:

$$\mathbf{p}_i^T\mathbf{p}_j = 1, \quad \text{if } i = j \quad \text{and} \quad \mathbf{p}_i^T\mathbf{p}_j = 0, \quad \text{if } i \neq j; \quad \text{and} \quad (2)$$

$$\mathbf{t}_i^T\mathbf{t}_j = 0 \quad \text{if } i \neq j \quad (3)$$

Also we can note that  $\mathbf{t}_i$  is the linear combination of the original  $\mathbf{X}$  data defined by the transformation vector  $\mathbf{p}_i$ :

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i \quad (4)$$

The vectors  $\mathbf{t}_i$  are known as the principal component *scores* and contain information on how the *samples* are related to each other. The  $\mathbf{p}_i$  vectors are the *eigenvectors* of the covariance of matrix  $\mathbf{X}$ . They are known as the principal component *loadings* and contain information on how *variables* are related to each other. In fact, the PCA splits the data matrix  $\mathbf{X}$  in two parts: one describes the system variation (the process model  $\mathbf{T}_k\mathbf{P}_k^T$ ) and the other captures the noise or unmodeled information (residual variance  $\mathbf{E}$ ). This division is not always perfect, but it routinely provides good results [1]. It is very important to distinguish the number of components (dimension) that are to be kept in the model.

The number of *principal components*  $k$ , to retain in the model must be less than or equal to the smaller dimension of  $\mathbf{X}$ , i.e.,  $k \leq \min\{m, n\}$  [2]. The goodness of the model depends on a good choice of how many PCs to keep.

There are different criteria to choose the number of PCs [6]. The eigenvalues associated with each eigenvector or principal component:  $\mathbf{p}_i$ , tell us how much information (variation) each PC explains. Then we can look at the cumulative percent variance captured by the first few PCs and choose a number of PCs that accounts for most of the variability of the data (i.e. 90% to 99%).

Alternatively, we can look for a *knee* point in the residual percent variance plotted against the number of principal components. This is thought to be the natural break between the useful PCs and residual noise.

Another criterion is to accept the PCs whose eigenvalues are above the average eigenvalue. In correlated-based PCA, the average eigenvalue is 1. It is advisable to investigate more than one criterion since there is no universally accepted methodology.

## PCA Model

The concept of principal components is shown graphically in Figure 1. The figure shows a PCA model constructed for a data set of three variables. We can see that the samples lie mainly on a plane, thus the data is well described by a two principal component model (2 PCs). The first PC aligns with the greatest variation in the data while the second PC aligns with the greatest remaining variance that is orthogonal to the first PC.

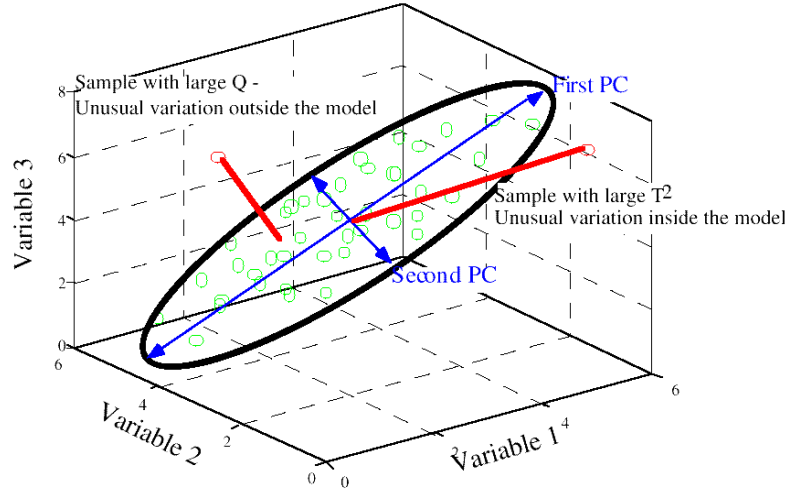


Figure 1 - Principal Component Model Staying on a Plane, Showing T<sup>2</sup> and Q Outliers

There are two statistics commonly employed in evaluating new data using a previously developed PCA model: Q statistic and Hotelling's T<sup>2</sup> statistic.

The Q statistic measures the *lack of model fit* for each sample. It indicates how well each sample conforms to the PCA model by measuring the distance a data point falls from the PC model. It is calculated as the difference between the data point and its projection on the PC model. It gives the lack of fit to the model [1].

$$\mathbf{Q}_i = \mathbf{e}_i \mathbf{e}_i^T = \mathbf{x}_i (\mathbf{I} - \mathbf{P}_k \mathbf{P}_k^T) \mathbf{x}_i^T \quad (5)$$

The Hotelling's T<sup>2</sup> measures the variation *within* the PCA model. T<sup>2</sup> is the sum of the normalized squared scores defined as [1]:

$$T_i^2 = \mathbf{t}_i (\mathbf{T}_k^T \mathbf{T}_k)^{-1} \mathbf{t}_i^T \quad (6)$$

Statistical limits can be developed for sample scores, T<sup>2</sup> and Q, and individual residuals. If some point falls outside the limits for a specific confidence interval (95%, for example), this point may be considered to be an outlier and may be not representative of the data used to develop the PCA model.

The PCA model of a data matrix includes mean and variance scaling vectors, eigenvalues, loadings, statistics limits on the scores, Q and T<sup>2</sup>. The model can be used with new process data to detect changes in the system that generated the original data. After detecting a probable outlier due to extreme T<sup>2</sup> or Q values, we can investigate the inputs that contribute to the extreme statistical value.

## Methodology

A PCA model was developed to describe correlations of the primary loop temperature variables of the IPEN research reactor located in Sao Paulo, Brazil (see Appendix). The reactor data acquisition system records a snapshot of data once a minute for a complete fuel cycle that usually lasts a couple of days. A representative data set that corresponds to 25 hours of reactor operation was used to construct the model. A second data set, corresponding to 15 hours of another cycle operation, is used to validate the PCA model. Lastly, an artificial drift is imposed on each sensor to test the sensitivity of the model.

The input data used to develop the model was arranged in a matrix with 1501 rows (samples) and 5 columns (temperature variables). The time plot of these variables is shown in Figure 2.

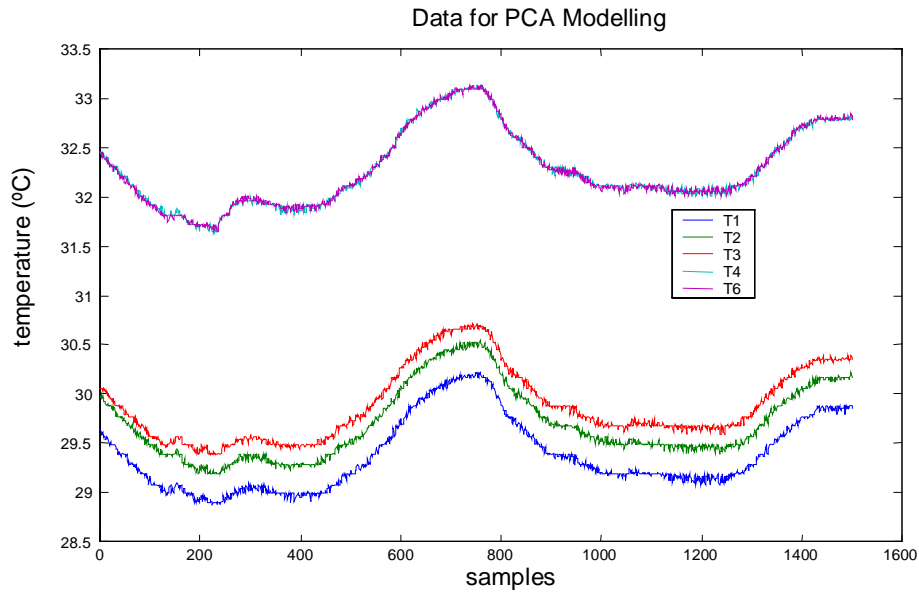


Figure 2 - Temperature Model Development Data

As shown in Table 1; T1, T2, T3, T4 and T6 are highly correlated, meaning that they vary together. This redundancy in the measurements allows us to build a PCA model that will retain the most of information in a few principal components.

corr. coef.	T1	T2	T3	T4	T6
T1	1.0000	0.9963	0.9960	0.9918	0.9903
T2	0.9963	1.0000	0.9945	0.9887	0.9862
T3	0.9960	0.9945	1.0000	0.9939	0.9927
T4	0.9918	0.9887	0.9939	1.0000	0.9975
T6	0.9903	0.9862	0.9927	0.9975	1.0000

Table 1 - Correlation Coefficients

To construct the PCA model, the input matrix was divided into a training set (to develop the model) and a test set in an odd-even manner. The input matrix was mean centered and scaled to a unit variance. This is necessary for PCA model development. PCA functions in MATLAB were used to calculate the principal components, the eigenvalues, and the amount of variance explained by each PC component.

Figure3 is a plot of the eigenvalues versus the PC number and is used to help to choose the number of PCs to keep in the model. The size of the eigenvalue equals the amount of variance explained in the corresponding PC. We use the log plot that can show the break when the eigenvalues cover several orders of magnitude, as in this case. This plot is used to identify a knee point where the PCs above the knee contain information and the PCs below the knee represent noise. The knee occurs between 1 and 2 PCs. The second PC does contribute useful information. We will keep 2 PCs that explain 99.80% of the variance.



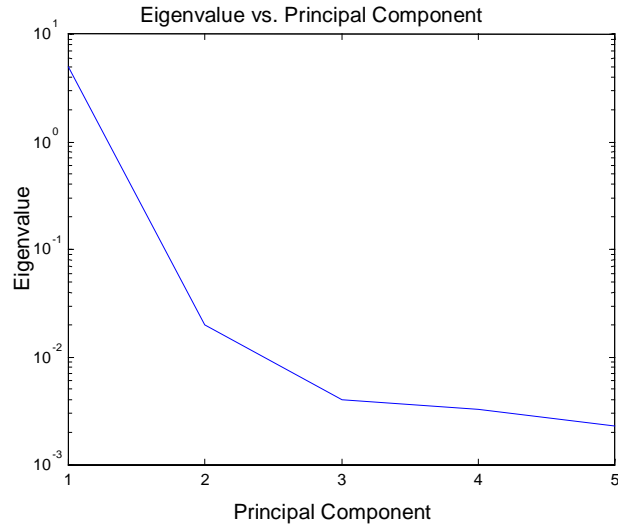


Figure 3 - Eigenvalues versus Principal Components

Next we plot and analyze the loadings on the retained principal components: PC1 and PC2.

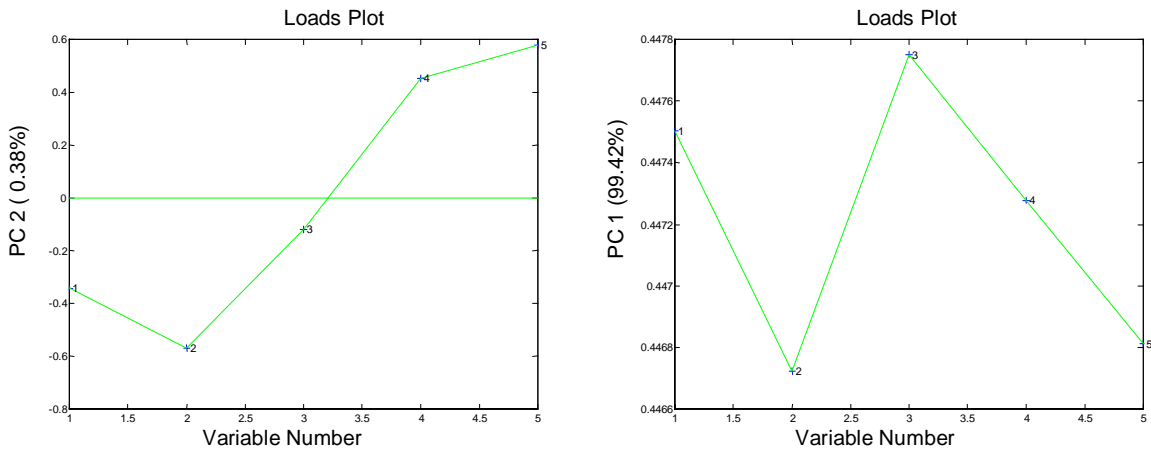


Figure 4 - Loadings on Principal Component 1 and 2

The principal component loadings are the weightings of each input to the specific PC. They show the underlying relationship among the variables. PC1 weights all the variables positively, so it is a gross measurement of the temperature in the reactor. PC2 accounts for the differences between variables {T1, T2, T3} and {T4, T6}. This makes sense, since the three first variables are located physically near each other and are related to the temperature inside the pool while the last two sensors are located outside the pool. Although PC2 accounts for a small amount of variation when compared to PC1, it is important to describe the differences between the two groups of variables.

Figure 5 is a plot of PC1 versus PC2 and shows that PC1 and PC2 are uncorrelated. If there were a noticeable relationship in this plot, it would be attributed to non-linear relationships in the data. The PC

technique removes all linear correlations and results in a scatter plot when the non-linear relationships are small or nonexistent.

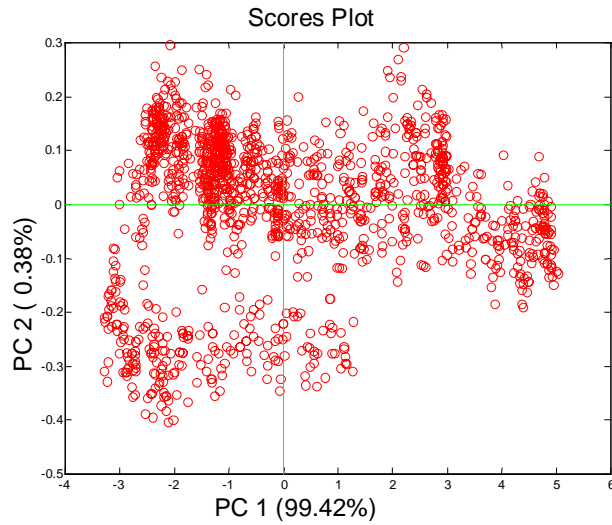


Figure 5 - Scores on PC1 versus Scores on PC2

The  $T^2$  and  $Q$  statistics are shown in Figure 6. The dashed lines represent a 95% confidence interval used to identify possible outliers. The  $T^2$  and  $Q$  residuals show the data fits the model well.

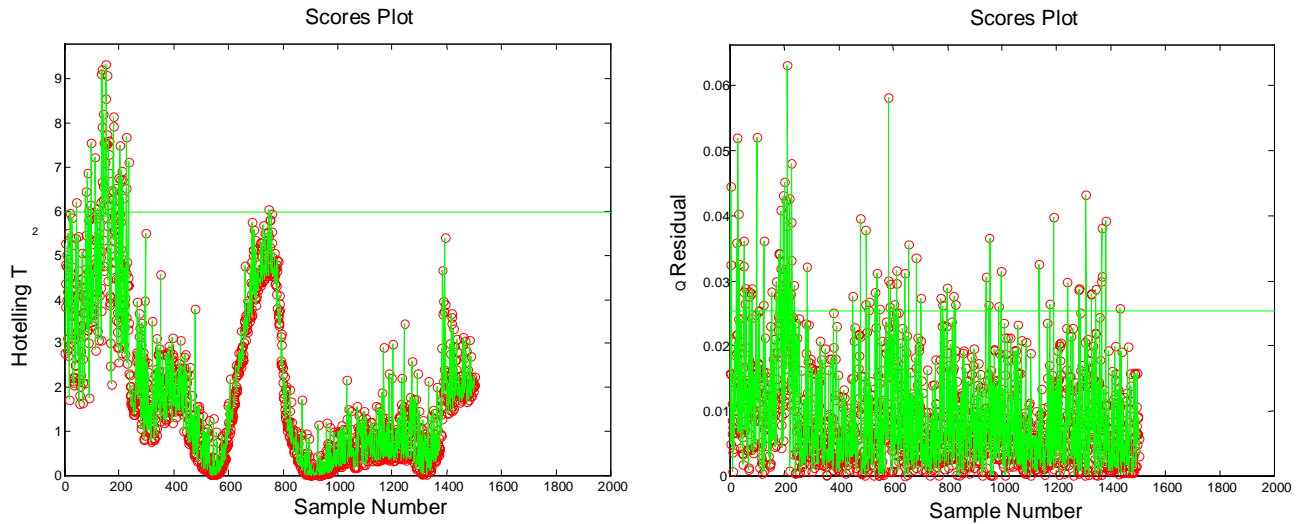


Figure 6 -  $T^2$  Statistic on 2 PC Model (left) and  $Q$  Statistic on 2 PC Model (right)

### Validation

To validate the PCA model, another data set corresponding to another week operation was applied to the PCA model. The resultant  $T^2$  and  $Q$  residuals are shown in Figure 7.

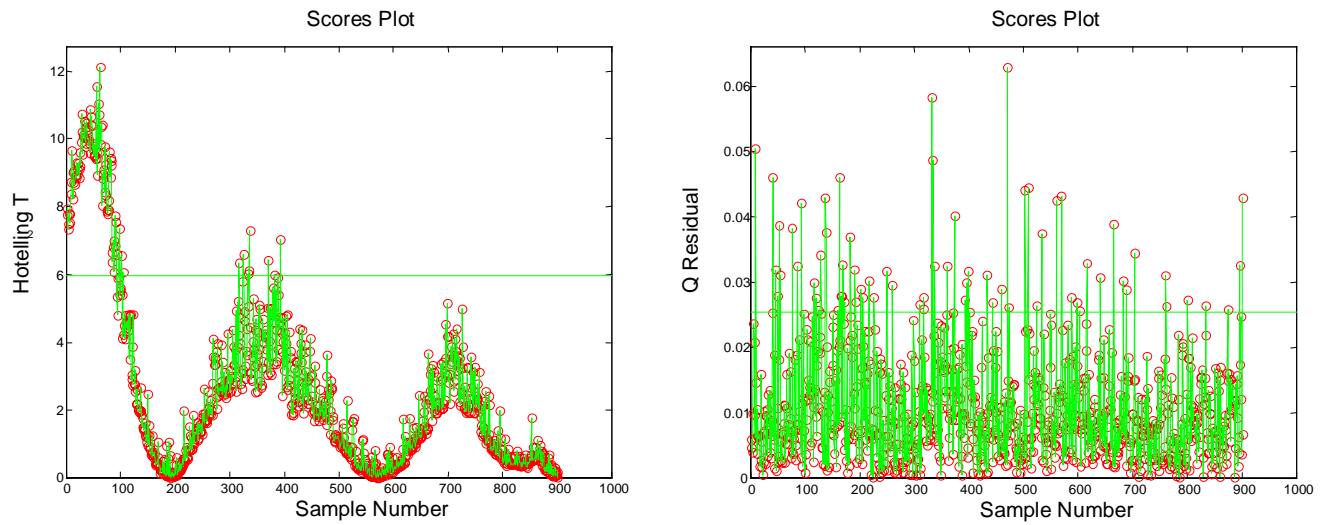


Figure 7 -  $T^2$  Statistic on Validation Data (left) and Q Statistic on Validation Data (right)

Since the  $T^2$  and Q statistics are within the confidence limits, the model represents the validation data set well. The PCA gives the best linear model in sense of minimum squared error.

### Drifted Sensor

To verify the model ability to detect and identify the drifting sensors, an artificial drift (ramp) was applied separately to each input variable. The drift simulates a common problem that affects process sensors and may result from aging. The simulated drift is a ramp that grows to 0.05°C (maximum value) for a temperature variable that originally varies from 28.87°C to 30.22°C. This small drift corresponds to a 0.17% change and is imperceptible in a time profile. Figure 8 shows the  $T^2$  and Q statistics for this case.

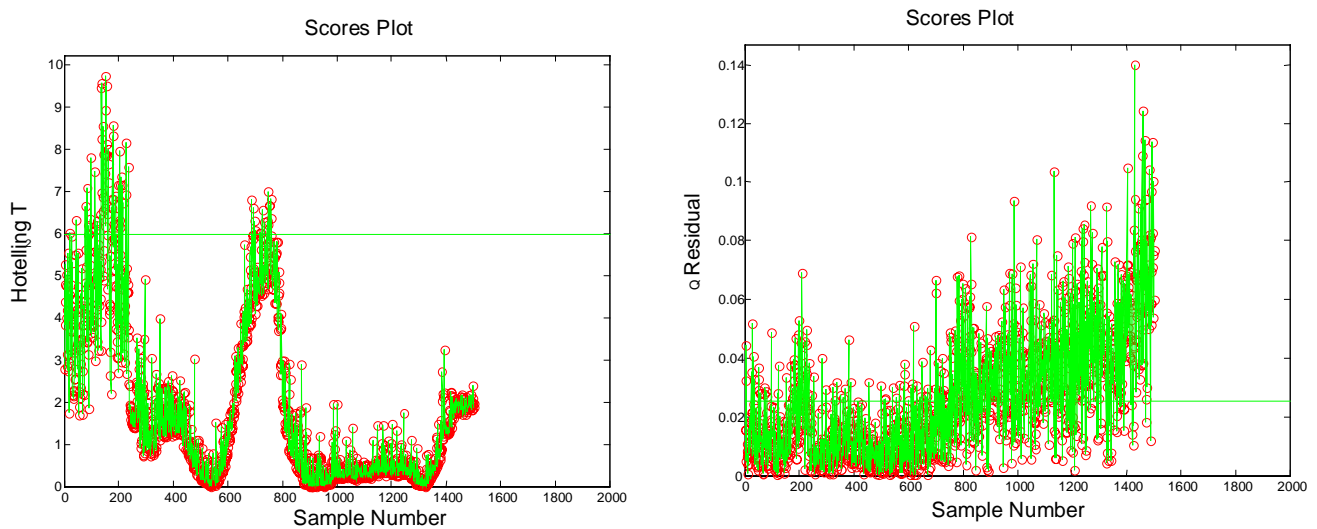


Figure 8 -  $T^2$  Statistic on Drifted Data (left) and Q Statistic on Drifted Data (right)

The  $T^2$  statistic doesn't seem to be out of the ordinary but the Q statistic plot shows that its values are increasing over time. This indicates that something is going on that is not in the original data. We can look at the contribution of each input to the large Q statistic. Through this analysis, it is possible to determine which variable is responsible for the unusual Q behavior. Samples 829, 1238 and 1430 were investigated. The contributions to the Q statistics are plotted in Figures 9 and 10.

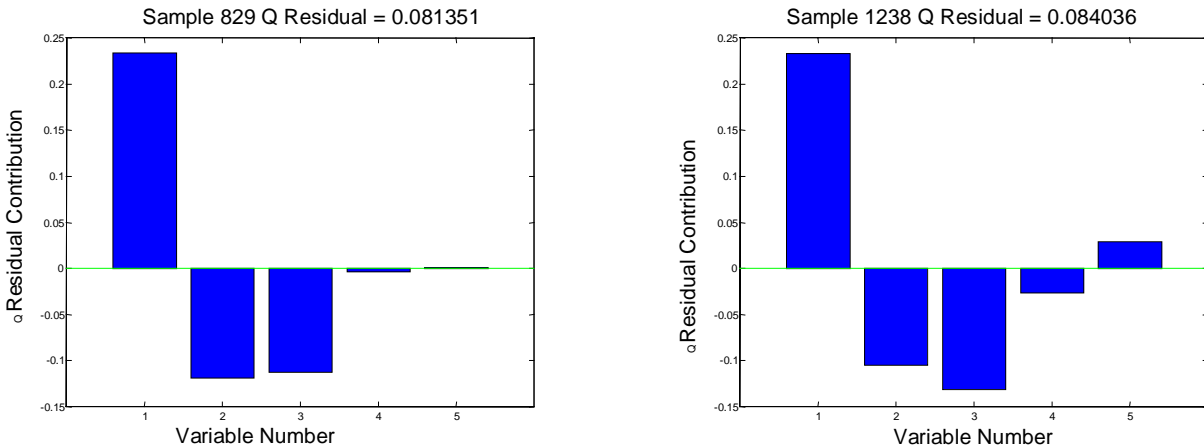


Figure 9 - Contributions (T1, T2, T3, T4, T6) to Q statistics of samples 829 (left) and 1238 (right)

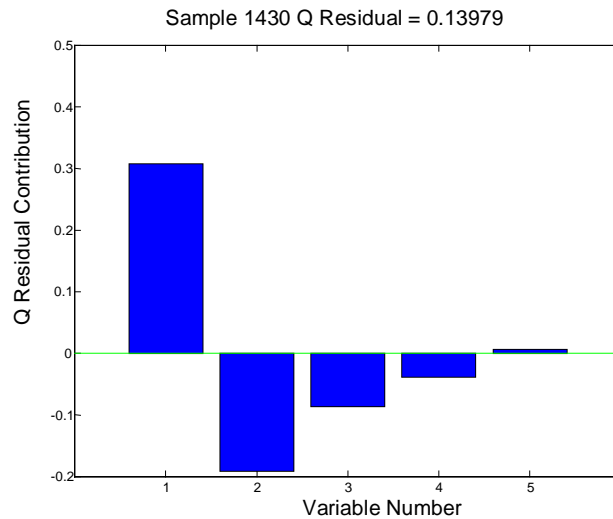


Figure 10 - Contributions (T1, T2, T3, T4, T6) to Q statistics of sample 1430

From Figures 9 and 10, it is easy to see that the variable T1 is the responsible for the unusually large Q statistic. This agrees with the fact that the drift was added to variable T1. Artificial drifts that were added to each of the other variables were detected and the drifted variable was identified using the Q statistic. When a ramp drift with 0.5°C maximum value is added to the T1 variable, the deviation of the Q statistic is even more evident as shown in Figure 11.

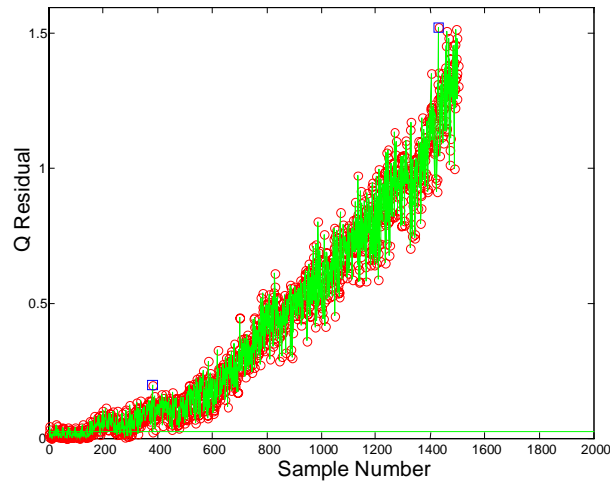


Figure 11 - Q Statistic on Drifted Data

The contributions of the variables to the Q residuals on the samples 379 and 1430 are shown in Figure 12.

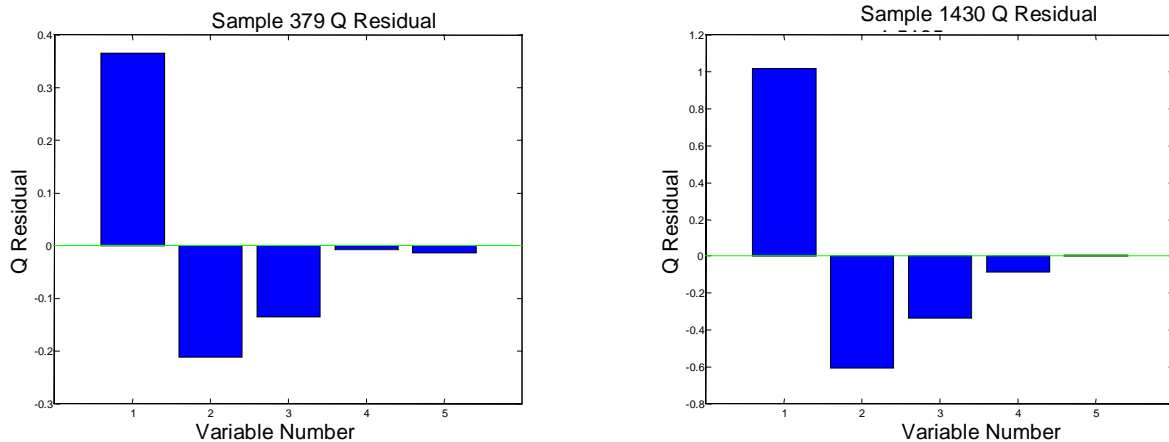


Figure 12 - Contributions of (T1, T2, T3, T4, and T6) to Q statistics of samples 379 and 1430

Again we observe the large contribution of sensor T1 to the Q statistic.

## Conclusions

A PCA model with two principal components was developed to describe five temperature sensors in a nuclear research reactor. The model fitted the data well, as shown by the  $T^2$  and Q statistics. The model was tested with a validation data set from a separate reactor fuel cycle and the model performed well. Artificial drifts were added to the variables and the model both detected and identifies the drifted variables. The PCA model was determined to be a good method to monitor the temperature sensors in this plant. This is due to the highly correlations in the data and to the insignificant non-linearities present.

If non-linearities or time delays were present in the system, other methods such as non-linear partial least squares or neural networks may be used.

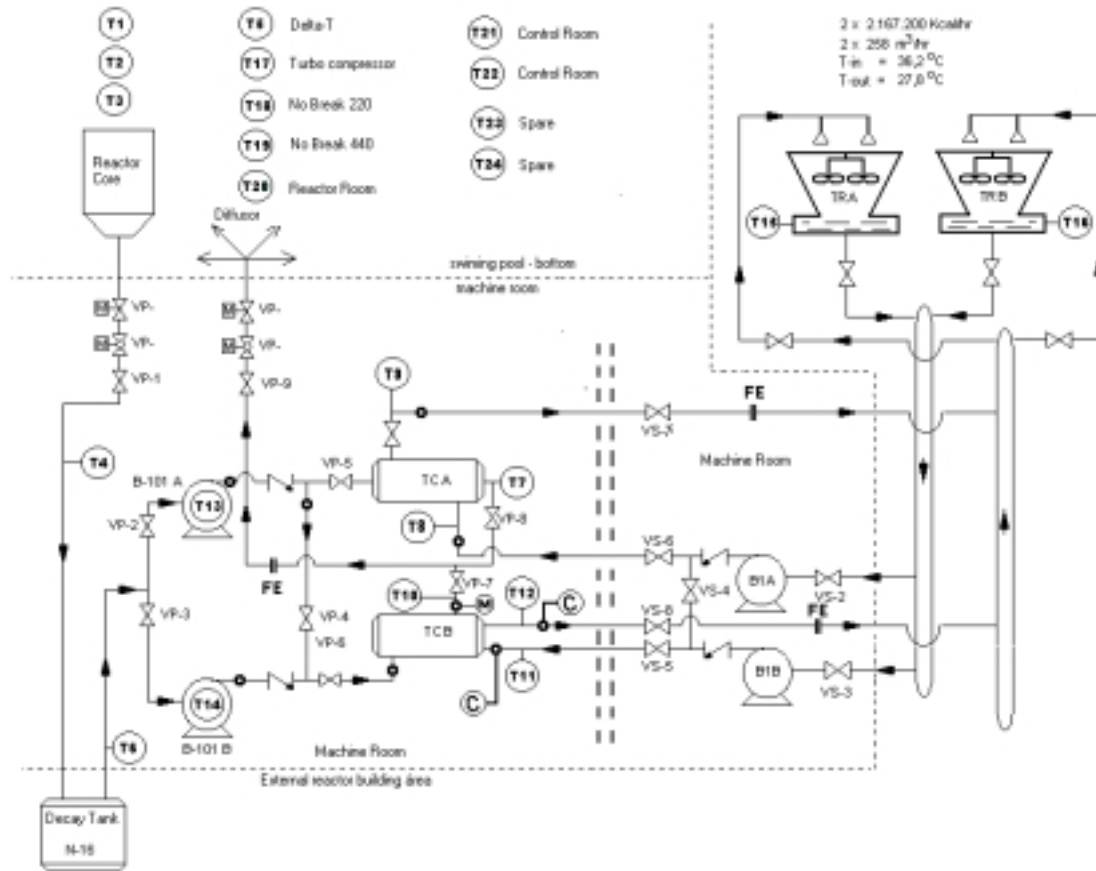
### **Acknowledgments**

The authors wish to thank the National Science Foundation (NSF) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) that gave support to this joint research project.

### **References**

- [1] Gallagher, N.B., B.M. Wise, S.W. Butler, D.D.White Jr. and G.G. Barna (1997), Development and Benchmarking of Multivariate of Statistical Process Control Tools for a Semiconductor Etch Process: Improving Robustness through Model Updating, ADCHEM 1997, Banff.
- [2] Wise, B.M., N.B. Gallagher, S.W. Butler, D.D.White Jr. and G.G. Barna (1996), Development and Benchmarking of Multivariate of Statistical Process Control Tools for a Semiconductor Etch Process: Impact of Measurement Selection and Data Treatment on Sensitivity, Safeprocess '97, Hull, England August 26-27.
- [3] Dunia R., Qin S. J. (1998), Joint Diagnosis of Process and Sensor Faults Using Principal Component Analysis. Control Engineering Practice 6 (1998) 457-469.
- [4] Hines J.W., (2000), *PCA models*. NE 589 class notes, The University of Tennessee.
- [5] Jolliffe, L.T., (1986), *Principal Components Analysis*, Springer-Verlag.
- [6] Valle S., Weihua Li and S.J. Qin. (1999), Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a comparison to Other Methods, *Ind. Eng. Chem. Res.*, 38, 4389-4401.

# Appendix - Schematic Diagram of the IEA-R1 Pool Research Reactor



Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali  
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY  
I YEAR

Course:

**EXPERIMENTAL DESIGN AND  
CHEMOMETRICS IN FOOD**  
(5 credits – 38 hours)

Teacher: Marcello Mascini  
([mmascini@unite.it](mailto:mmascini@unite.it))

The Teacher is available to answer questions at the end of the lesson, or on request by mail



# The course is split in 4 units

## UNIT 1: statistical regression

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the  $\chi^2$  method, Validation of the model

## UNIT 3: Data Matrices and sensor arrays

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

## UNIT 2: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

## UNIT 4: Elements of Pattern recognition

cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

# **UNIT 4: Elements of Pattern recognition**

cluster analysis

Normalization

The space representation (PCA)

Examples of PCA

Discriminant analysis (DA) PLS-DA

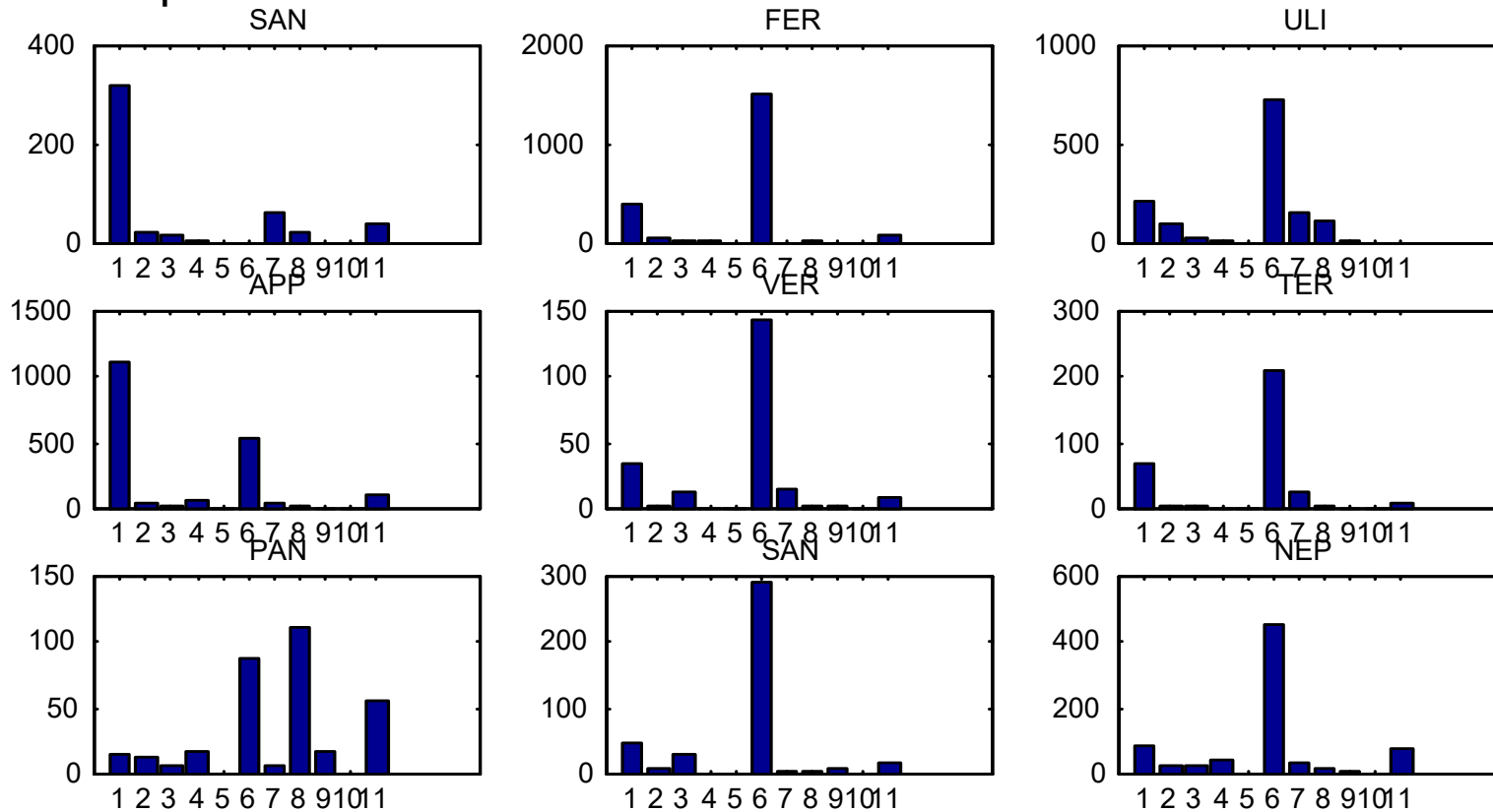
Examples of PLS-DA

# Definitions

- Pattern: set of features defining the properties of a complex object
- Class: a set of objects having some important properties in common
- Classifying: mathematical operation for which a sample, described by a number of features, is assigned to a particular class.
- The set of features is called Pattern, the classification operation is called Pattern Recognition.
- The relationship between the sample and the class is not explicit, it depends on the features chosen to describe the objects
- The pattern recognition problem is "interesting" when the individual features are not able to identify samples
  
- Fruits:
- Features: weight, shape, color, sugar, acid, .....

# Pattern of mineral water

- Ionic profile of mineral water



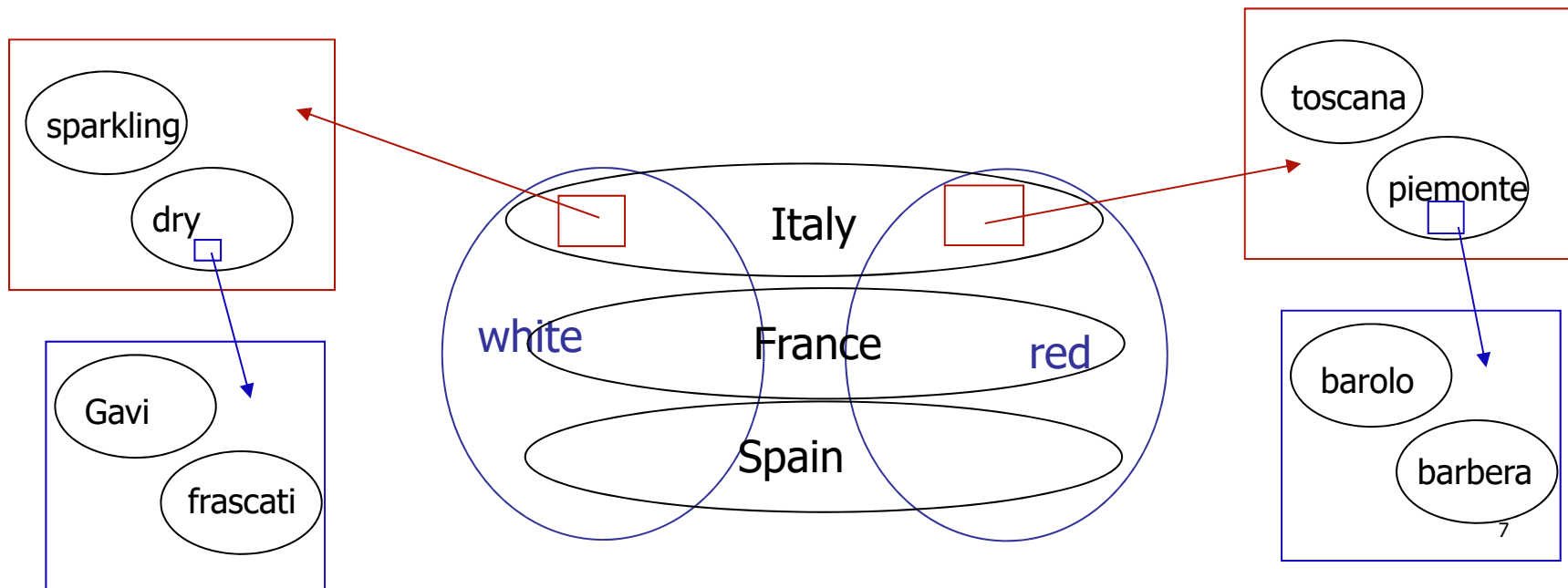
Ca	Na	Mg
K	NH4	HCO
SO4	Cl	NO3
F	SiO	

# Pattern analysis

- The pattern recognition is a method to have information on the sample described by patterns
- Mathematically, a pattern is a vector that corresponds to the feature describing some aspects of a sample
- The features are linked to the type of information you want to be described
- The last operation of the pattern recognition is to assign a sample to a class (membership class)

# Membership class

- The membership class is a theoretical set of elements sharing a global feature
- Items can be grouped according to different classification schemes depending on the global feature
- Example:
  - Wine classification

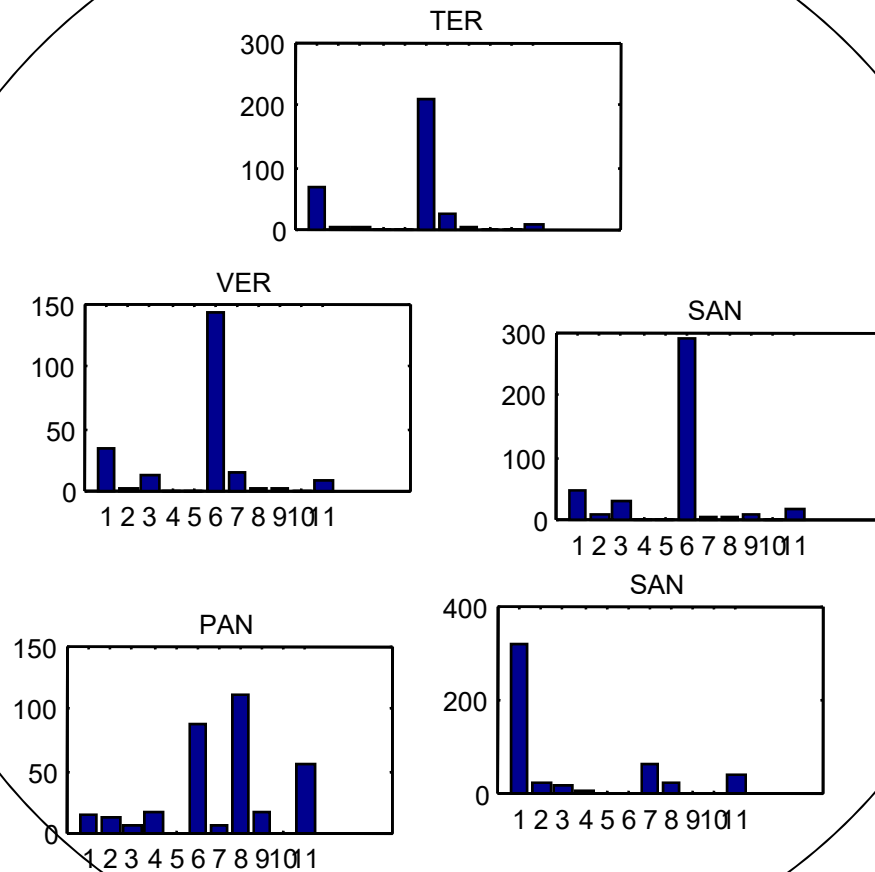


# Patterns and membership class

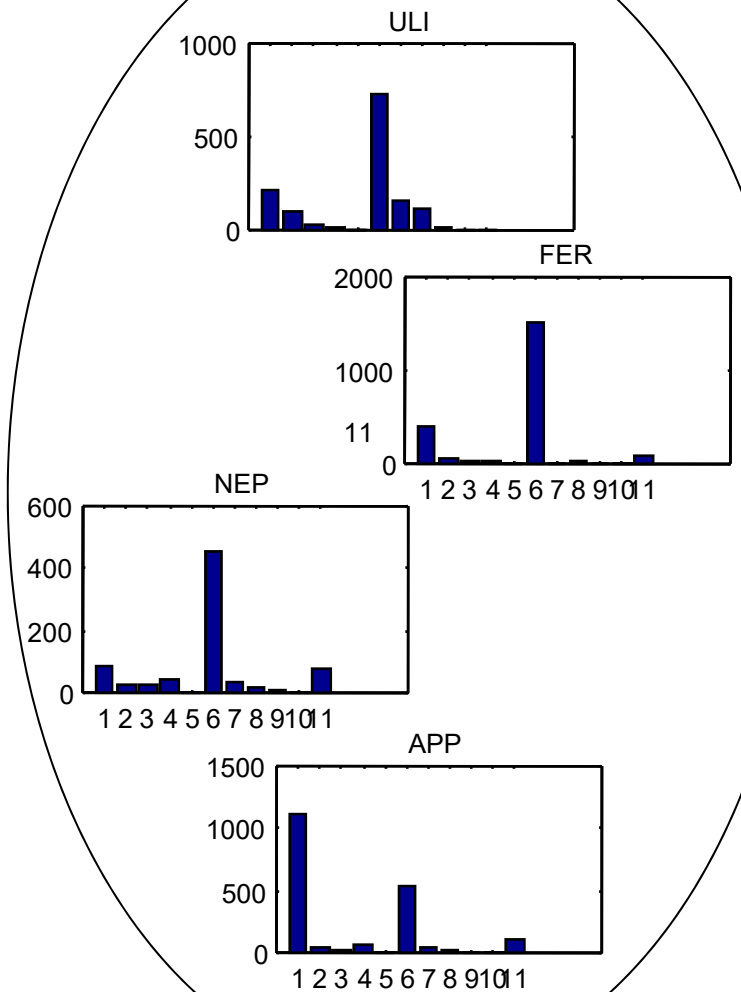
- If the features are appropriate for the classification then the elements belonging to a same class have similar patterns.
- The similarity between patterns can be detected with a visual inspection of suitable graph representation
- The simplest: column chart and radar-plot

# Column chart and membership class

## oligo minerals



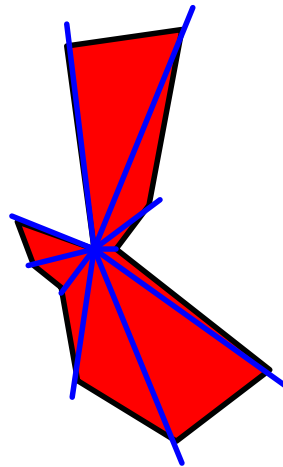
## minerals



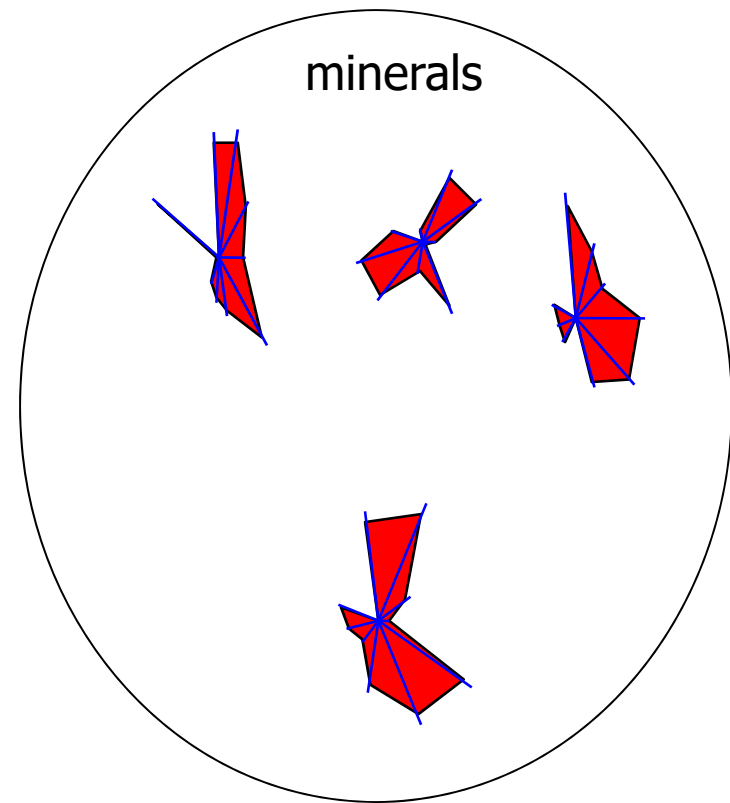
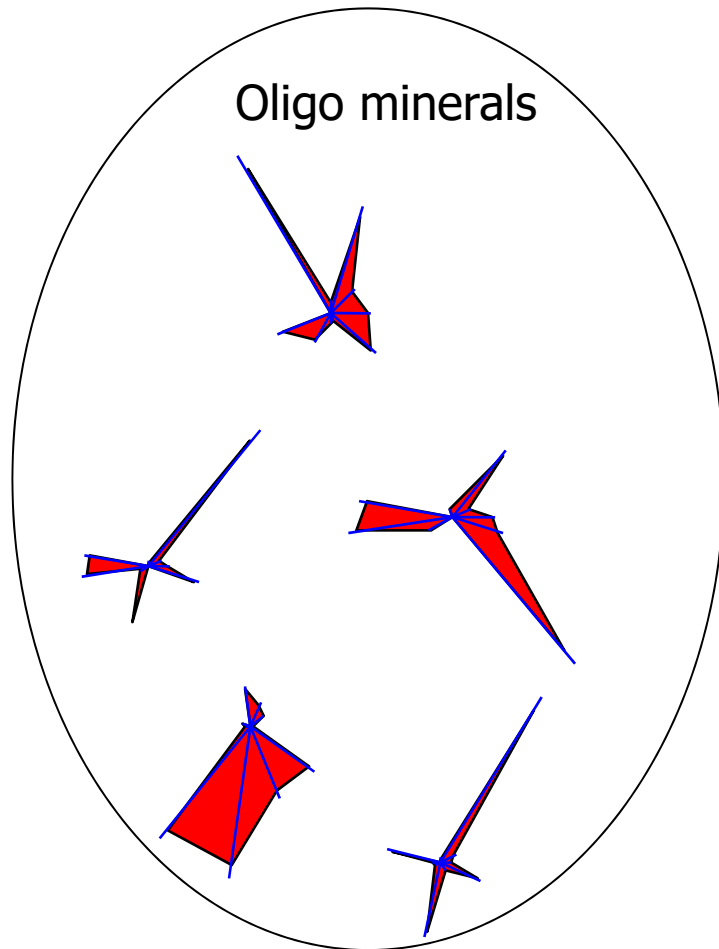


# Radar plot

- A simple way to visualize multidimensional patterns
- the pattern variables are the same of the directions (axes)
- The axes are organized to form a circle
- Along each axis is plotted the value of the variable
- Joining the points on the axes you get a figure that forms the "profile" of the pattern
- often used in sensory analysis to define the sensory profile of foods.



# Radar plot and class membership



# distance Criteria

- Each pattern is a point in N-dimensional space
- The space is defined by the features that describe the pattern
- For each feature is assigned an axis
- All axes define an Orto-normal basis
- "class-membership" – distance relation
- Two close points (patterns) probably belong to the same class two far points belong to different classes

# Classification criteria

- Criteria "unsupervised"
- Determining, on the basis of an a priori criterion, an internal classification scheme
- The criterion used is generally that of the distance
  
- Clusters of analysis
- hierarchical method used to form classes with more and more undefined.
- Exotic Methods
- Potential Method

# Cluster Analysis

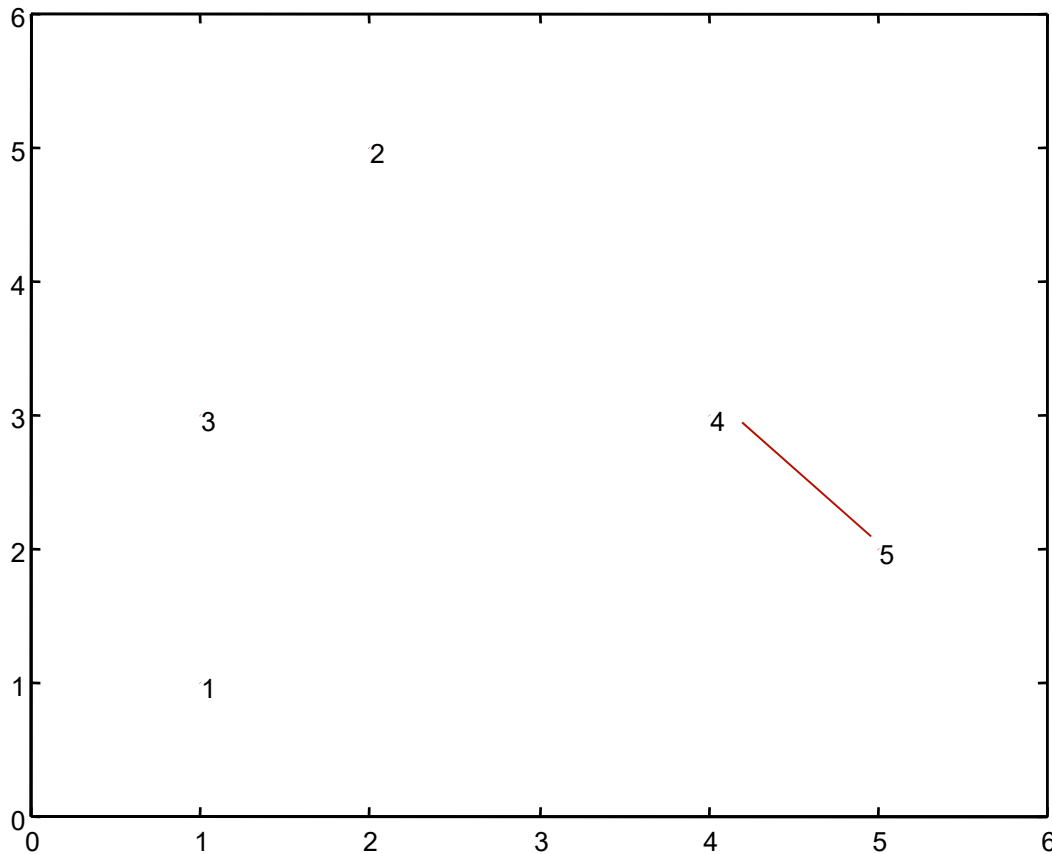
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

- Cluster analysis groups in hierarchy set of points
- Depending on the distance the points are grouped together to form more and more large groups. Eventually all data can be grouped together
- The basic instrument of the cluster analysis is the distance matrix
- Given a set of  $X_i$  pattern, the distances  $d_{ij}$  is the following matrix:

$$d_{ij} = \|X_i - X_j\|$$

- The matrix is symmetric

# Distance matrix



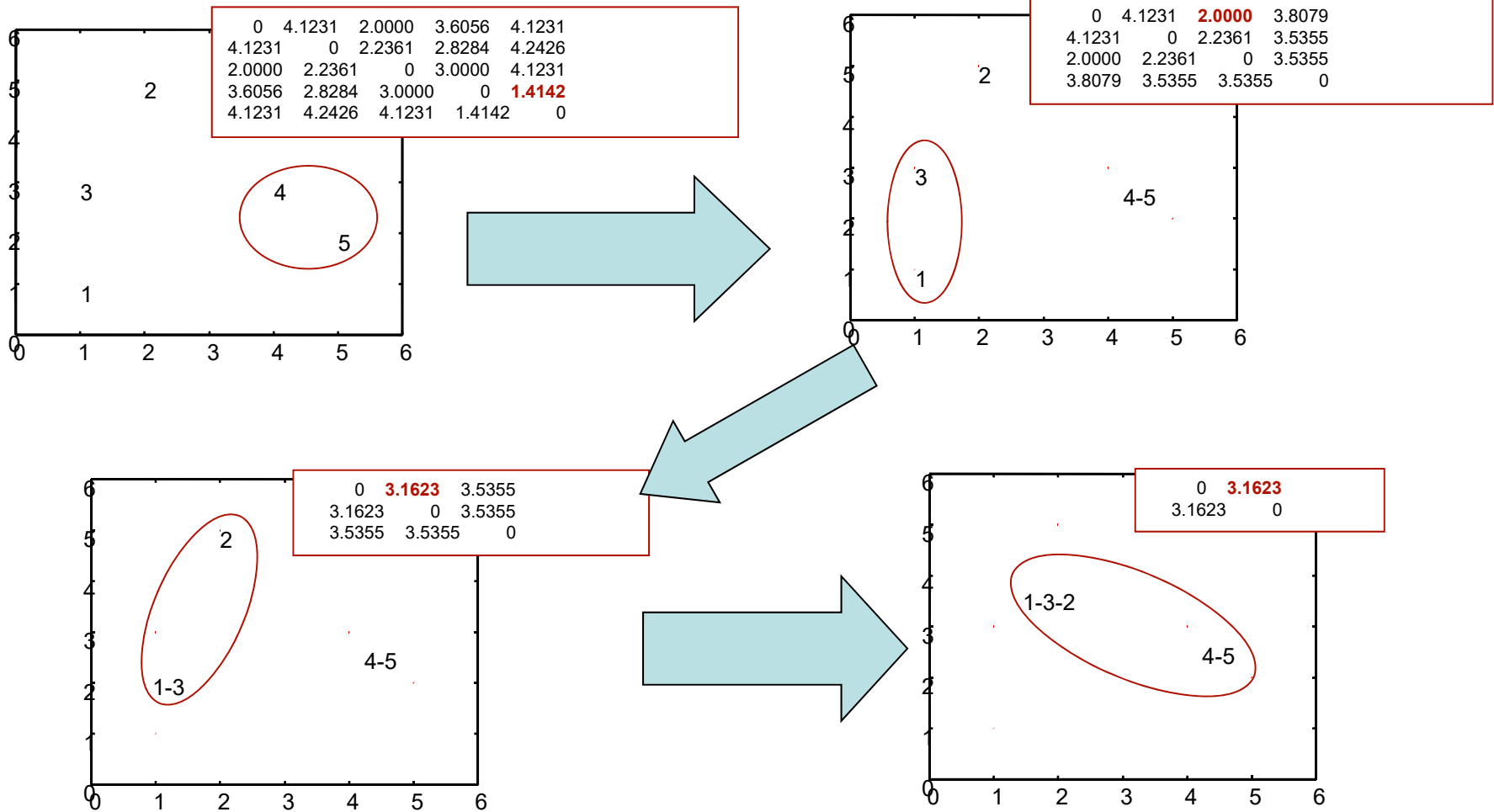
	1	2	3	4	5
1	0	4.1231	2.0000	3.6056	4.1231
2	4.1231	0	2.2361	2.8284	4.2426
3	2.0000	2.2361	0	3.0000	4.1231
4	3.6056	2.8284	3.0000	0	<b>1.4142</b>
5	4.1231	4.2426	4.1231	1.4142	0

- Points 4 and 5 are the closest
- The pair 4-5 is isolated from 1-2-3
- Probably there are two sets of data:
  - 1-2-3
  - 4-5
- The cluster analysis makes the analysis rational and it allows to operate on N-dimensional space

# Cluster analysis

- The hierarchical cluster analysis is an iterative process by making a dendrogram that defines the classes depending on the distance.
- Step i
- distance matrix calculation
- Detecting points with smaller distance
- Formation of clusters by combining Points
- Replacement of the cluster with the average point
- Refining the procedure until there is only one point

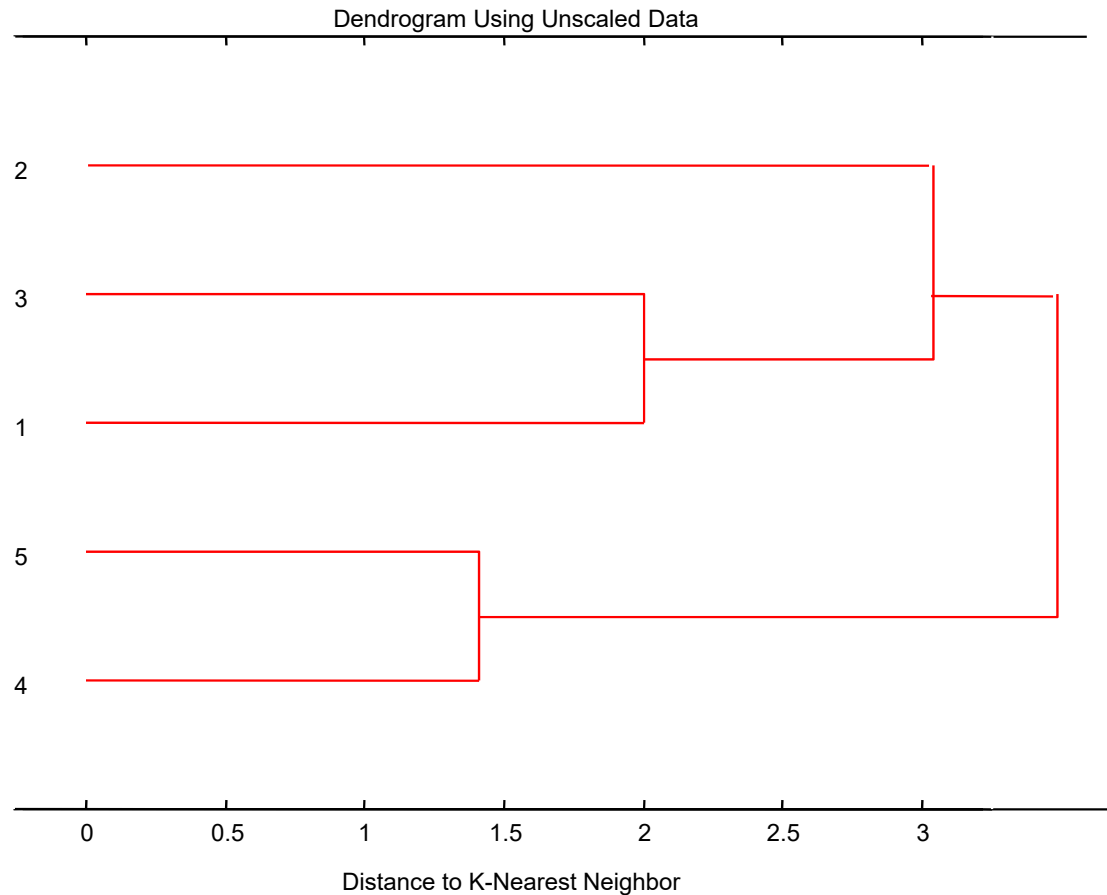
# Example



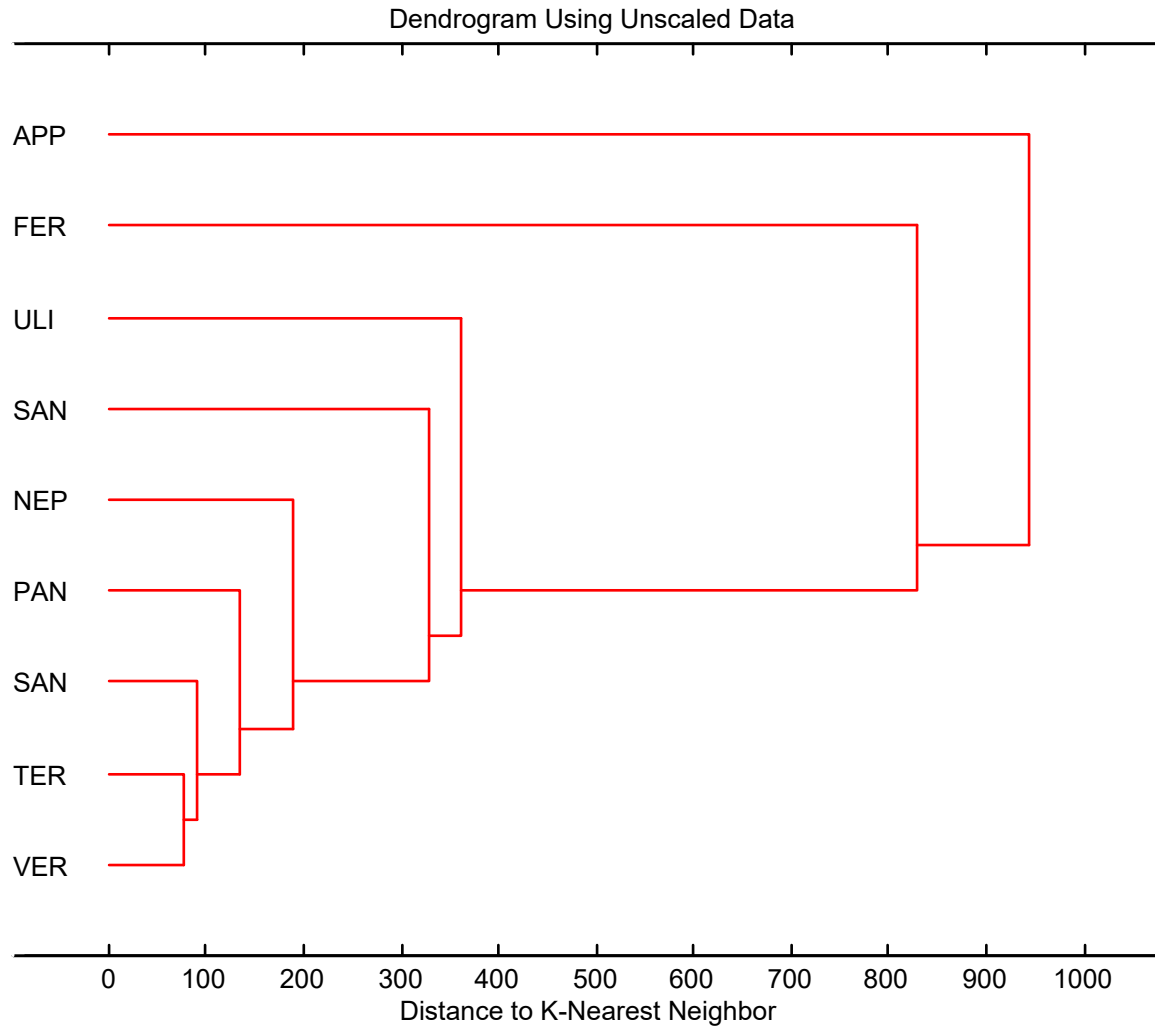


# Dendrogram

- The result of the cluster analysis is depicted in a tree diagram (dendrogram) where all the points are joined to form clusters, as a function of distance.

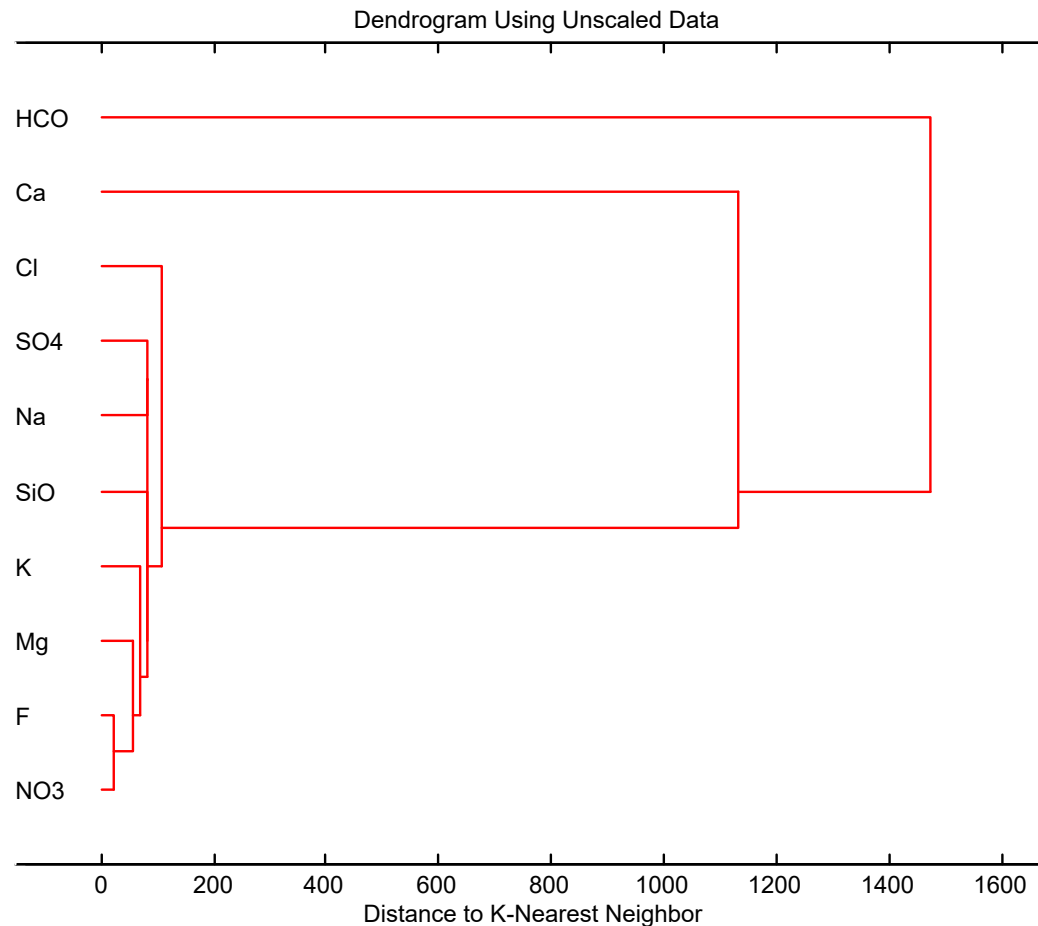


# Cluster analysis mineral waters

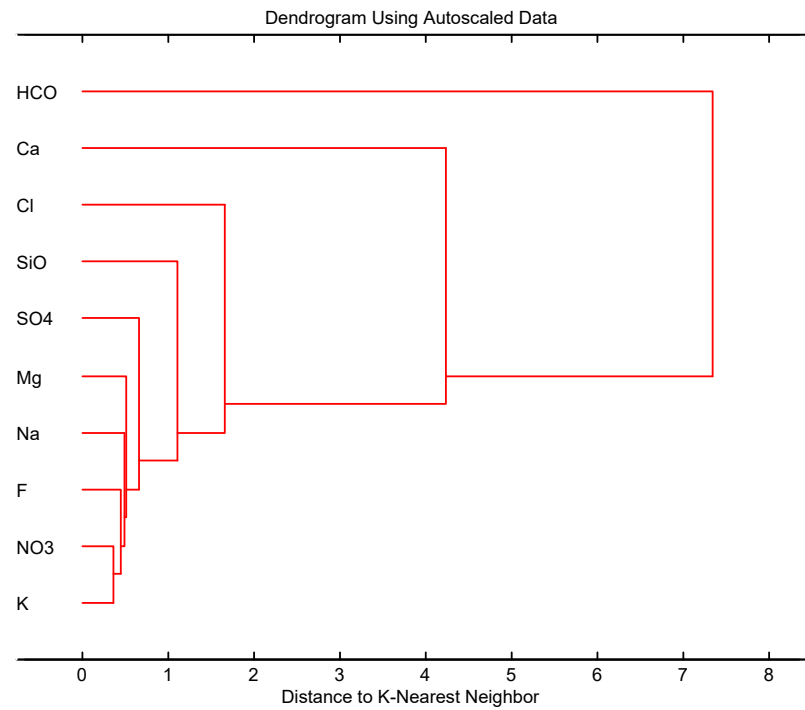
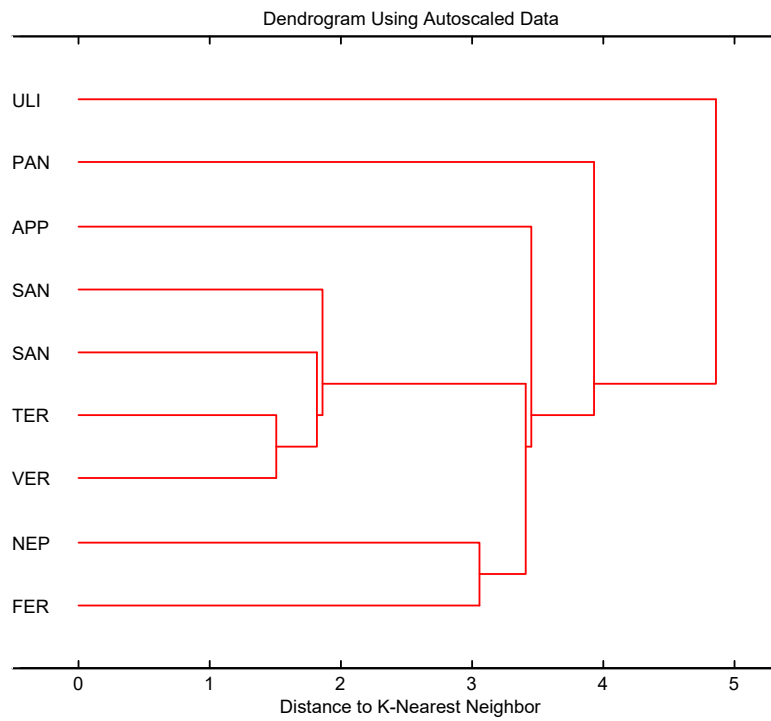


# "inverse" Pattern recognition

- To study the role of the features we can study the problem in a transposed manner where the features become samples and samples the features.
- Example: cluster analysis of mineral waters



# Normalization of the tree diagram for mineral waters

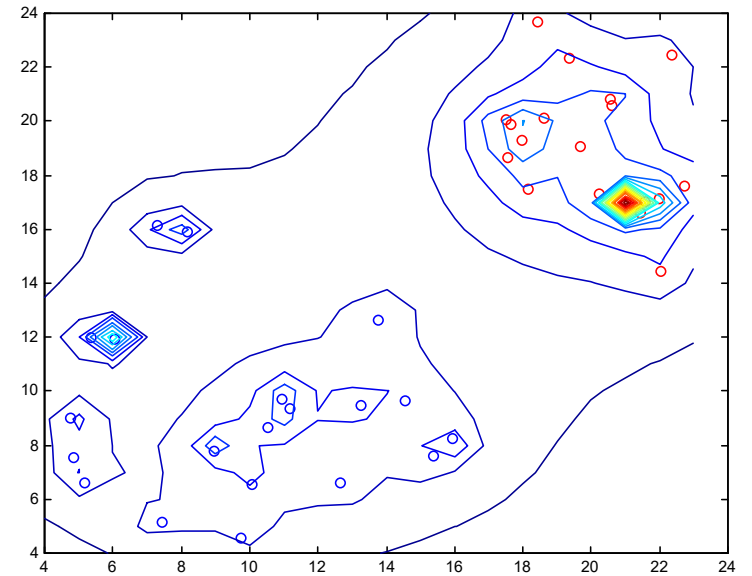
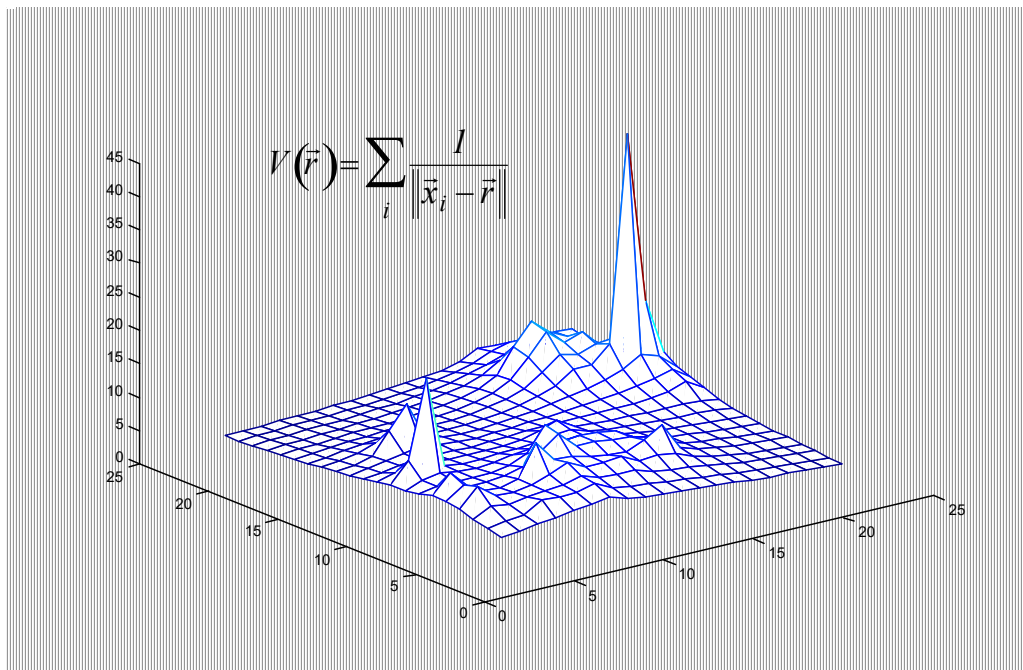
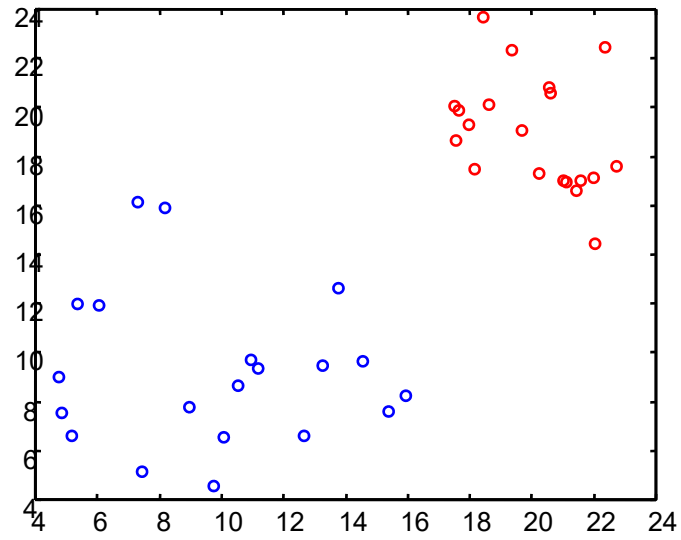


# Gravitational clustering

- An interesting "exotic" method of "unsupervised" classification is based on the analogy with the force fields.
- Suppose that each point possesses a mass  $M$  (equal for all). This mass will generate a potential  $V$  point in the space. Where The points are grouped (in a class) will generate a higher potential, and then studying the evolution of potential, and in particular Its top it is possible to identify the spatial regions of maximum densification (classes).
- The analogy with the masses is just an example you can use any potential function.
- If gravitational analogy,  $N$  data points  $x_i$  the potential at the point  $r$  is given by:

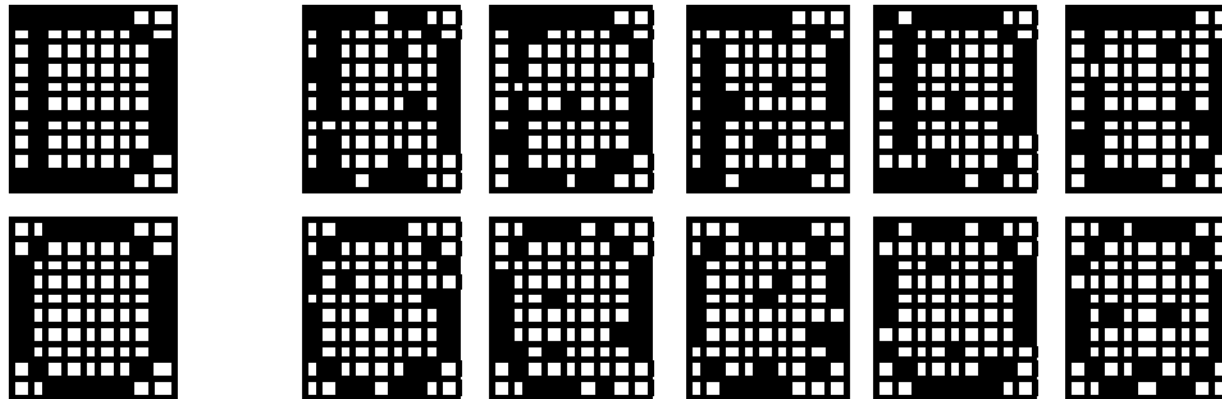
$$V(\vec{r}) = \sum_i \frac{1}{\|\vec{x}_i - \vec{r}\|}$$

# Gravitational clustering



# Template Matching

- This method is efficient when each class contains only a pattern. The measured patterns are affected by additive noise (no translation, rotation, deformation of pattern)
- Typical example: Image Recognition
- For each class, the pattern with low noise is taken as the class template
- Two methods to assign a pattern to a class:
  - Count the number of agreements: maximum correlation
  - Count the number of disagreements: lowest error



templates

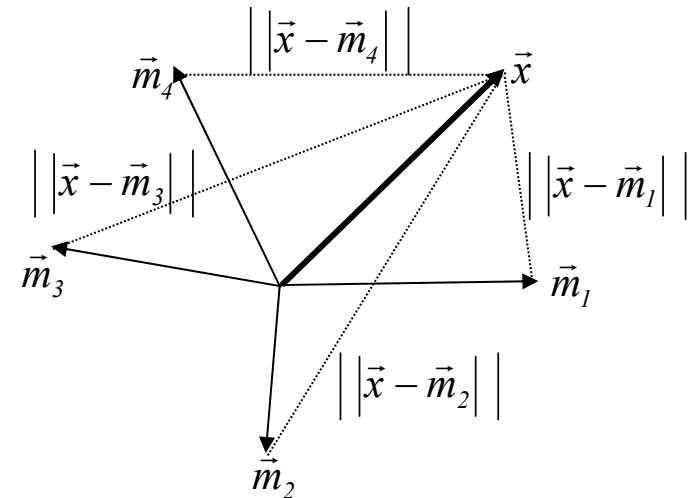
Measured Patterns

# K-nearest neighbour

- mathematical calculation of template matching
- The templates are represented by vectors  $\mathbf{m}_j$
- For class  $\mathbf{K}$ , the distance between a pattern ( $\mathbf{x}$ ) and the corresponding template ( $\mathbf{m}$ ) is :

$$\varepsilon_k = \left\| \vec{x} - \vec{m}_k \right\|$$

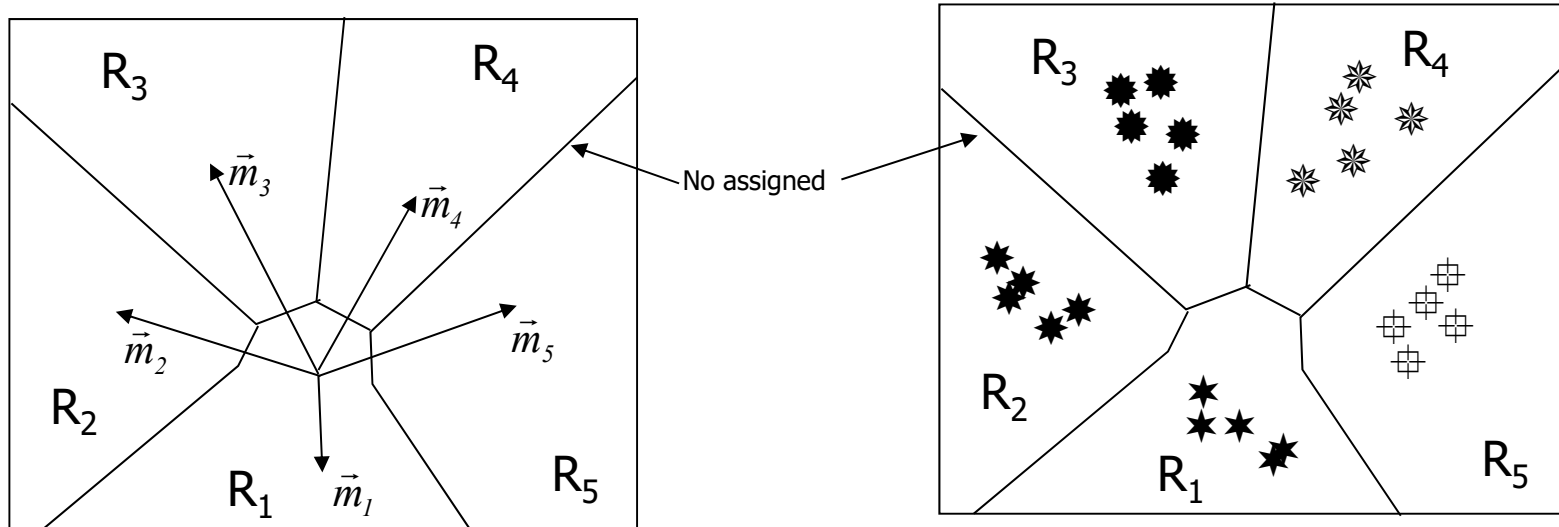
- The pattern  $\mathbf{x}$  is assigned to the class to which the distance from the corresponding template is minimum
- The transaction requires the definition of a measurement system, a rule for the distances calculation





# Cluster analysis contours

- The discriminant functions divides the space into class regions
- The contours are points at the same distance from two or more template
- The Linear functions produce polygonal contours.

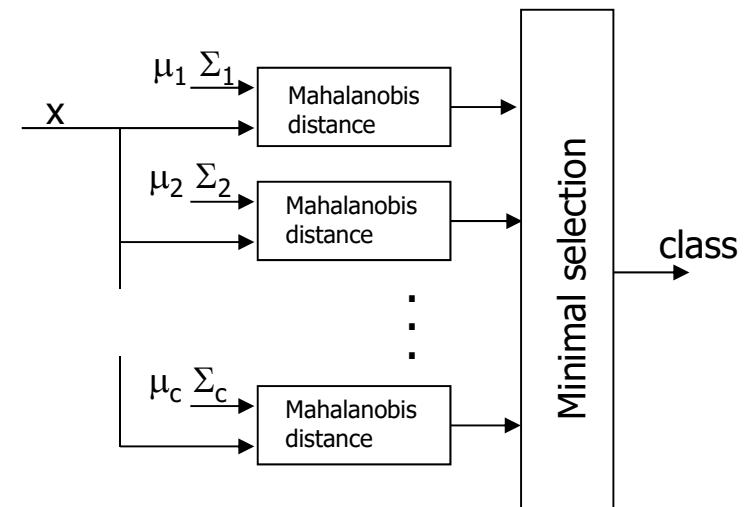


# Mahalanobis distance

- For a Gaussian distribution, the iso-probability points are given by the following quadratic form:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- The probability value is called the Mahalanobis distance, or statistical distance,
- the probability that a vector  $\mathbf{x}$  belonging to a Gaussian distribution is defined by  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$
- By evaluating the Mahalanobis distance of a pattern is possible to assign a pattern to each class to which the Mahalanobis distance is smaller, that is, towards which the probability is greater



# Nonparametric classification

- If the probability distribution of the patterns is not known or is not Gaussian, you should use a non-parametric statistical classification that is independent of the classes and in which you search for a combination of variables that allows the identification of classes.
- This operation is similar to the discriminant analysis
- The calculation is similar to multiple linear regression, using the linear relation between the matrix X (set of patterns of samples measured) and a Y matrix (numerical coding of class membership)
- to solve the problem it is necessary to find the regression matrix B

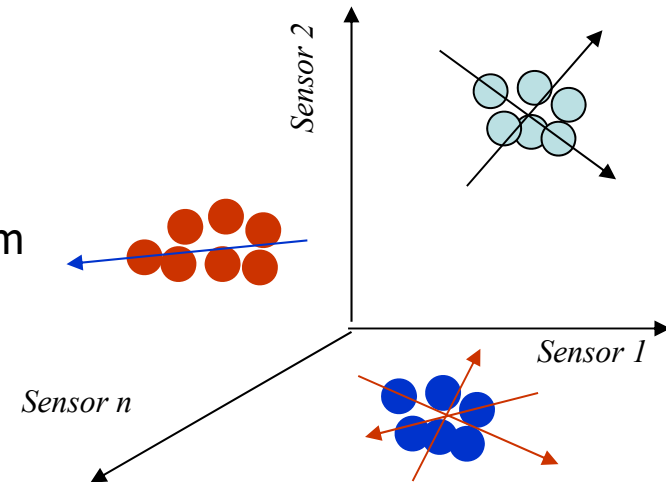
$$Y = X \cdot B^T \Rightarrow B^T = \left( X^T \cdot X \right)^{-1} \cdot X^T \cdot Y$$

# “one-of-many”

- The Y-matrix is constructed with a number of columns equal to the number of classes in the problem  
Each Y column then identifies a class  
Given a pattern  $X_i$  the corresponding  $Y_i$  line is made by setting to zero all the elements except the one corresponding to the class of  $X_i$  which is set to 1  
When identifying the regression model, we will have a finite accuracy, so the pattern is assigned to the class whose corresponding value is larger.

# Soft Independent Modeling of Class Analogy (SIMCA)

- For each class of samples, a PCA model is constructed. The PC bases are different for each class and the number of meaningful components is also different
- Each class defines a proper hypervolume
- Unknown samples are identified applying them to each model and looking for the matching one.
- A probability of membership is obtained.



# A case of neural network paradigm

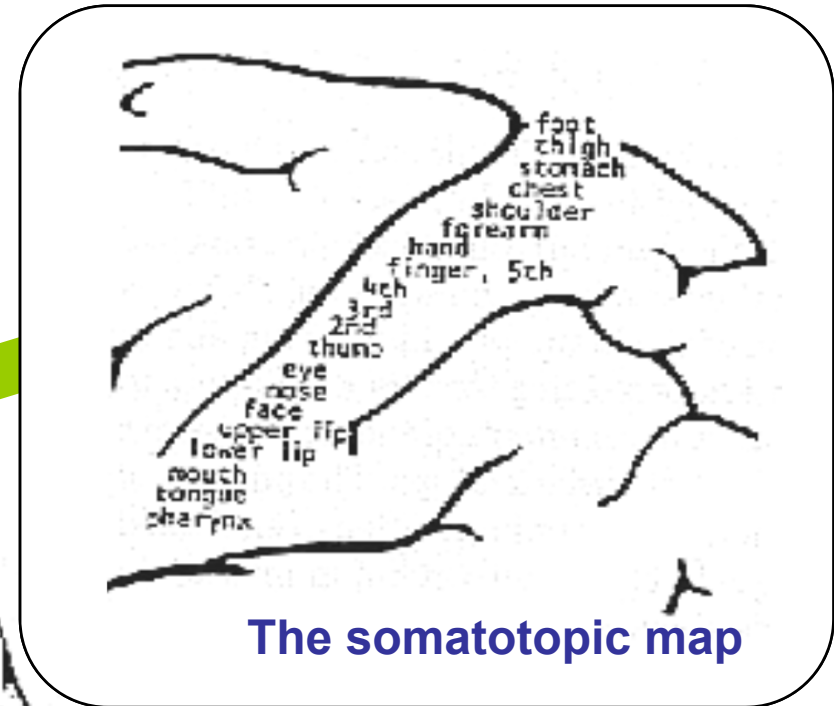
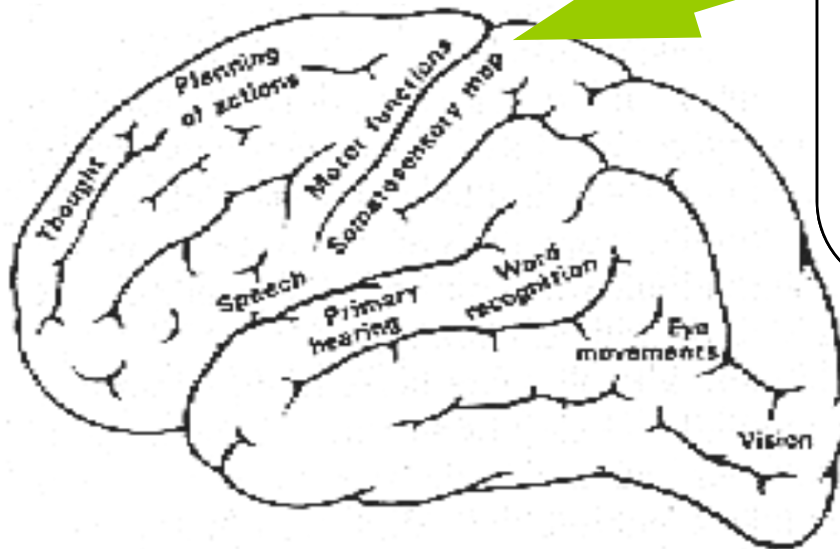
Self Organizing Map  
Linear Vector Quantization

# Self Organizing Map

- it is a neural network based on strong biological similarities . It aims at mimicking the functionality of the cerebral cortex
  - sensorial map
- Principal features:
  - it learns from the experience; unsupervised; adaptive
- It provides a powerful tool to map a phenomena (represented as a multidimensional system) into a bidimensional grid discovering any intrinsic classification property
- The map is formed by a bidimensional grid of neurons (discrete space)
- Each neuron is identified with a codebook vector belonging to the sensor space and representing the link between the SOM and the input space
- Learning algorithm (*Kohonen algorithm*) is structured in two step
  - Response
  - Adaptation

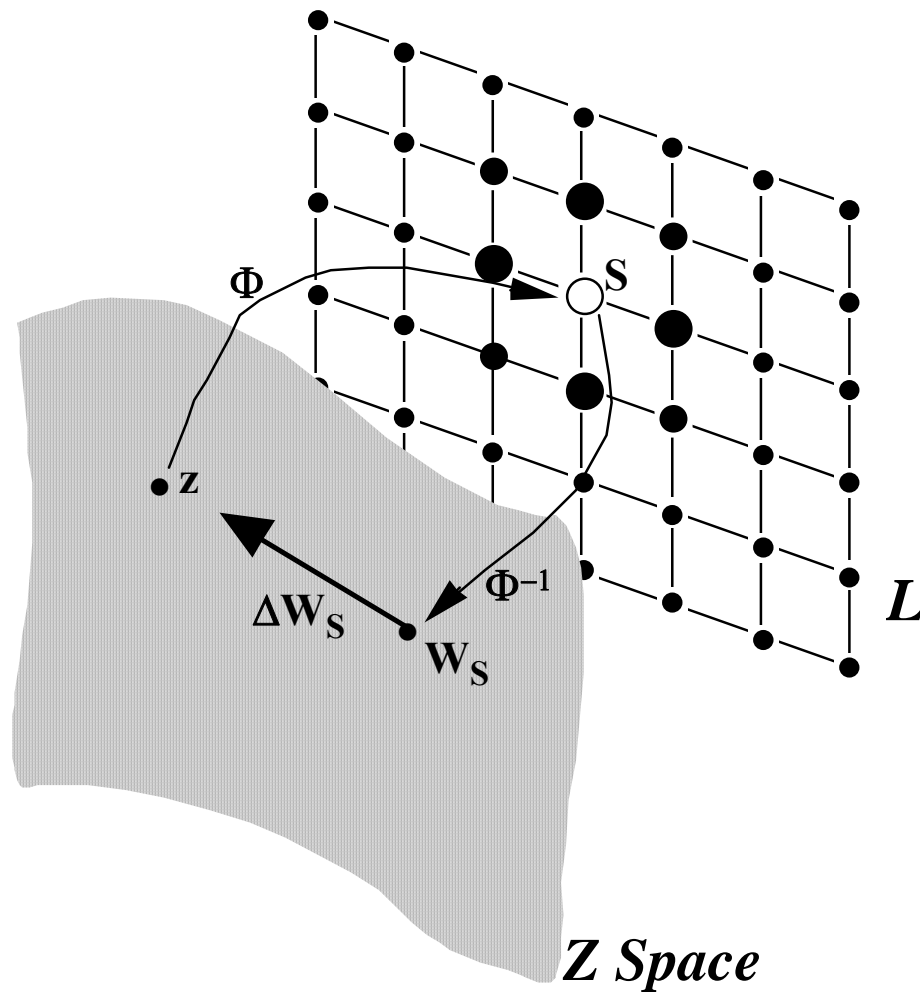
# SOM: the Biological Paradigm

Brain areas





# SOM: the Learning Algorithm



## 1. Response

$$\|z - w_s\| \leq \|z - w_r\|$$

## 2. Adaptation

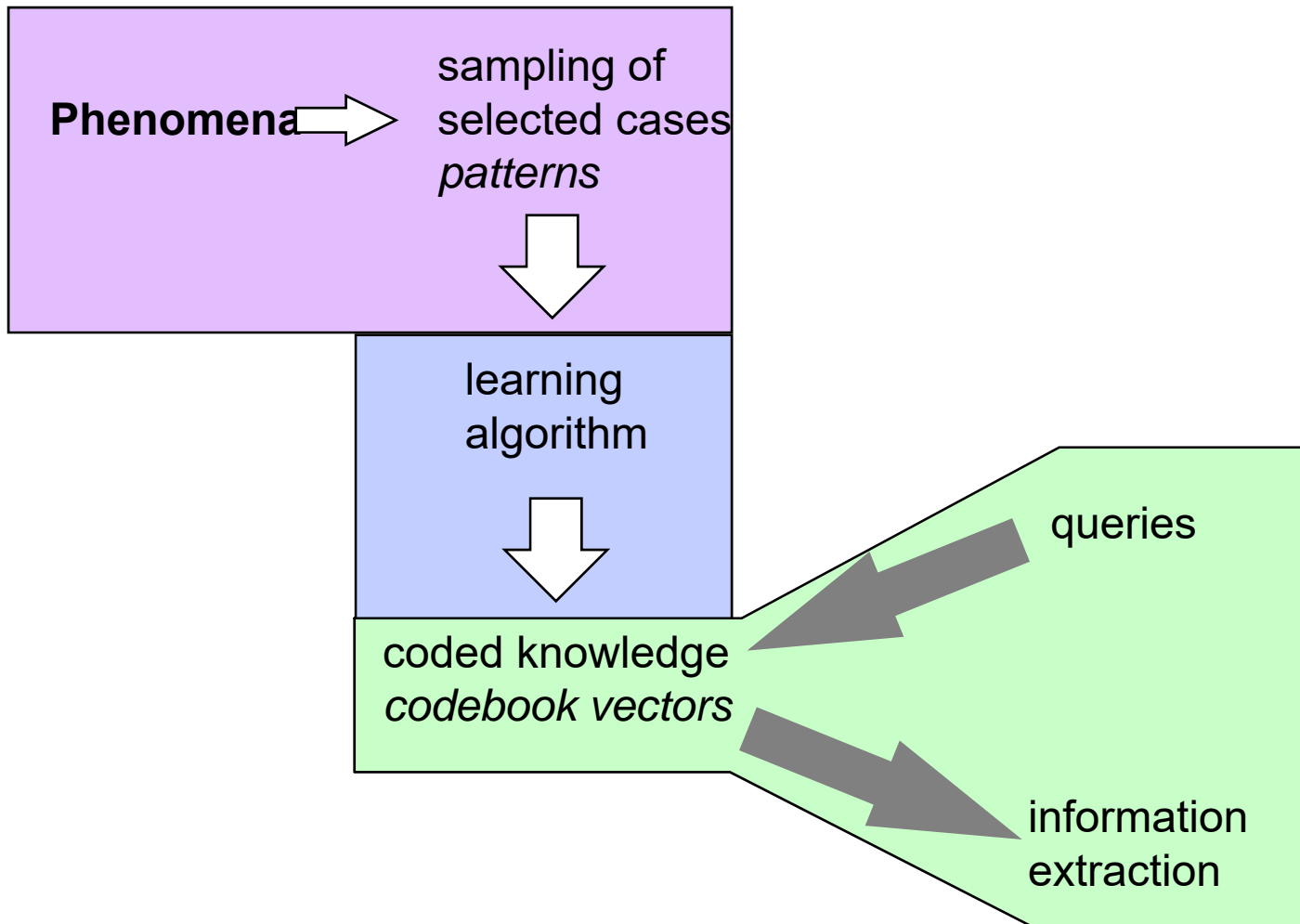
$$w_r^{new} = w_r^{old} + \alpha \cdot h_{rs} (z - w_r^{old})$$

**h** is the neighbour function defining the extension of SOM which participate to the adaptation process

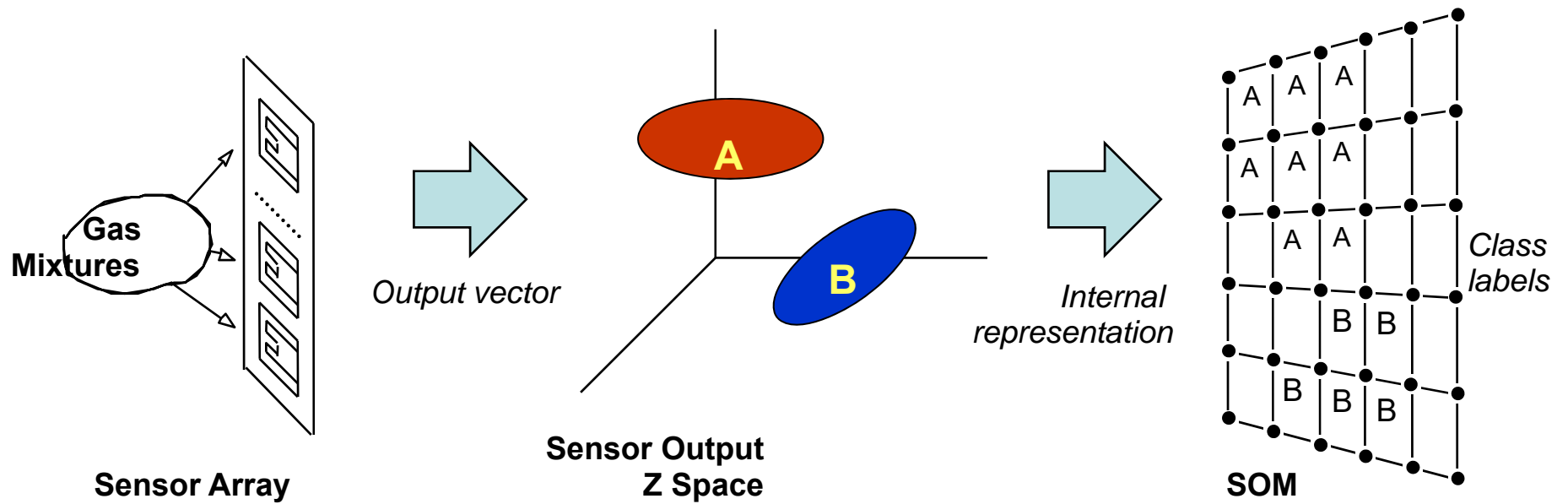
$$h_{rs} = \exp\left(\frac{-\|r-s\|^2}{2\sigma^2}\right)$$

$\alpha$  is a learning rate

# SOM: Flow of Data

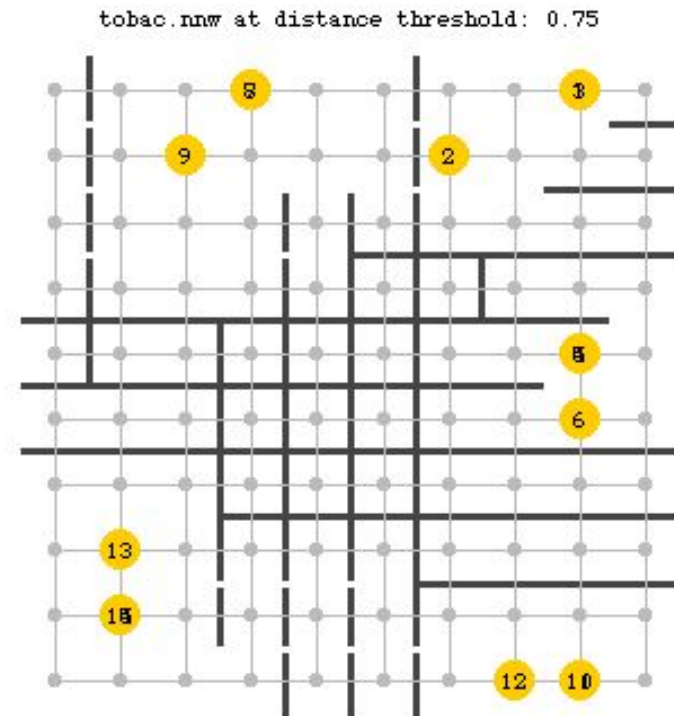
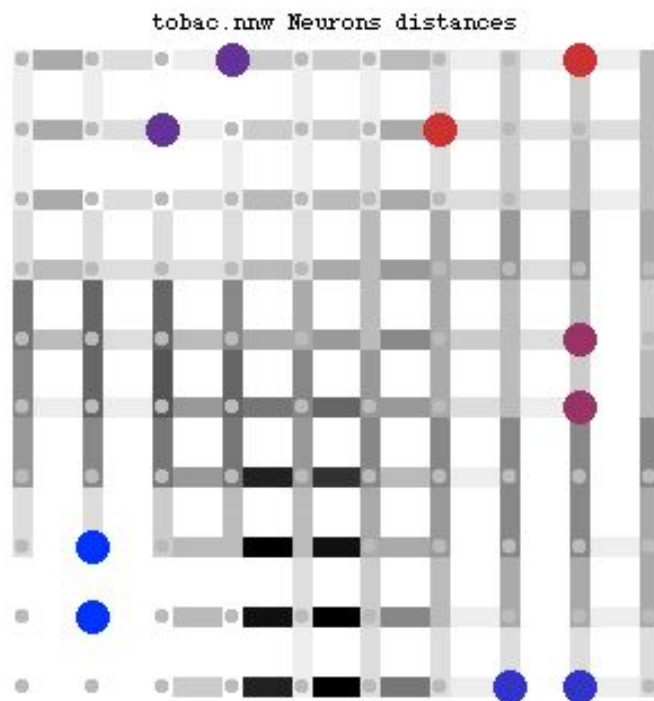


# SOM and Sensor Array



# SOM: Representation of Clustering

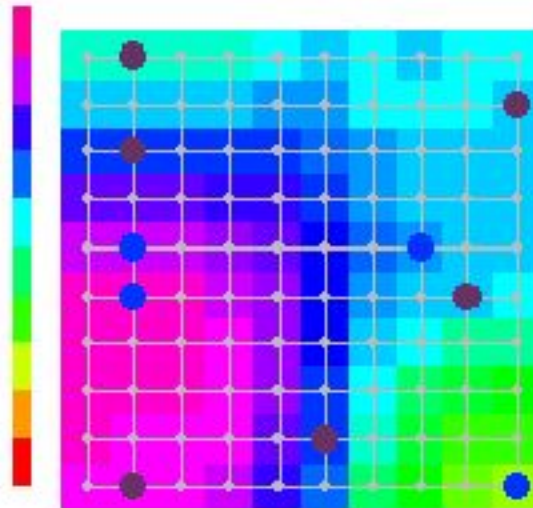
- Distance between neurons can be represented drawing lines, connecting adjacent neurons, in a color-scale proportional to the distance between the codebook vectors. Clusters can be formed fixing a threshold value to the distance.



# SOM: Component Planes

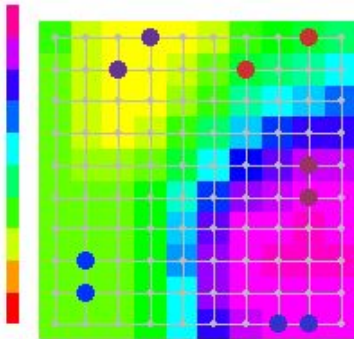
- The components of the codebook vectors are related to the sensors composing the array. These components can be plotted onto the SOM grid giving information about the behaviour of single sensors.

studlin.nnw Comp. Plane sensor:4

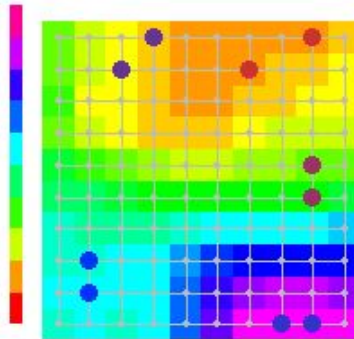


# SOM: Component Planes

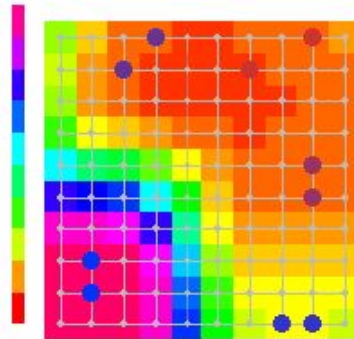
tobac.nnw Comp. Plane sensor:1



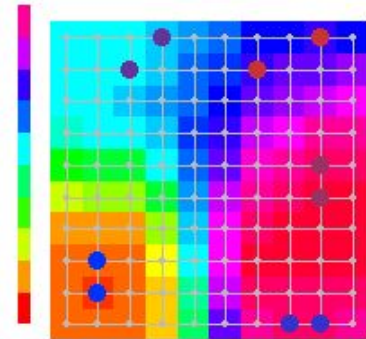
tobac.nnw Comp. Plane sensor:2



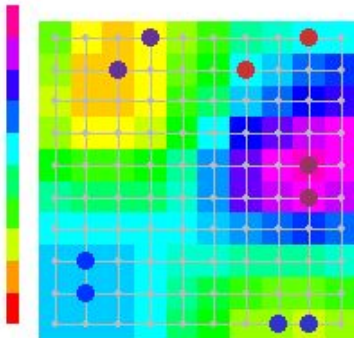
tobac.nnw Comp. Plane sensor:3



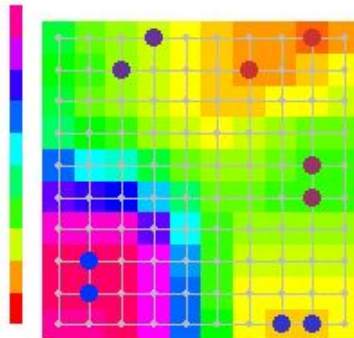
tobac.nnw Comp. Plane sensor:4



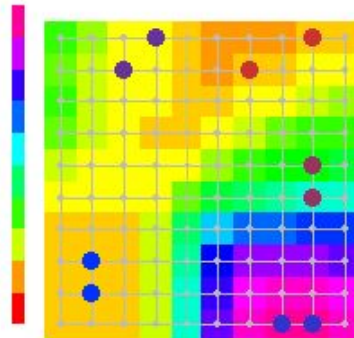
tobac.nnw Comp. Plane sensor:5



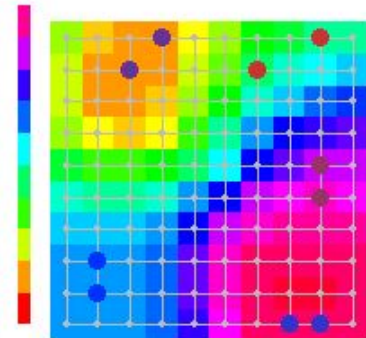
tobac.nnw Comp. Plane sensor:6



tobac.nnw Comp. Plane sensor:7



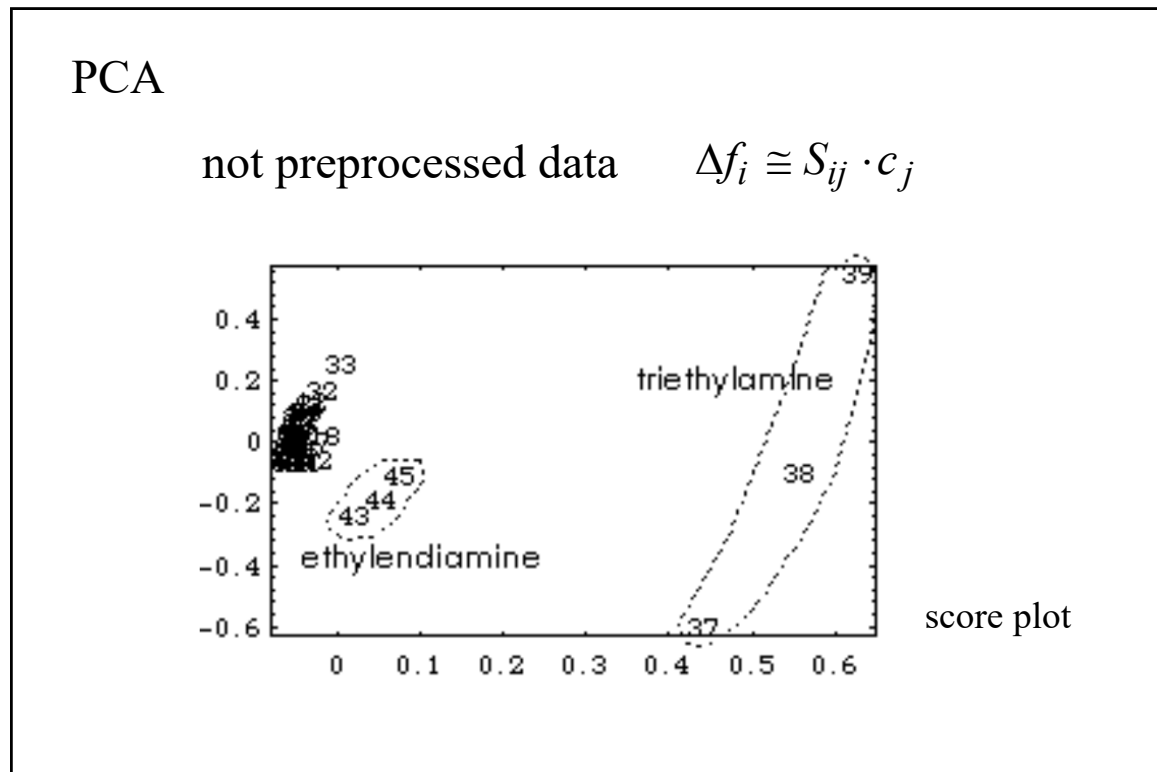
tobac.nnw Comp. Plane sensor:8



# PCA - SOM comparison:

## Array of QMB for the detection of VOC

Sensor responses have been measured, at different concentrations, for a number of different volatile compounds chosen as representatives of the following classes: alkanes, aldehydes, alcohols, aromatics and amines.

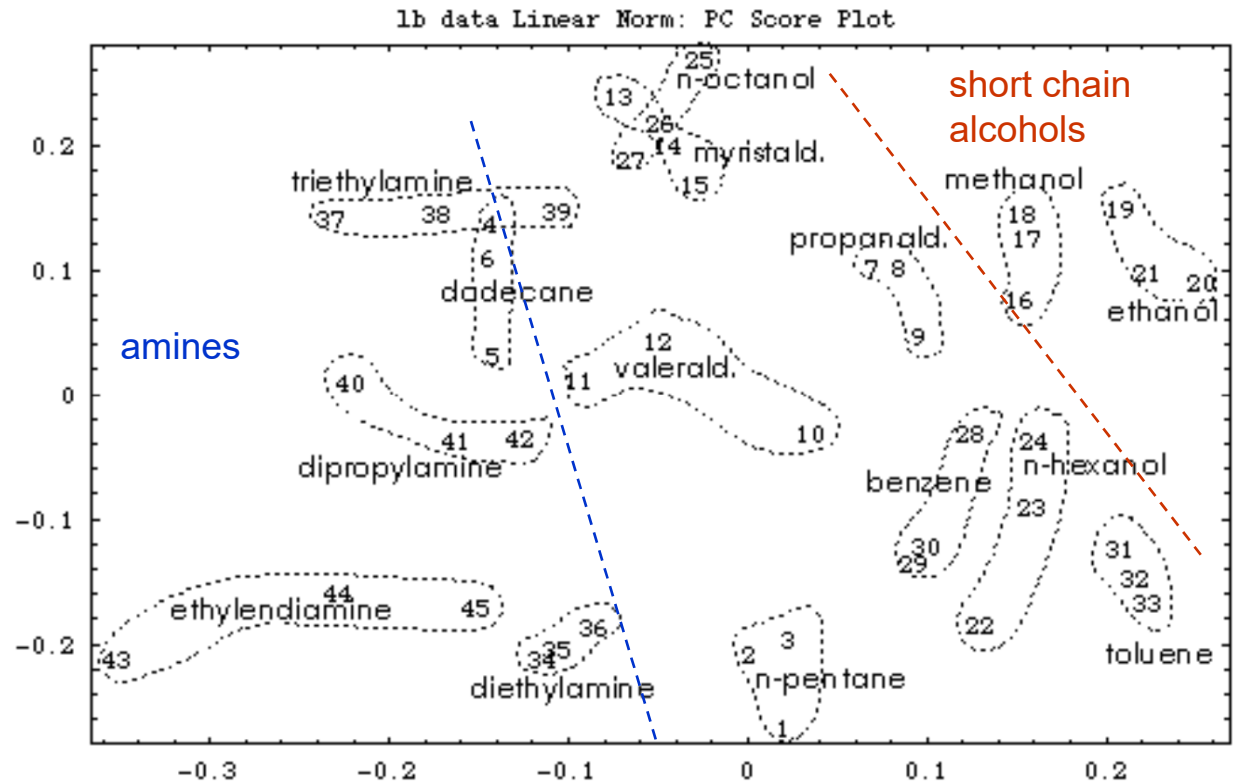


# PCA - SOM comparison:

Array of QMB for the detection of VOC

## Linear Normalization

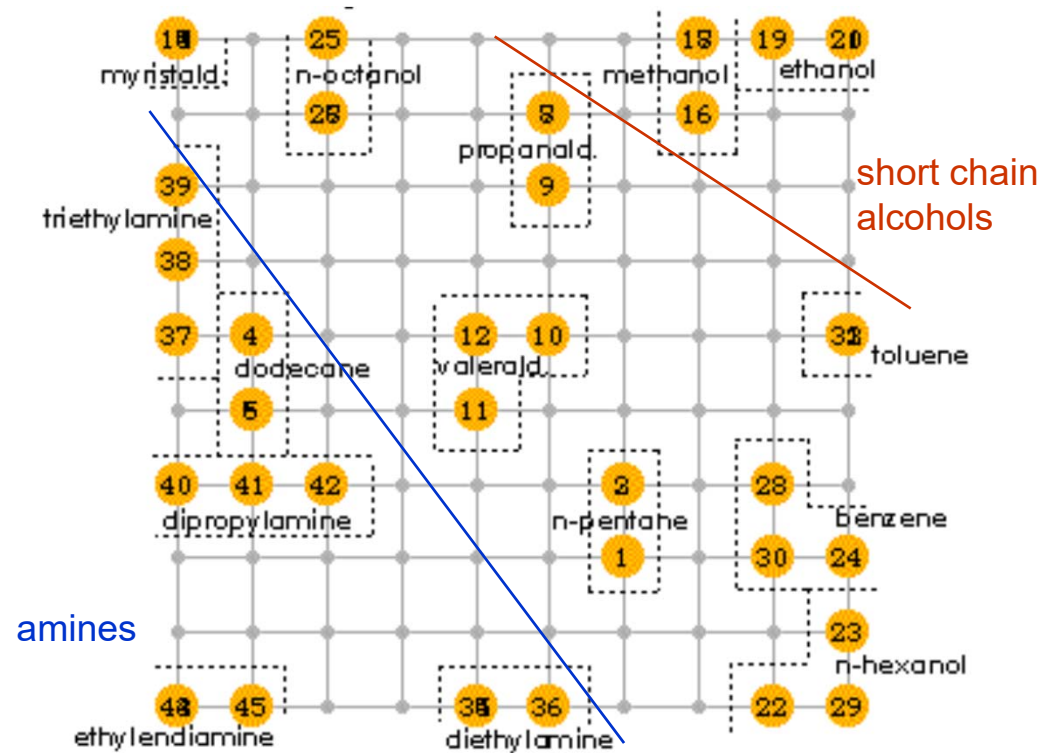
$$\Delta f_i \Rightarrow \frac{\Delta f_i}{\sum_k f_k} \cong \frac{S_{ij} \cdot c_j}{\sum_k S_{kj} \cdot c_j} = \frac{S_{ij}}{\sum_k S_{kj}}$$



PCA score plot



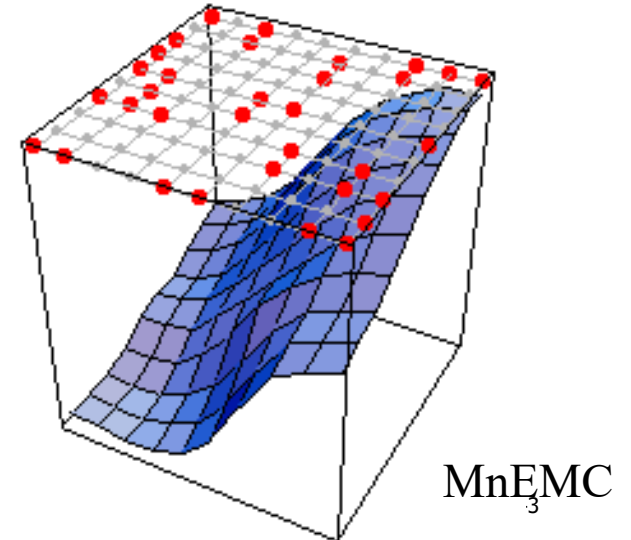
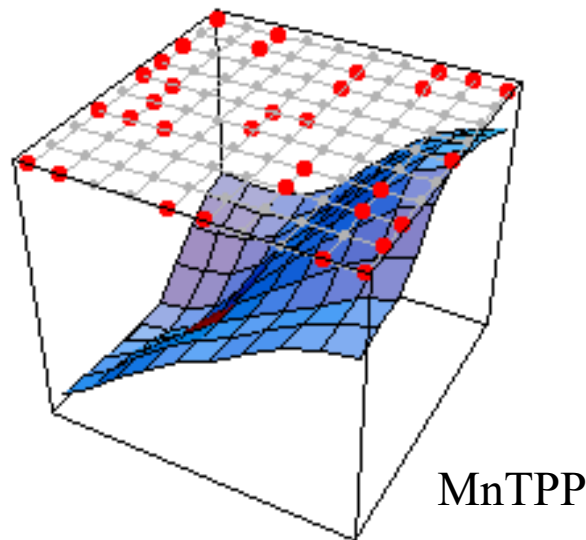
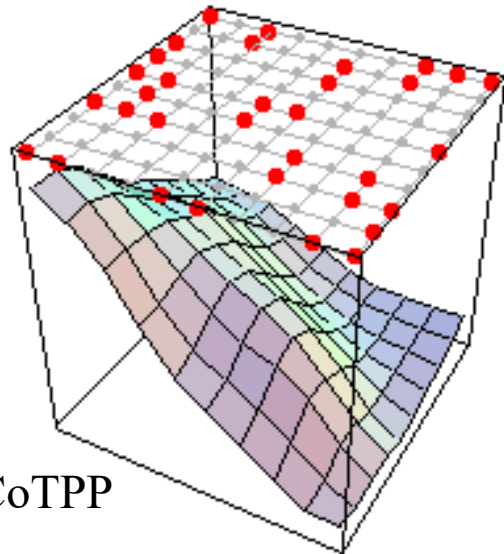
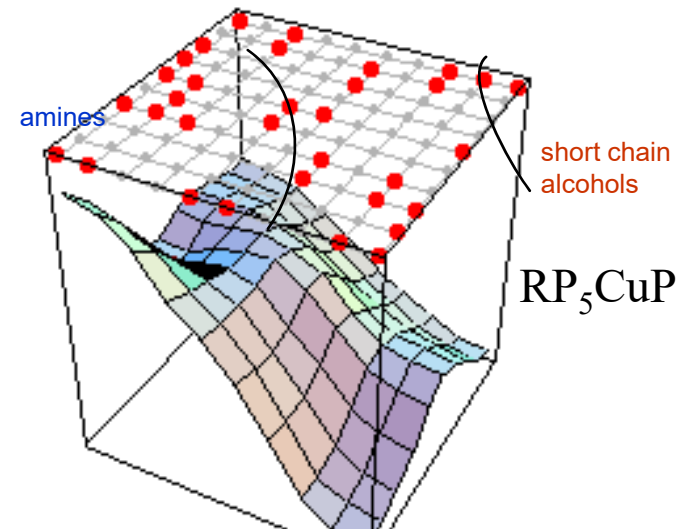
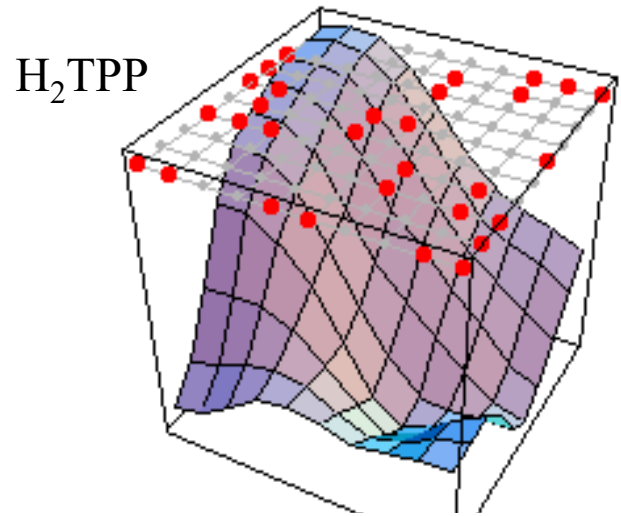
# PCA - SOM comparison: Array of QMB for the detection of VOC



# PCA - SOM comparison:

Array of QMB for the detection of VOC

- **Study of the Component Planes**



# Nonlinear classification methods: Learning Vector Quantization

- Vector Quantization: an approximation of the probability density functions of vectorial variables by finite sets of codebook vectors.
- Basic LVQ algorithm:

$m_i$  codebook vectors assigned to each class

an input  $\mathbf{x}$  is assigned to the class to which the closest  $m_i$  belongs

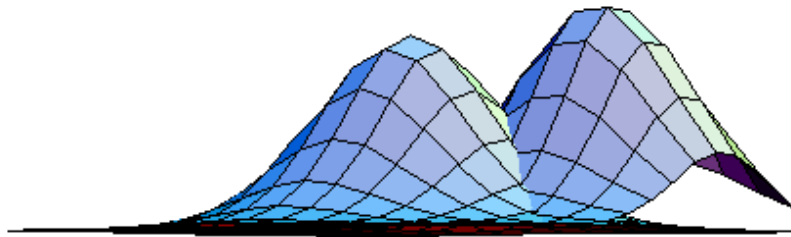
$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)] \quad \text{If } x \text{ belongs to the class of } m_c$$

$$m_c(t+1) = m_c(t) - \alpha(t)[x(t) - m_c(t)] \quad \text{If } x \text{ does not belong to the class of } m_c$$

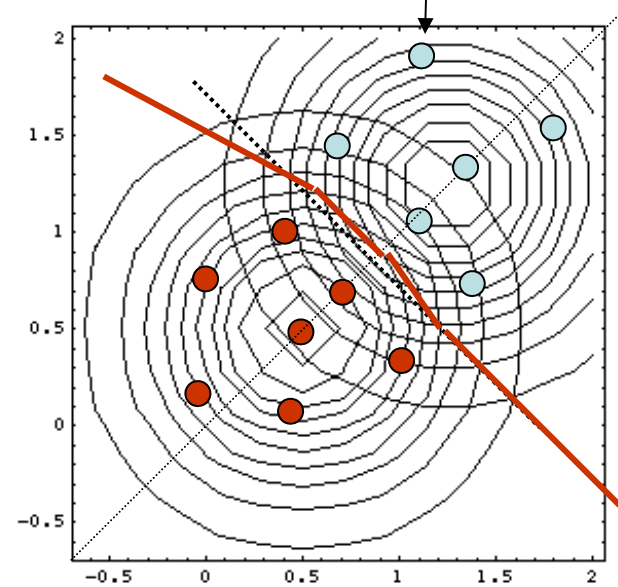
$$m_i(t+1) = m_i(t) \quad \text{If } i \neq c$$

# Nonlinear classification methods: Learning Vector Quantization

Example with two classes with 2D vectors



*Codebook vectors*



Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali  
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY  
I YEAR

Course:  
**EXPERIMENTAL DESIGN AND  
CHEMOMETRICS IN FOOD**  
(5 credits – 38 hours)

Teacher: Marcello Mascini  
([mmascini@unite.it](mailto:mmascini@unite.it))

The Teacher is available to answer questions at the  
end of the lesson, or on request by mail

# The course is split in 4 units

## UNIT 1: Univariate analysis

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the  $\chi^2$  method, Validation of the model

## UNIT 2: Multivariate analysis

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

## UNIT 3: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

## UNIT 4: Elements of Pattern recognition

cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

## **UNIT 3: Design of Experiments**

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs  
(Plackett–Burman designs, D-optimal designs,  
Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields  
of food science

## Factors

Silver laydown,  
Finish time...

Time,  
Catalyst...

Transport speed,  
Capture lens...

**Film  
Building**

**Chemical  
Process**

**Digital  
Imaging**

## Responses

Speed,  
Contrast

Yield,  
Purity

Image resolution,  
Banding



**Factors**

**Responses**

Compensation plan,  
Sales training



Sales revenue,  
Volume of new sales

Method of shipping,  
Order entry method



Shipping cost,  
Inventory level

Product positioning,  
Price



Trial purchase,  
Share of market

# Topics

- Review of Error Analysis
- Theory & Experimentation in Engineering
- Some Considerations in Planning Experiments
- Review of Statistical formulas and theory
- Begin Statistical Design of experiments (“DOE” or “DOX”)

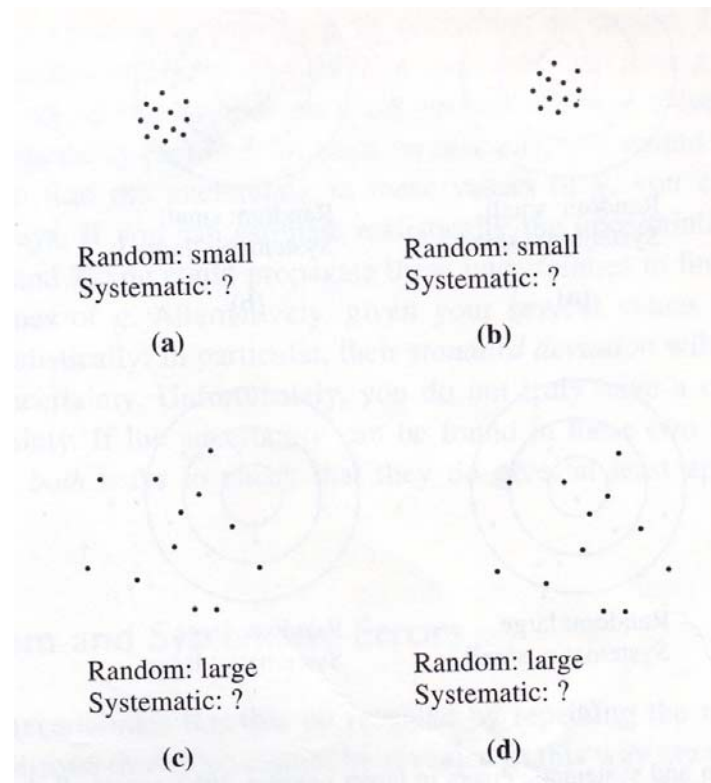
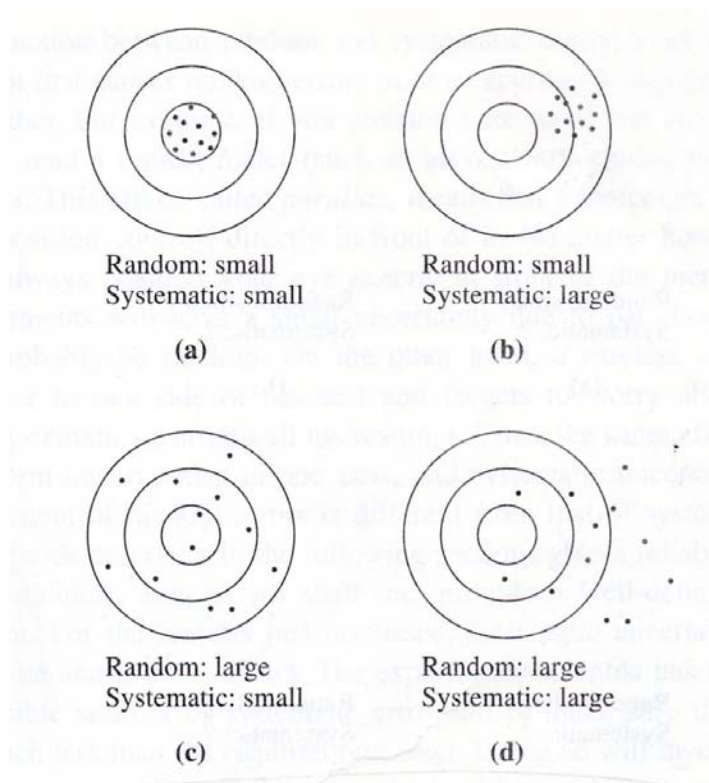
# Review of Error Analysis

- Uncertainty or “random error” is inherent in all measurements
  - Statistical basis
  - Unavoidable- seek to *estimate* and take into account
  - Can minimize with better instruments, measurement techniques, etc.

# Review of Error Analysis

- *Systematic* errors (or “method errors”) are mistakes in assumptions, techniques etc. that lead to non-random bias
  - Careful experimental planning and execution can minimize
  - Difficult to characterize; can only look at evidence after the fact, troubleshoot process to find source and eliminate

# Graphical Description of Random and Systematic Error



Why do we need to estimate uncertainty and include in stated experimental values?

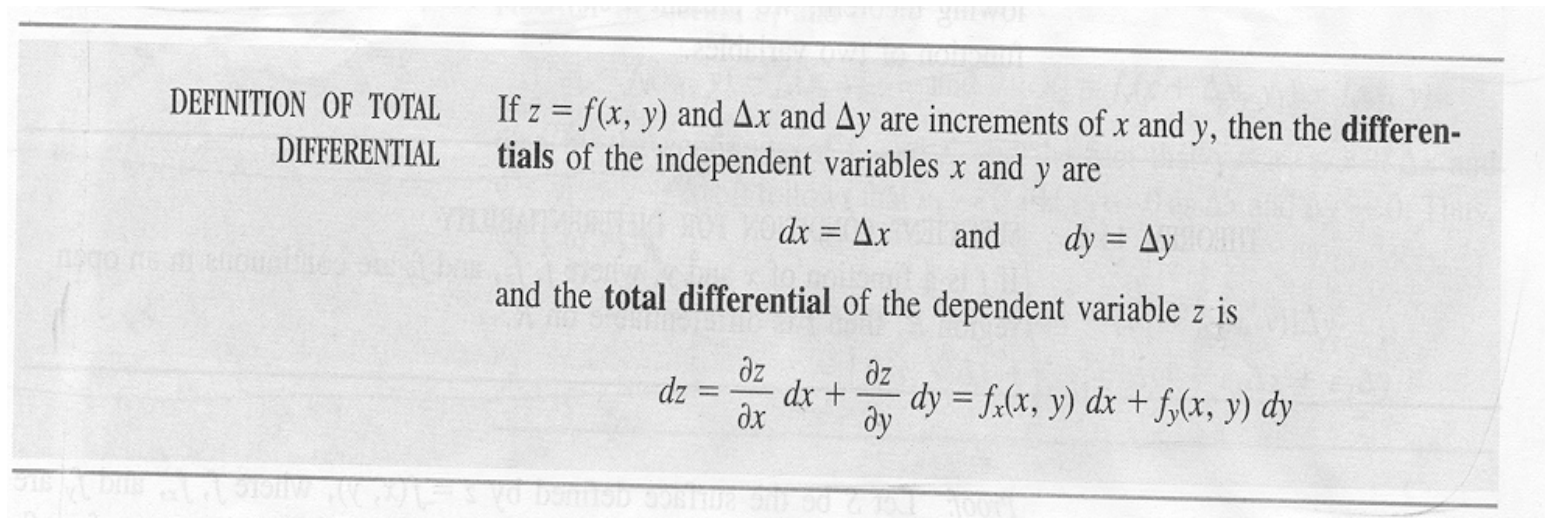
- Probability of being wrong will influence process and/or financial decisions
  - Cost / benefit of accepting result as “fact”?
  - What would be the effect downstream as the uncertainty propagates through the process?
- When comparing two values and determining if they are different
  - Overlap of uncertainty?
  - What is the probability that the difference is *significant*?

# Stating Results +/- Uncertainty

- Rule for Stating Uncertainties
  - Experimental uncertainties should almost always be rounded to one significant figure.
- Rule for Stating Answers
  - The last significant figure in any stated answer should usually be of the same order of magnitude (in the same decimal position) as the uncertainty.
  - Express Uncertainty as error bars and confidence interval for graphical data and curve fits (regressions) respectively

# Determining *Propagated* Error: Non-statistical Method

- Compute from total differential





# Propagated error

- OR Can do *sensitivity analysis* in spreadsheet of other software program
  - Compute possible uncertainty in calculated result based on varying values of inputs according to the uncertainty of each input
  - Example: Use “Solver” optimization tool in Excel to find maximum and minimum values of computed value in a cell by varying the value of each input cell
    - Set constraint that the input values lie in the range of uncertainty of that value

Or Can Use *repeat measurements* to estimate uncertainty in a result using *probability and statistics* for *random* errors:

- mean
- standard deviation of each measurement
- standard deviation of the mean of the measurements
- Confidence intervals on dependant variable
- Confidence intervals on regression parameters

# Statistical Formulas from chapter 4 of *Taylor*

## THE STANDARD DEVIATION

The average uncertainty of the individual measurements  $x_1, x_2, \dots, x_N$  is given by the standard deviation, or SD:

$$\sigma_x = \sqrt{\frac{1}{N-1} \sum (x_i - \bar{x})^2}. \quad [\text{See (4.9)}]$$

This definition of the SD, often called the *sample* standard deviation, is the most appropriate for our purposes. The *population* standard deviation is obtained by replacing the factor  $(N - 1)$  in the denominator by  $N$ . You will usually want to calculate standard deviations using the built-in function on your calculator; be sure you know which definition it uses.

The detailed significance of the standard deviation  $\sigma_x$  is that approximately 68% of the measurements of  $x$  (using the same method) should lie within a distance  $\sigma_x$  of the true value. (This claim is justified in Section 5.4.) This result is what allows us to identify  $\sigma_x$  as the *uncertainty* in any one measurement of  $x$ ,

$$\delta x = \sigma_x,$$

and, with this choice, we can be 68% confident that any one measurement will fall within  $\sigma_x$  of the correct answer.

## THE STANDARD DEVIATION OF THE MEAN

As long as systematic uncertainties are negligible, the uncertainty in our best estimate for  $x$  (namely  $\bar{x}$ ) is the standard deviation of the mean, or SDOM,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}. \quad [\text{See (4.14)}]$$

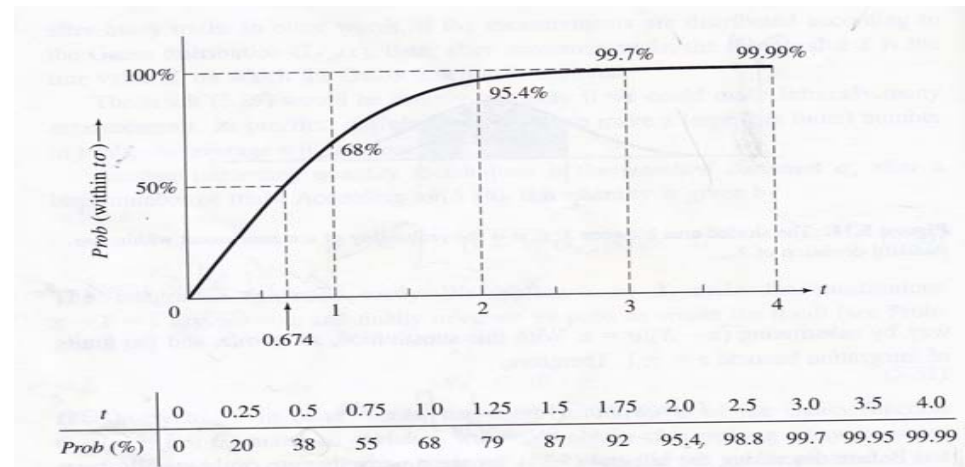
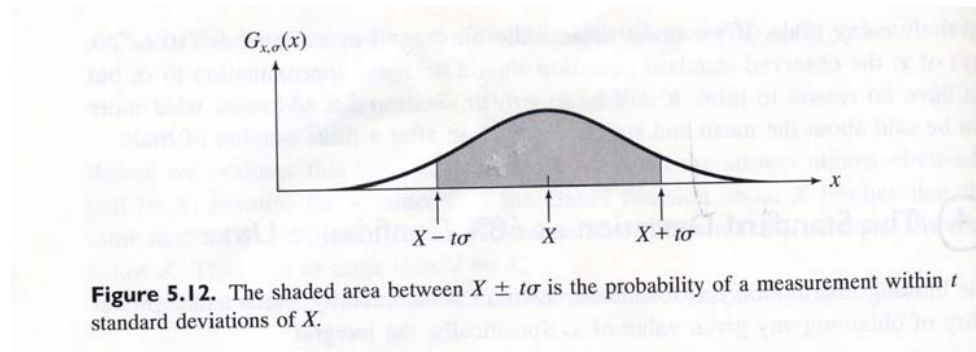
If there *are* appreciable systematic errors, then  $\sigma_{\bar{x}}$  gives the *random component* of the uncertainty in our best estimate for  $x$ :

$$\delta x_{\text{ran}} = \sigma_{\bar{x}}.$$

If you have some way to estimate the systematic component  $\delta x_{\text{sys}}$ , a reasonable (but not rigorously justified) expression for the total uncertainty is the quadratic sum of  $\delta x_{\text{ran}}$  and  $\delta x_{\text{sys}}$ :

$$\delta x_{\text{tot}} = \sqrt{(\delta x_{\text{ran}})^2 + (\delta x_{\text{sys}})^2}. \quad [\text{See (4.26)}]$$

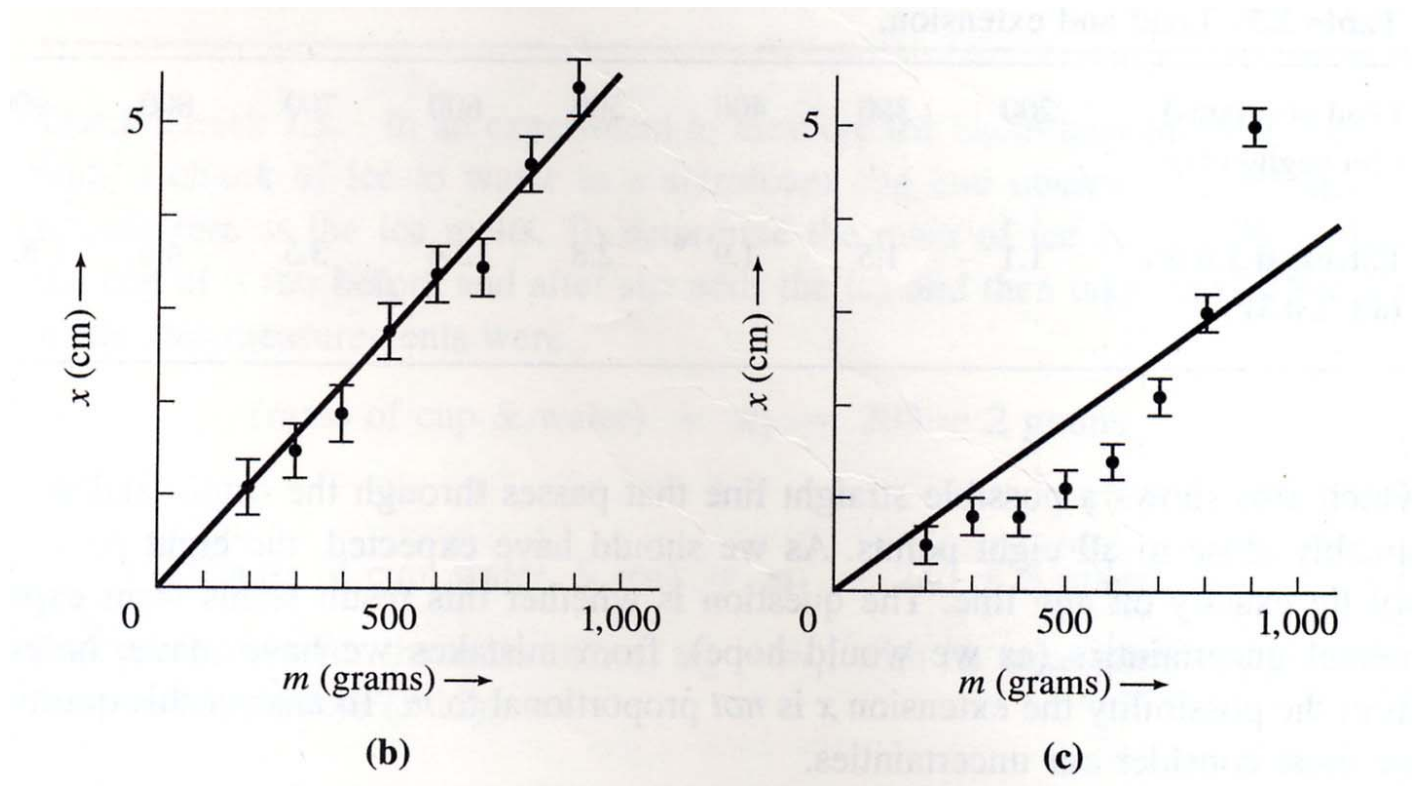
# Relationship of standard deviation to confidence intervals



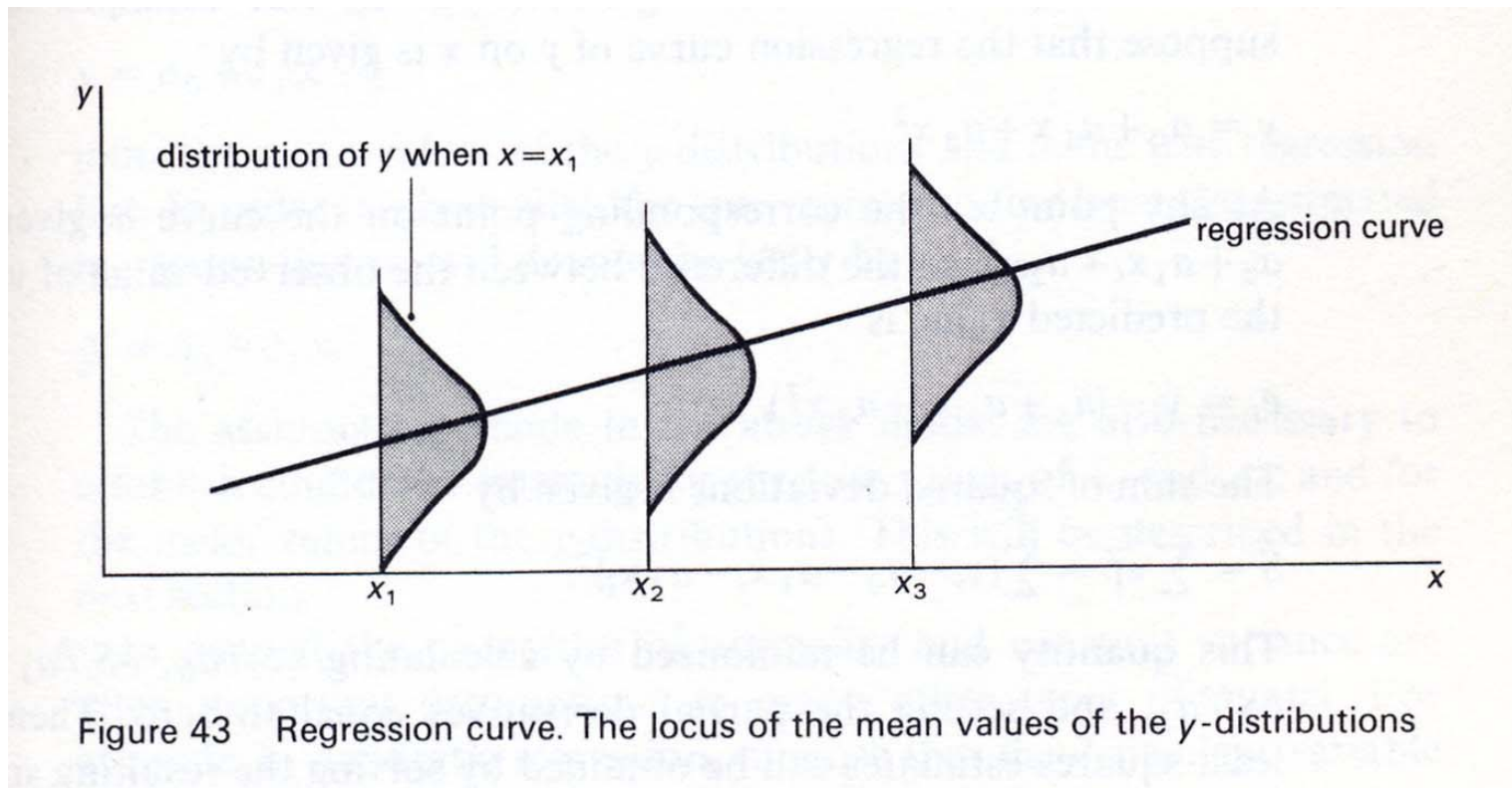
# Confidence intervals on non-linear regression coefficients

- Can be complex- use software but understand theory of how calculated for linear case

Error bars that represent uncertainty in the dependant variable



How measurements at a given  $x, y$  would be distributed for multiple measurements



# Determining Slope and Intercept In Linear Regression

## A STRAIGHT LINE, $y = A + Bx$ ; EQUAL WEIGHTS

If  $y$  is expected to lie on a straight line  $y = A + Bx$ , and if the measurements of  $y$  all have the same uncertainties, then the best estimates for the constants  $A$  and  $B$  are:

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

and

$$B = \frac{N \sum xy - \sum x \sum y}{\Delta},$$

where the denominator,  $\Delta$ , is

$$\Delta = N \sum x^2 - (\sum x)^2. \quad [\text{See (8.10) to (8.12)}]$$

Based on the observed points, the best estimate for the uncertainty in the measurements of  $y$  is

$$\sigma_y = \sqrt{\frac{1}{N-2} \sum_{i=1}^N (y_i - A - Bx_i)^2}. \quad [\text{See (8.15)}]$$



# Confidence intervals (SD) on slope B and Intercept A

## Chapter 8: Least-Squares Fitting

The uncertainties in  $A$  and  $B$  are:

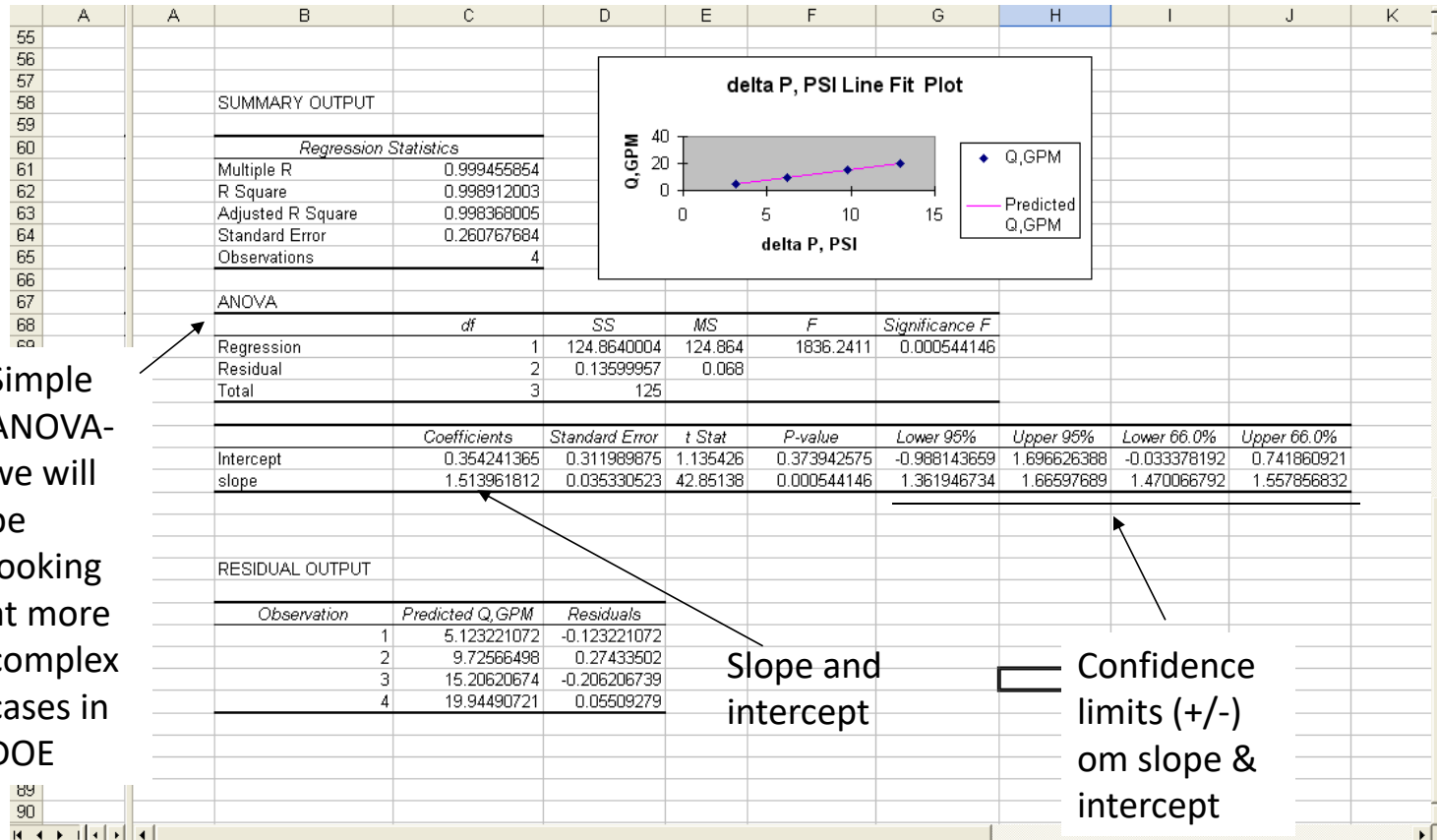
$$\sigma_A = \sigma_y \sqrt{\frac{\sum x^2}{\Delta}}$$

and

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}$$

[See (8.16) & (8.17)]

# Regression Output in Excel

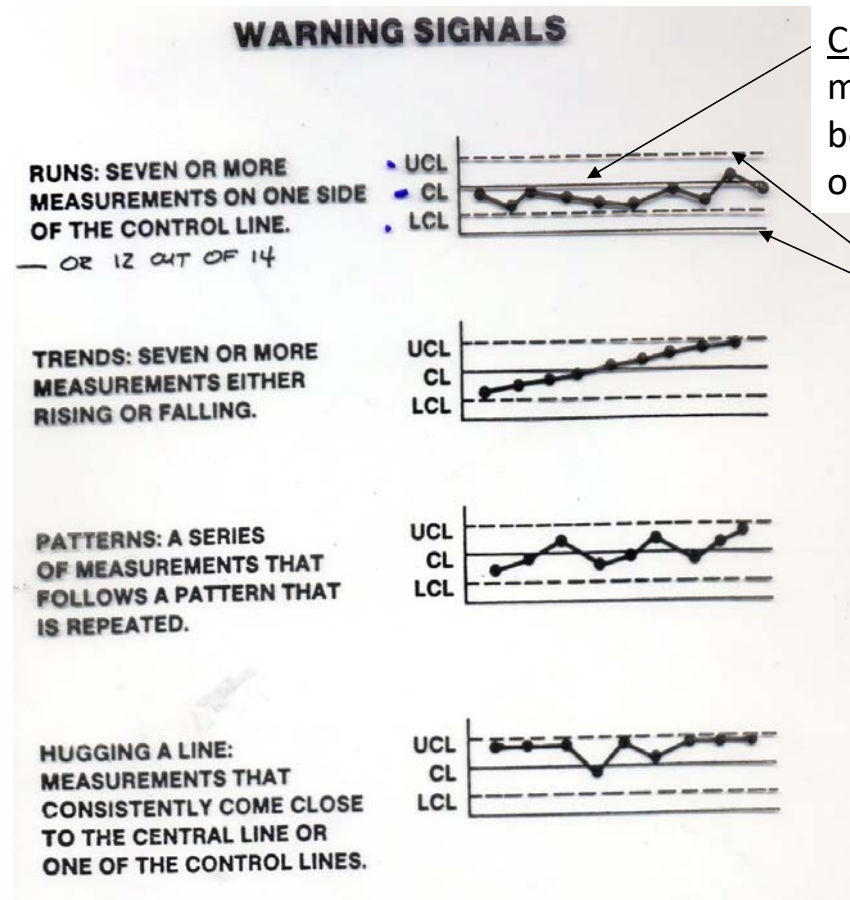


# Statistical Process Control

- Very Widely Used
- Used for quality control and **in conjunction with DOE** for process improvement
- *Control Charts* provide statistical evidence
  - That a process is behaving normally or if something wrong
  - **Serve as data output (dependant variable )from process in designed statistical experiments**

Variation from expected behavior in control charts- similar to regression and point statistics

Expect random deviations from mean just like in regression



Control Limit is the mean of a well behaved process output (i.e. product)

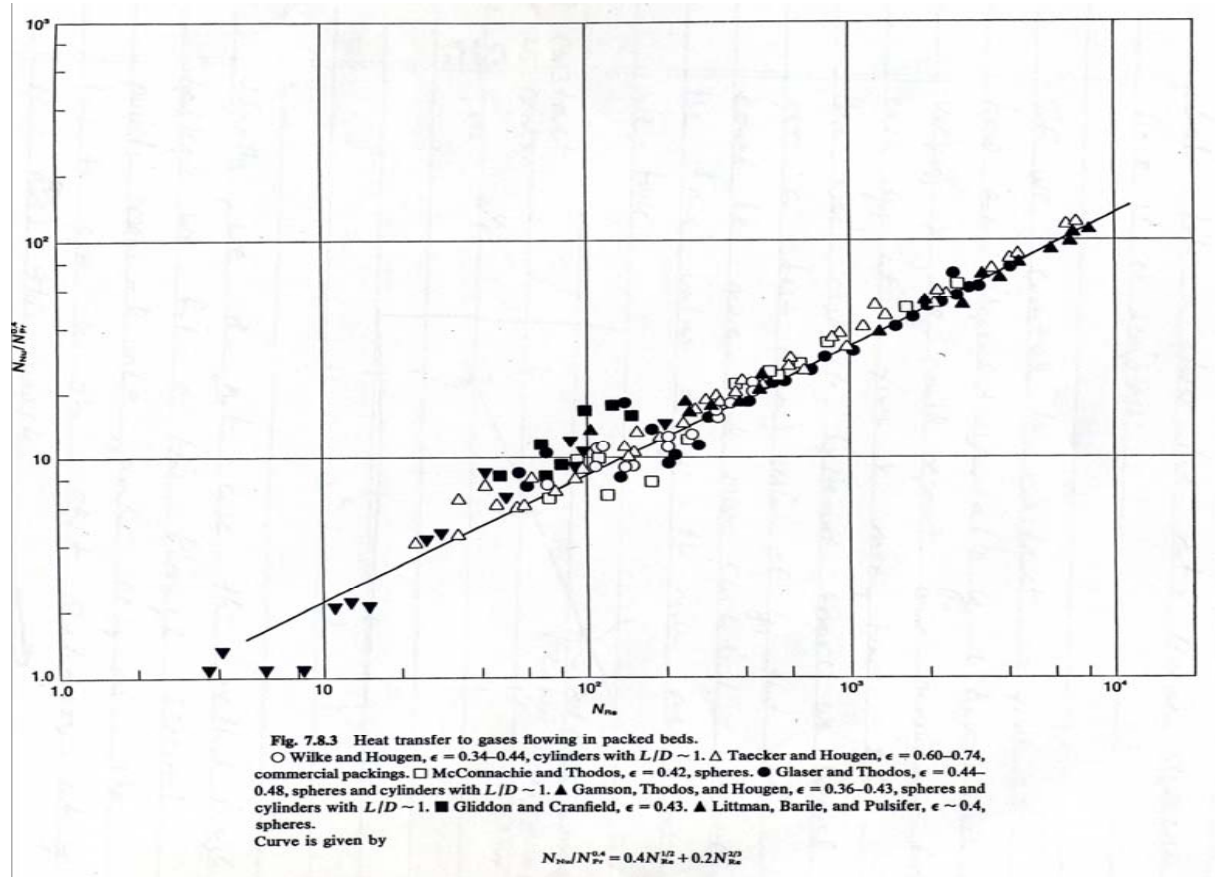
Upper and lower Control Limits represent confidence limit on mean of "well behaved" process output

# Theory and Experimentation

- Two fundamental approaches to problem solving problems in the discovery of knowledge:
  1. Theoretical (physical/mathematical modeling)
  2. Experimental measurement

(Most often a combination is used)

Example of combination of theory and experimentation to get semi-empirical correlation



# Features of alternative methods

- Theoretical Models

- Simplifying assumptions needed
- General results
- Less facilities usually needed
- Can start study immediately

- Experimental approach

- Study the “real world”-no simplifying assumptions needed
- Results specific to apparatus studied
- High accuracy measurements need complex instruments
- Extensive lab facilities maybe needed
- Time delays from building apparatus, debugging

# Functional Types of Engineering Experiments

1. Determine material properties
2. Determine component or system performance indices
3. Evaluate/improve theoretical models
4. Product/process improvement by testing
5. Exploratory experimentation
6. Acceptance testing
7. Teaching/learning through experimentation



# Some important classes of Experiments

1. Estimation of parameter mean value
2. Estimate of parameter variability
3. Comparison of mean values
4. Comparison of variability
5. Modeling the dependence of dependant Variable on several quantitative and/or qualitative variables

# Practical Experimental Planning

## Experimental design:

- Consider *goals*
- Consider what data can be collected.
- Difficulty of obtaining data
- What data is *most* important
- What measurements can be ignored
- Type of data: Categorical? Quantitative?
- Test to make sure that measurements/apparatus are reliable
- Collect data carefully and document fully in ink using bound notebooks. Make copies and keep separately

# Preview of Uses for DOE

- Lab experiments for research
- Industrial process experiments

## Four engineering problem classes to which DOE is applied in manufacturing

1. Comparison
2. Screening/ characterization
3. Modeling
4. Optimization

# Comparison

- Compares to see if a change in a single “factor” (variable) has resulted in a process change (ideally an improvement)

# Screening/Characterization

- Used when you want to see the effect of a whole range of factors so as to know which one(s) are most important.
- Two factorial experiments usually used

# Modeling

- Used when you want to be able to construct a mathematical model that will predict the effect on a process of manipulating a variables or multiple variables

# Optimization

- When you want to determine the optimal settings for all factors to give an optimal process response.



# Introduction to experimental design

# Contents

- planning experiments
- regression analysis
- types of experiments
- software
- literature

$$A = \pi r^2$$

## Example of Experiment : synthesis of T8-POSS

- context: development of new synthesis route for polymer additive
- goal: optimize yield of reaction
- synthesis route consists of elements that are not uniquely determined (control variables):
  - time to let reaction run
  - concentration water
  - concentration silane
  - temperature
  - ...

## Issues in example T8-POSS synthesis

- how to measure yield
  - what to measure (begin/end weight,...)
  - when to measure (reaction requires at least one day)
- how to vary control variables
  - which values of pH, concentrations, ... (*levels*)
  - which combinations of values
  - equipment only allows 6 simultaneous reactions, all with the same temperature
- how many combinations can be tested
  - reaction requires at least one day
  - only 4 experimentation days are available

## Necessity of careful planning of experiment

- limited resources
  - time to carry out experiment
  - costs of required materials/equipment
- avoid reaching suboptimal settings
- avoid missing interesting parts of experimental region
- protection against external uncontrollable/undetectable influences
- getting precise estimates

Traditional approach to experimentation:

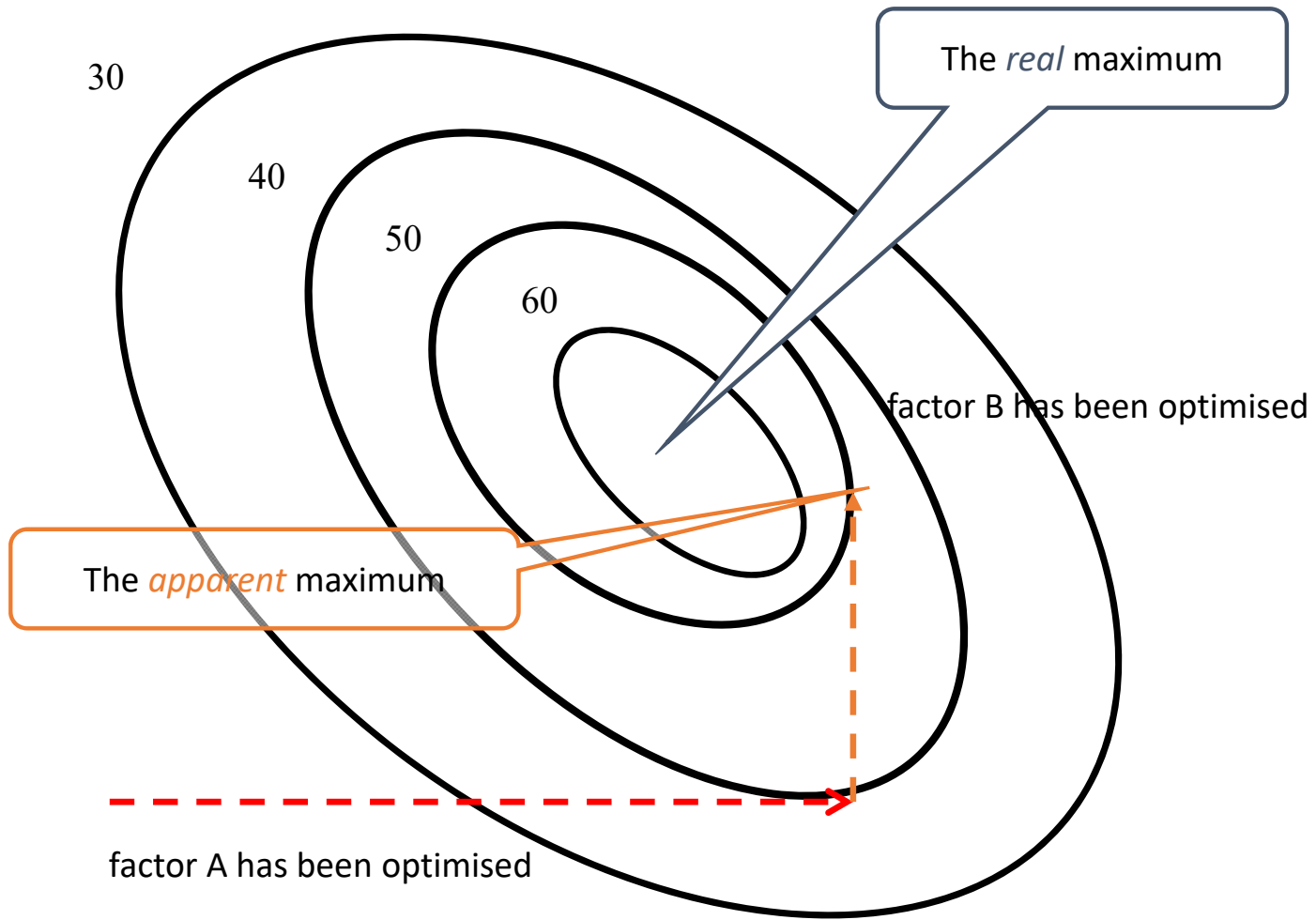
T8-POSS example

- set  $T = 40\text{ }^{\circ}\text{C}$ ,  $\text{H}_2\text{O}$  concentration = 10%; try  $c_{\text{Si}} = 0.1, 0.2, 0.3, 0.8, 0.9, 1.0\text{ M}$
- set  $T = 60\text{ }^{\circ}\text{C}$ ,  $c_{\text{Si}} = 0.5\text{ M}$ ,  $\text{H}_2\text{O}$  concentration = 5, 10, 12.5, 15, 17.5, 20%
- ...

This is called a One-Factor-At-a-Time (OFAT) or Change-One-Separate-factor-at-a-Time (COST) strategy.

Disadvantages:

- may lead to suboptimal settings (see next slide)
- requires too many runs to obtain good coverage of experimental region (see later)



Statistical terminology for experiments:  
illustrated by T8-POSS example

- response variable: yield
- factors: time, temperature,  $c_{Si}$ , H<sub>2</sub>O concentration
- levels: actual values of factors (e.g., T=30 °C, 40 °C, 50 °C)
- runs: one combination of factor settings (e.g., T=30 °C,  $c_{Si}$ =0.5M, H<sub>2</sub>O concentration = 15%)
- block: 6 simultaneous runs with same temperature in reaction station



Modern approach: DOE

- DOE = Design of Experiments
- key ideas:
  - change several factors simultaneously
  - carefully choose which runs to perform
  - use regression analysis to obtain effect estimates
- statistical software (Statgraphics, JMP, SAS,...) allows to
  - choose or construct designs
  - analyse experimental results

Example of analysis

simple experiment:

- response is conversion
- goal is screening (are time and temperature influencing conversion?)
- 2 factors (time and temperature), each at two levels
- 5 centre points (both time and temperature at intermediate values)

Statgraphics demo with conversion.sfx. (choose Special -> Experimental Design etc. from menu)

More advanced (5 factors, not all  $2^5$  combinations): colour.sfx

Example of construction: T8-POSS example

- 36 runs
  - 2 reactors available each day (each reactor 6 places)
  - 3 experimental days
- factors:
  - H<sub>2</sub>O concentration
  - temperature
  - $c_{Si}$
- goal is optimization of response
- choose in Statgraphics: Special -> Experimental Design -> Create Design -> Response Surface

## Goals in experimentation

- there may be more than one goal, e.g.:
  - yield
  - required reaction time until equilibrium
  - costs of required chemical substances
  - impact on environment (waste)
- these goals may contradict each other
- goals must be converted to explicitly measurable quantities

# Types of experimental designs

- “screening designs”

These designs are used to investigate *which* factors are important (“significant”).

- “response surface designs”

These designs are used to determine the *optimal* settings of the significant factors.

# Interactions

Factors may *influence each other*. E.g, the optimal setting of a factor may depend on the settings of the other factors.

When factors are optimised *separately*, the overall result (as function of all factors) may be *suboptimal*

...

# Interaction effects

Cross terms in linear regression models cause interaction effects:

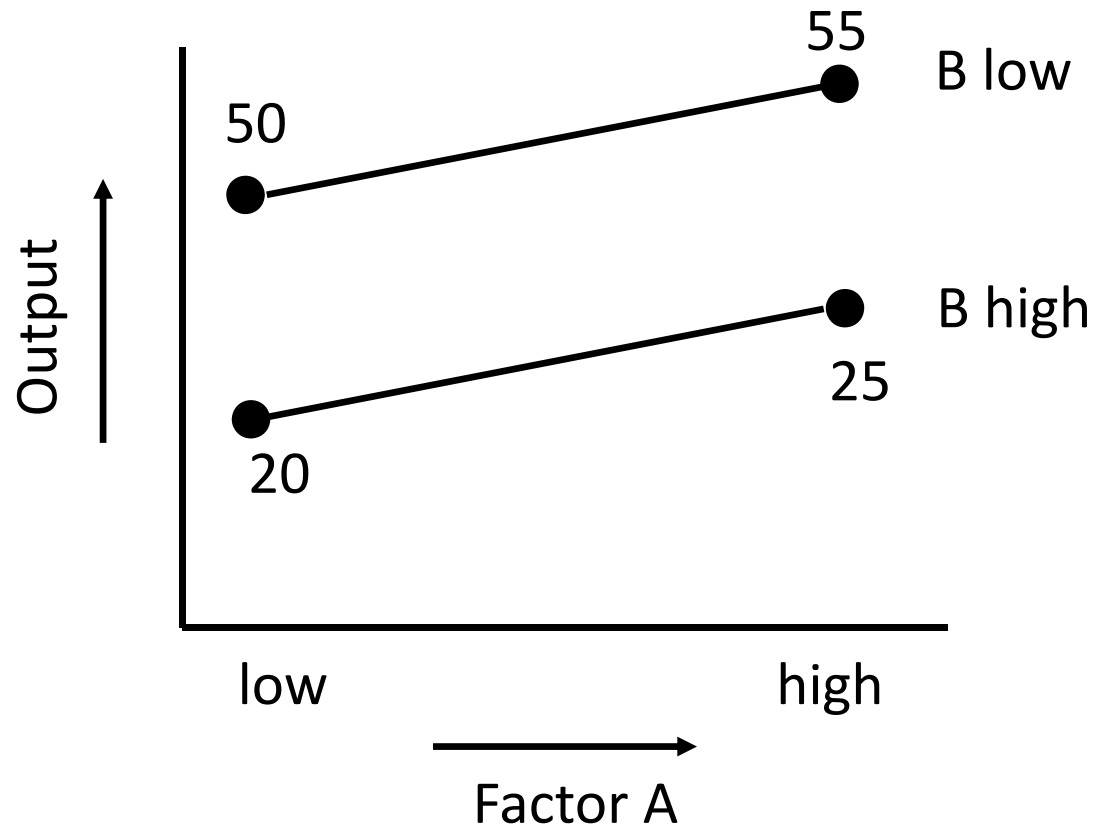
$$Y = 3 + 2x_A + 4x_B + 7x_Ax_B$$

$$x_A \rightarrow x_A + 1 \Rightarrow Y \rightarrow Y + 2 + 7x_B,$$

so increase depends on  $x_B$ . Likewise for  $x_B \rightarrow x_B + 1$

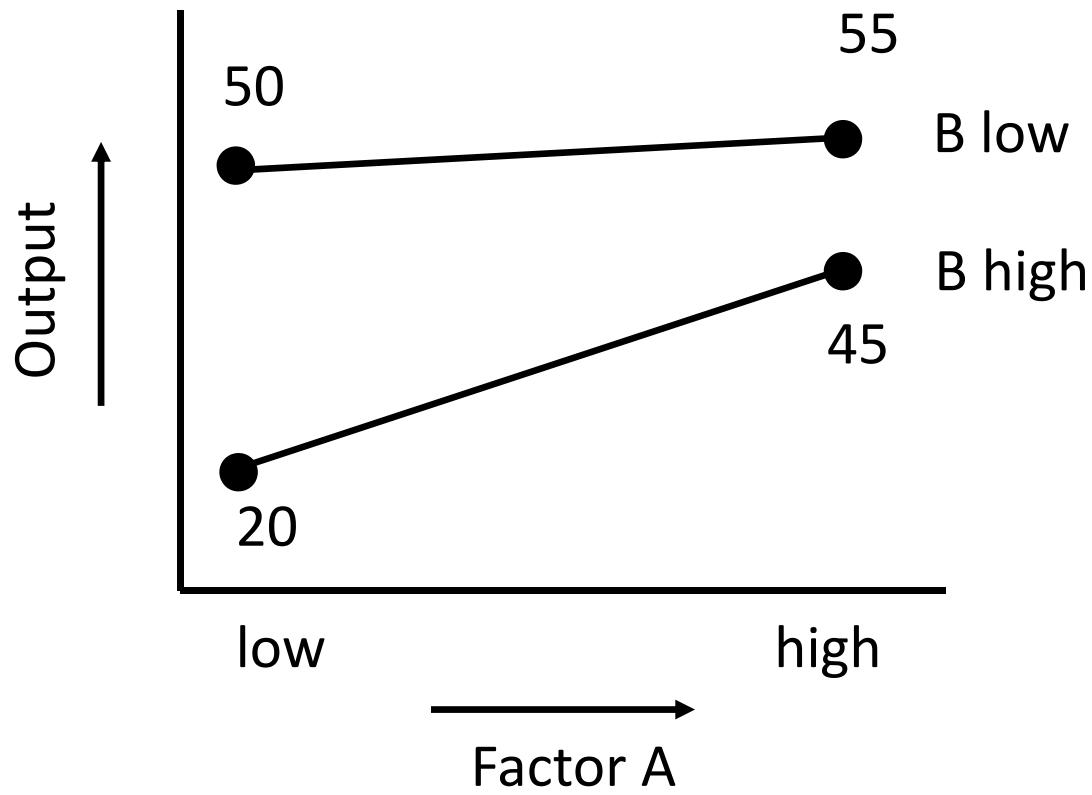
This explains the notation AB for the interaction of factors A and B.

No interaction

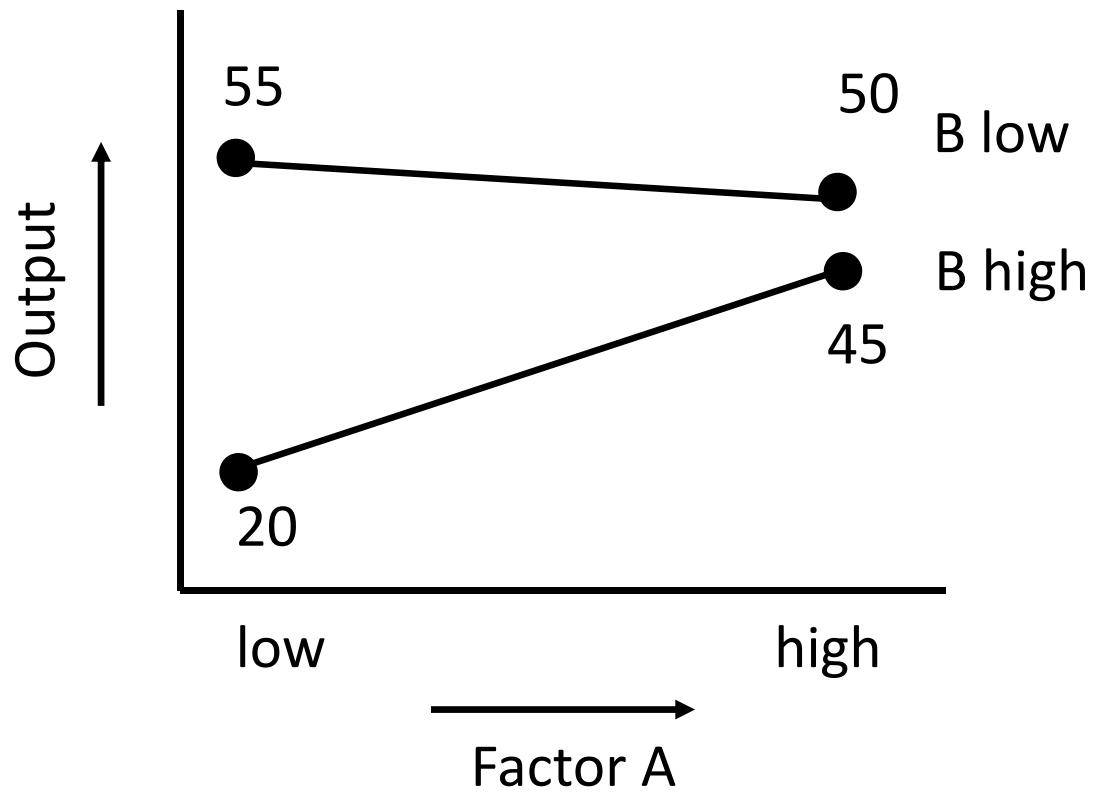




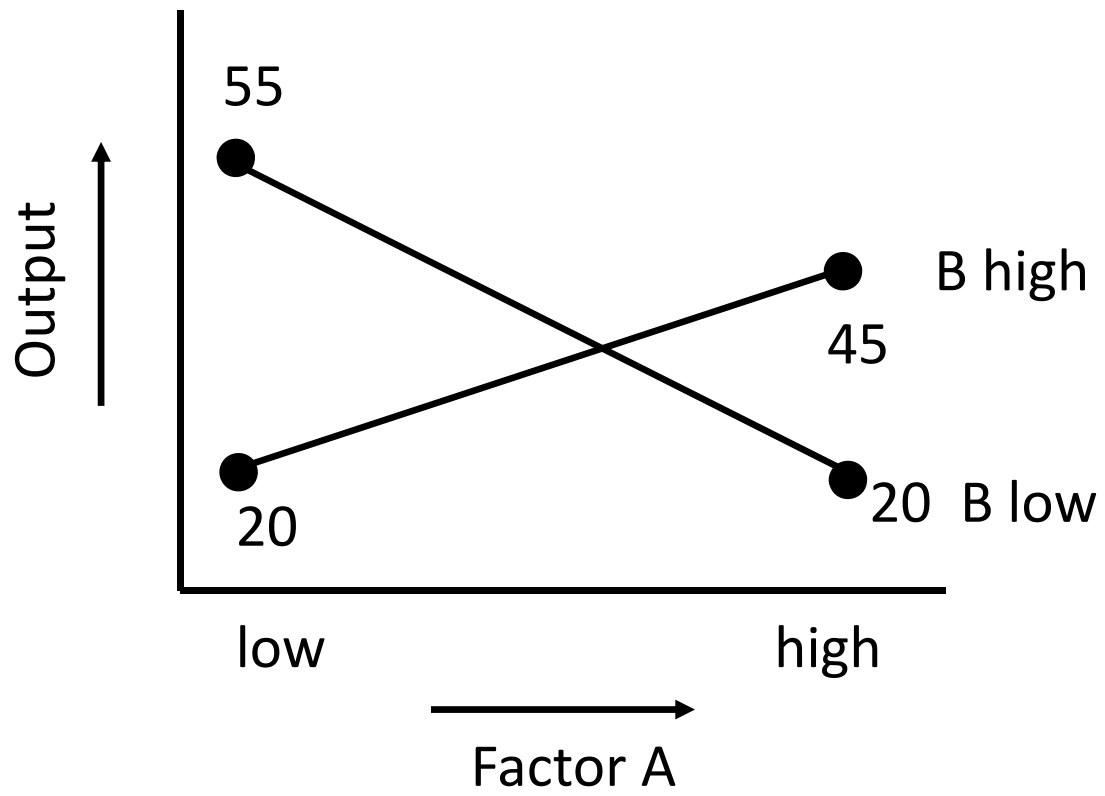
# Interaction I



## Interaction II



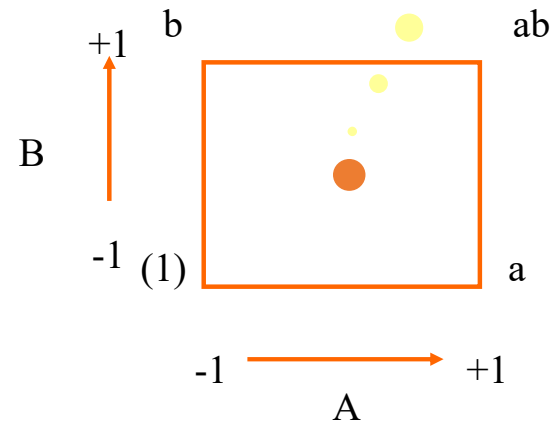
### Interaction III



# Centre points and Replications

If there are not enough measurements to obtain a good estimate of the variance, then one can perform replications. Another possibility is to add *centre points* .

Adding centre points serves two purposes:  
better variance estimate  
allow to test curvature using  
a lack-of-fit test



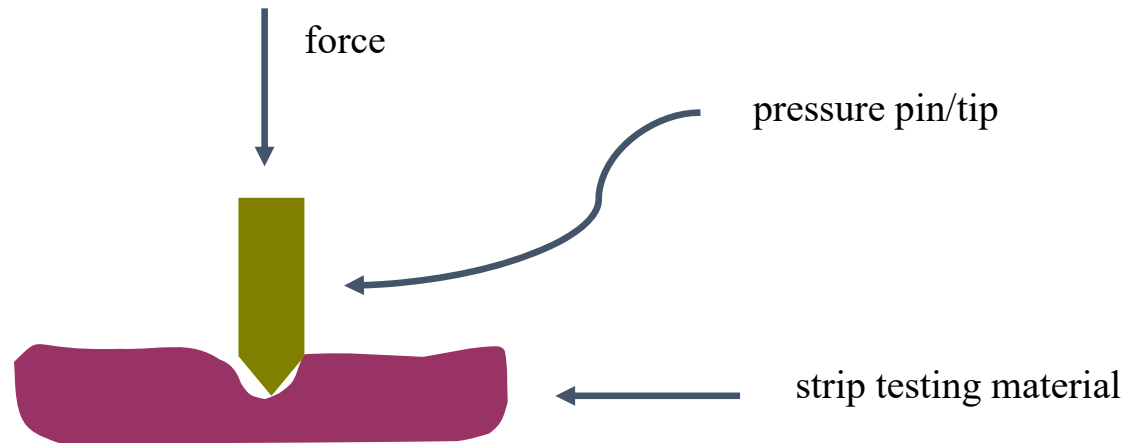
# Multi-layered experiments

Experiments are not one-shot adventures. Ideally one performs:

- an initial experiment
  - check-out experimental equipment
  - get initial values for quantities of interest
- main experiment
  - obtain results that support the goal of the experiment
- confirmation experiment
  - verify results from main experiment
  - use information from main experiment to conduct more focussed experiment (e.g., near computed optimum)

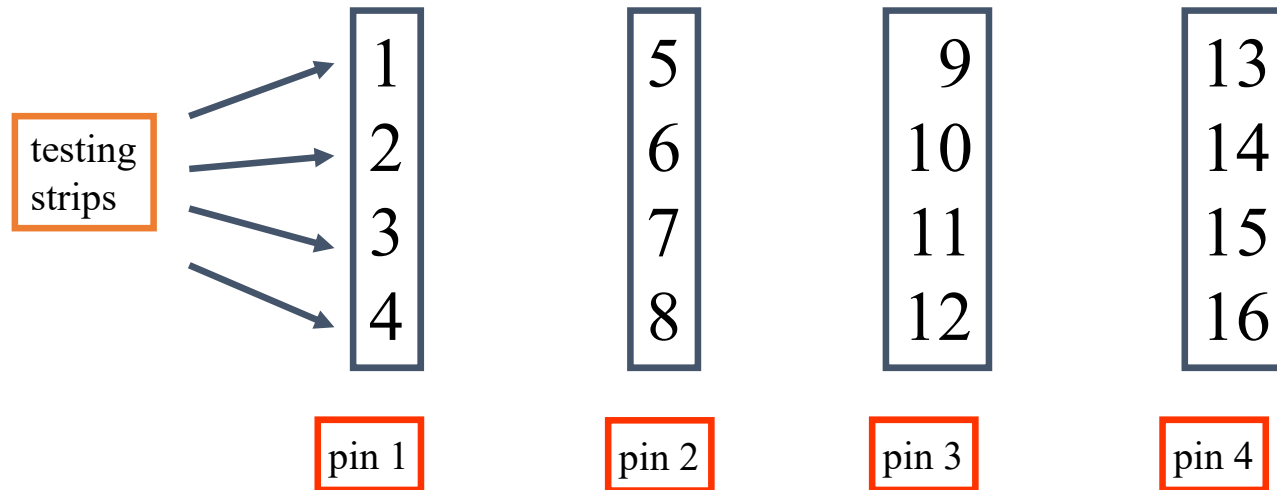
# Example

- *testing method for material hardness* :



*practical problem*: 4 types of pressure pins  
⇒ do these yield the same results?

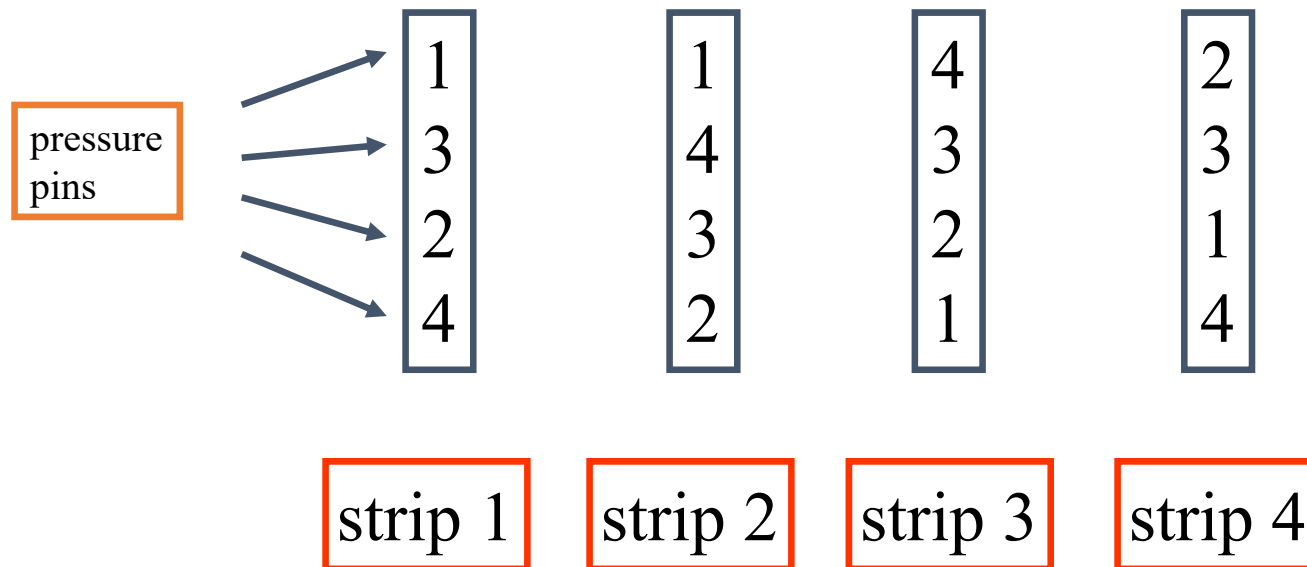
# Experimental design 1



*Problem:* if the measurements of strips 5 through 8 differ, is this caused by the strips or by pin 2?

# Experimental design 2

- Take 4 strips on which you measure (in random order) *each* pressure pin once :





# Blocking

- *Advantage* of *blocked* experimental design 2:  
differences between strips are filtered out

- *Model*:  $Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$

factor  
pressure pin

block effect  
strip

error term

*Primary goal*: reduction error term



## Short checklist for DOE (see protocol)

- clearly state objective of experiment
- check constraints on experiment
  - constraints on factor combinations and/or changes
  - constraints on size of experiment
- make sure that measurements are obtained under constant external conditions (if not, apply blocking!)
- include centre points to validate model assumptions
  - check of constant variance
  - check of non-linearity
- make clear protocol of execution of experiment (including randomised order of measurements)

## Introduction: What is meant by DOE?

- Experiment -
  - a test or a series of tests in which purposeful changes are made to the *input variables or factors* of a system so that we may observe and identify the reasons for changes in the *output* response(s).
- Question: 5 factors, and 2 response variables
  - Want to know the effect of each factor on the response and how the factors may interact with each other
  - Want to predict the responses for given levels of the factors
  - Want to find the levels of the factors that optimizes the responses - e.g. maximize  $Y_1$  but minimize  $Y_2$
  - Time and budget allocated for 30 test runs only.

# Strategy of Experimentation

- Strategy of experimentation
  - Best guess approach (trial and error)
    - can continue indefinitely
    - cannot guarantee best solution has been found
  - One-factor-at-a-time (OFAT) approach
    - inefficient (requires many test runs)
    - fails to consider any possible interaction between factors
  - Factorial approach (invented in the 1920's)
    - Factors varied together
    - Correct, modern, and most efficient approach
    - Can determine how factors interact
    - Used extensively in industrial R and D, and for process improvement.

- This course will focus on three very useful and important classes of factorial designs:
  - 2-level full factorial ( $2^k$ )
  - fractional factorial ( $2^{k-p}$ ), and
  - response surface methodology (RSM)
- I will also cover split plot designs, and design and analysis of computer experiments if time permits.
- Dimensional analysis and how it can be combined with DOE will also be briefly covered.
- All DOE are based on the same statistical principles and method of analysis - ANOVA and regression analysis.
- *Answer to question: use a  $2^{5-1}$  fractional factorial in a central composite design = 27 runs (min)*

# Statistical Design of Experiments

- All experiments should be designed experiments
- Unfortunately, some experiments are poorly designed - valuable resources are used ineffectively and results inconclusive
- Statistically designed experiments permit efficiency and economy, and the use of statistical methods in examining the data result in scientific objectivity when drawing conclusions.

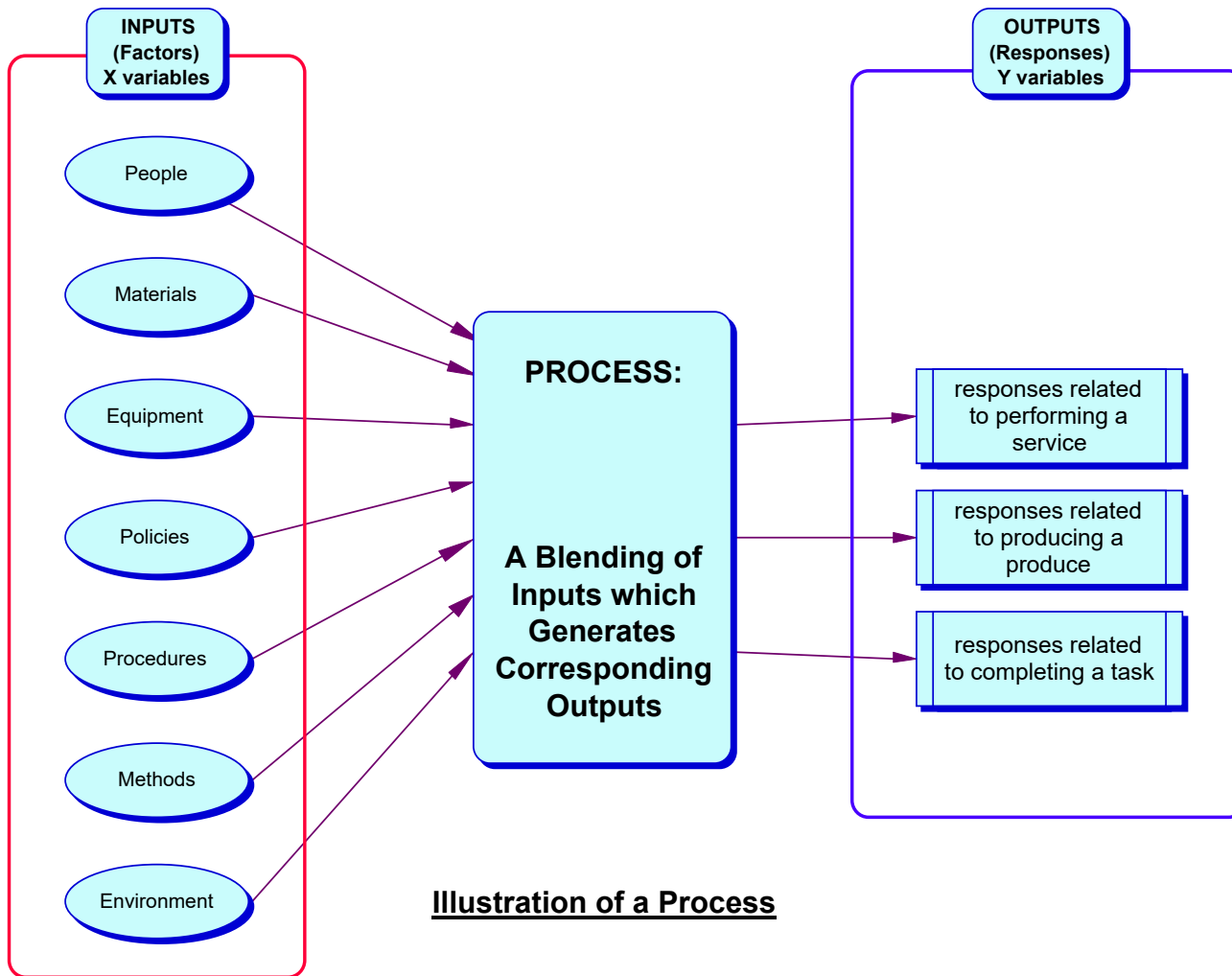
- DOE is a methodology for systematically applying statistics to experimentation.
- DOE lets experimenters develop a mathematical model that predicts how input variables interact to create output variables or responses in a process or system.
- DOE can be used for a wide range of experiments for various purposes including nearly all fields of engineering and even in business marketing.
- Use of statistics is very important in DOE and the basics are covered in a first course in an engineering program.

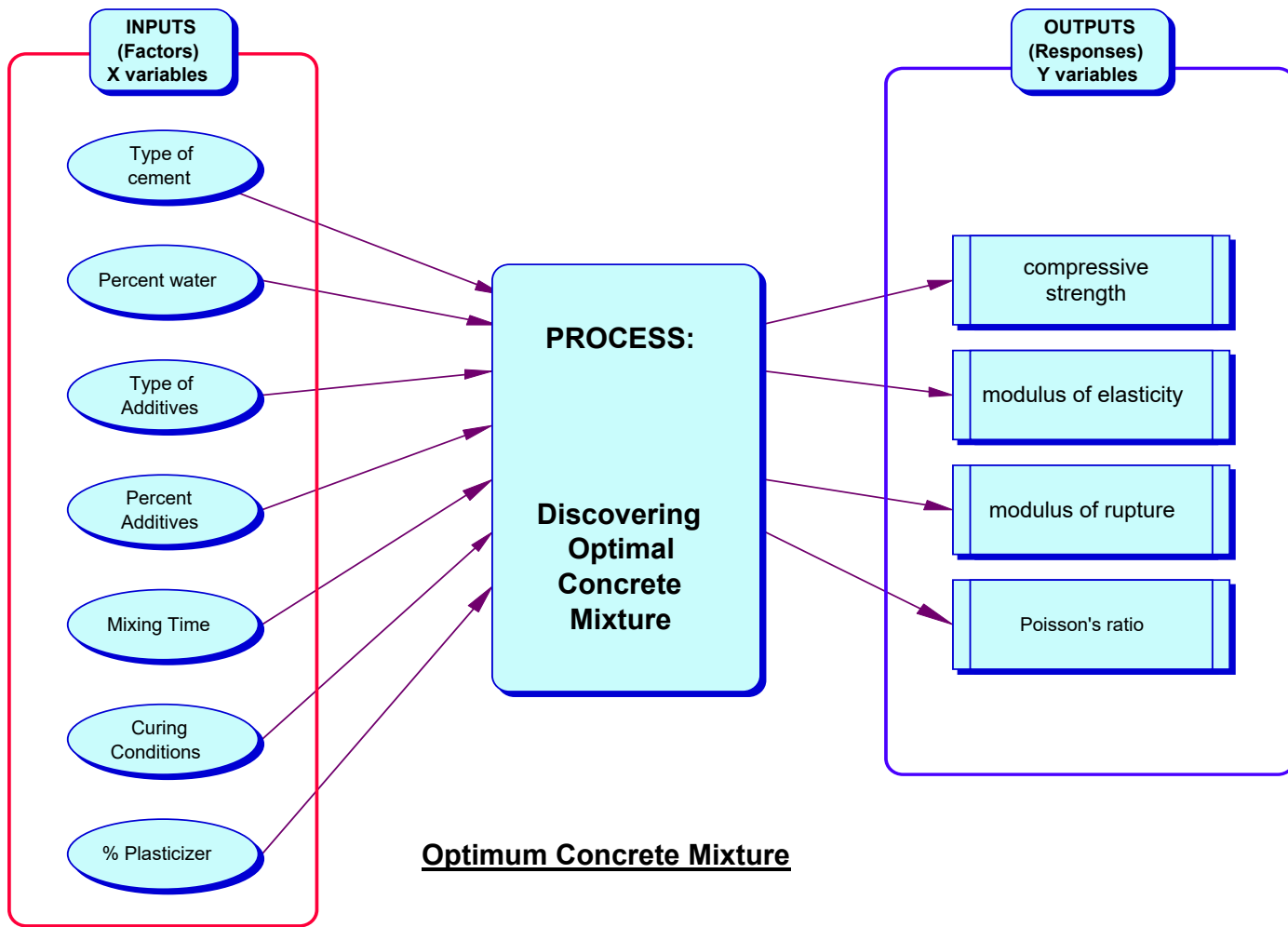
- In general, by using DOE, we can:
  - Learn about the process we are investigating
  - Screen important variables
  - Build a mathematical model
  - Obtain prediction equations
  - Optimize the response (if required)
- Statistical significance is tested using **ANOVA**, and the prediction model is obtained using **regression analysis**.

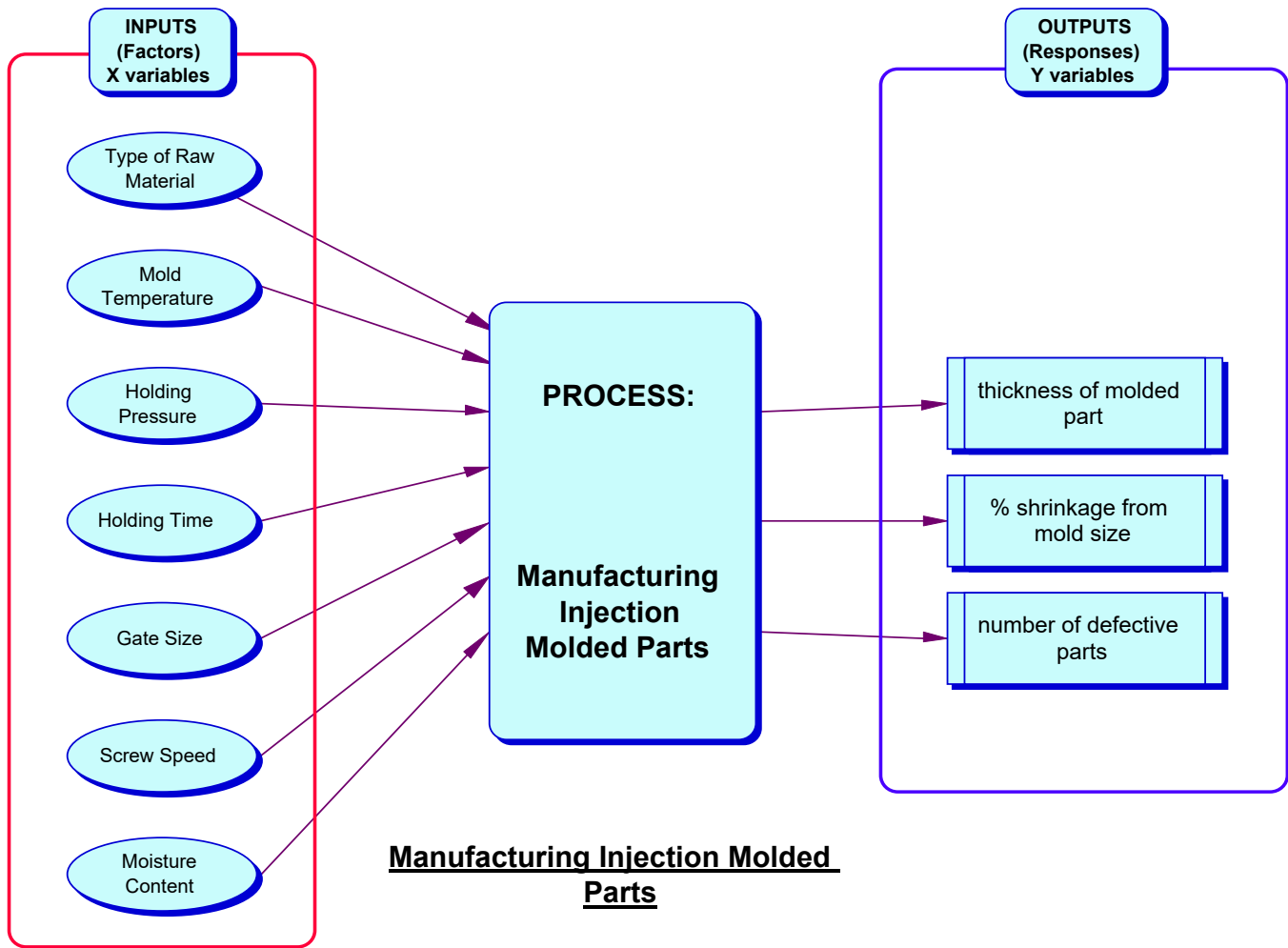


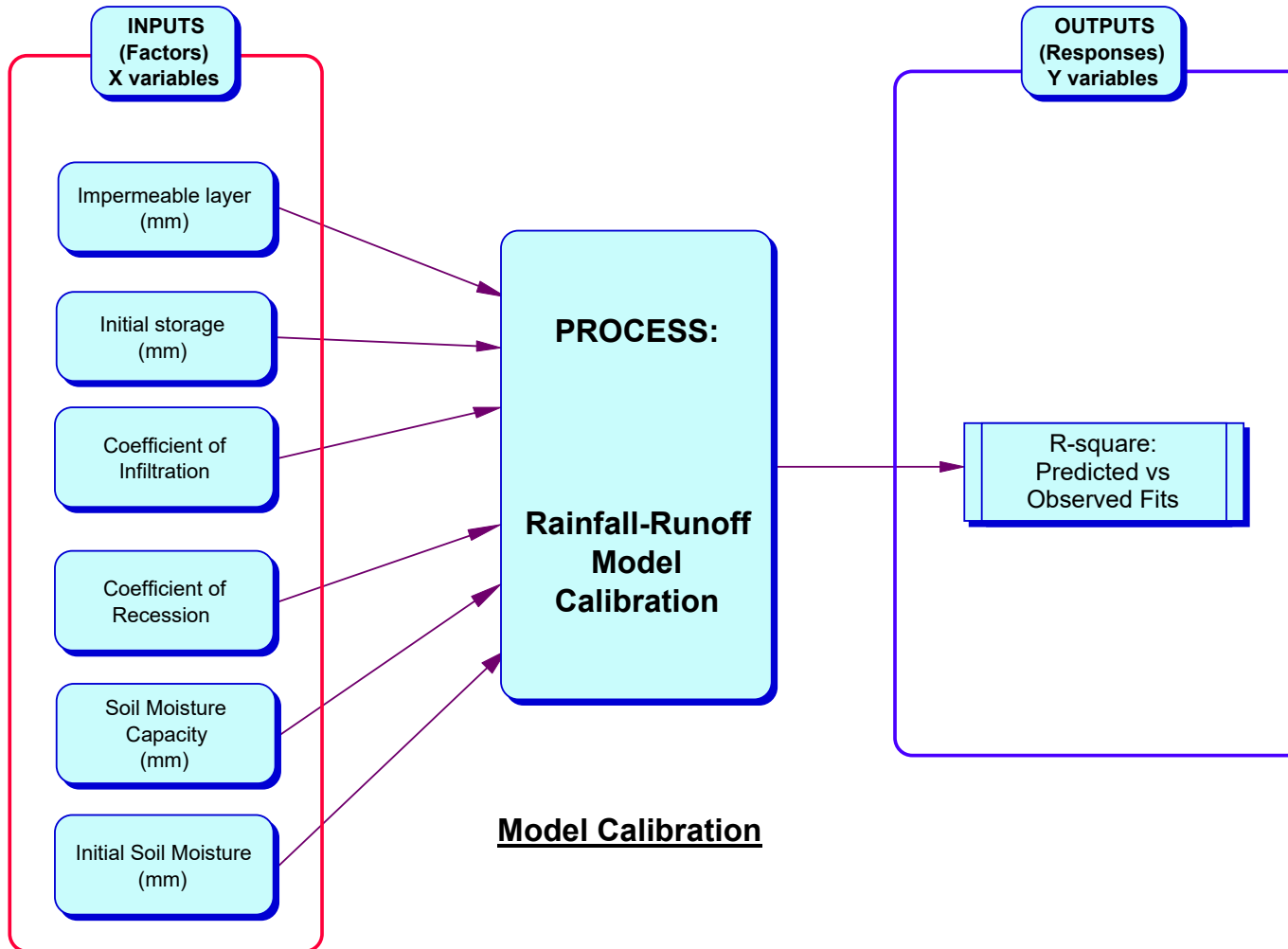
## Applications of DOE in Engineering Design

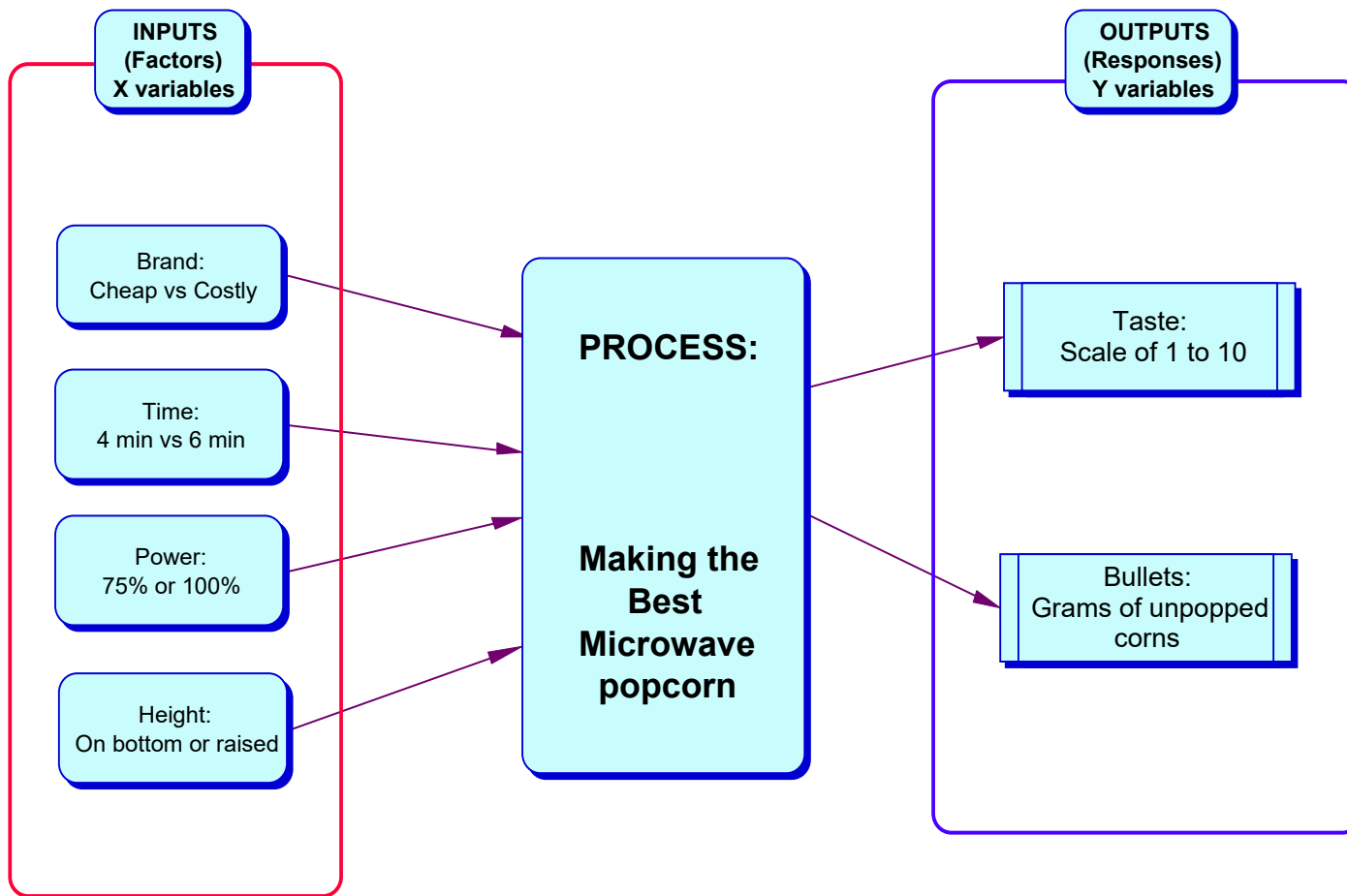
- Experiments are conducted in the field of engineering to:
  - evaluate and compare basic design configurations
  - evaluate different materials
  - select design parameters so that the design will work well under a wide variety of field conditions (robust design)
  - determine key design parameters that impact performance











**Making microwave popcorn**

## Examples of experiments from daily life

- Photography
  - Factors: speed of film, lighting, shutter speed
  - Response: quality of slides made close up with flash attachment
- Boiling water
  - Factors: Pan type, burner size, cover
  - Response: Time to boil water
- D-day
  - Factors: Type of drink, number of drinks, rate of drinking, time after last meal
  - Response: Time to get a steel ball through a maze
- Mailing
  - Factors: stamp, area code, time of day when letter mailed
  - Response: Number of days required for letter to be delivered

## More examples

- Cooking
  - Factors: amount of cooking wine, oyster sauce, sesame oil
  - Response: Taste of stewed chicken
- Sexual Pleasure
  - Factors: marijuana, screech, sauna
  - Response: Pleasure experienced in subsequent you know what
- Basketball
  - Factors: Distance from basket, type of shot, location on floor
  - Response: Number of shots made (out of 10) with basketball
- Skiing
  - Factors: Ski type, temperature, type of wax
  - Response: Time to go down ski slope



# Basic Principles

- Statistical design of experiments (DOE)
  - the process of planning experiments so that appropriate data can be analyzed by statistical methods that results in valid, objective, and meaningful conclusions from the data
  - involves two aspects: design and statistical analysis

- Every experiment involves a sequence of activities:
  - Conjecture - hypothesis that motivates the experiment
  - Experiment - the test performed to investigate the conjecture
  - Analysis - the statistical analysis of the data from the experiment
  - Conclusion - what has been learned about the original conjecture from the experiment.

## Three basic principles of Statistical DOE

- **Replication**
  - allows an estimate of experimental error
  - allows for a more precise estimate of the sample mean value
- **Randomization**
  - cornerstone of all statistical methods
  - “average out” effects of extraneous factors
  - reduce bias and systematic errors
- **Blocking**
  - increases precision of experiment
  - “factor out” variable not studied

# Guidelines for Designing Experiments

- Recognition of and statement of the problem
  - need to develop all ideas about the objectives of the experiment - get input from everybody - use team approach.
- Choice of factors, levels, ranges, and response variables.
  - Need to use engineering judgment or prior test results.
- Choice of experimental design
  - sample size, replicates, run order, randomization, software to use, design of data collection forms.

- Performing the experiment
  - vital to monitor the process carefully. Easy to underestimate logistical and planning aspects in a complex R and D environment.
- Statistical analysis of data
  - provides objective conclusions - use simple graphics whenever possible.
- Conclusion and recommendations
  - follow-up test runs and confirmation testing to validate the conclusions from the experiment.
- Do we need to add or drop factors, change ranges, levels, new responses, etc.. ???

## Using Statistical Techniques in Experimentation - things to keep in mind

- Use non-statistical knowledge of the problem
  - physical laws, background knowledge
- Keep the design and analysis as simple as possible
  - Don't use complex, sophisticated statistical techniques
  - If design is good, analysis is relatively straightforward
  - If design is bad - even the most complex and elegant statistics cannot save the situation
- Recognize the difference between practical and statistical significance
  - statistical significance  $\neq$  practical significance

- Experiments are usually iterative
  - unwise to design a comprehensive experiment at the start of the study
  - may need modification of factor levels, factors, responses, etc.. - too early to know whether experiment would work
  - use a sequential or iterative approach
  - should not invest more than 25% of resources in the initial design.
  - Use initial design as learning experiences to accomplish the final objectives of the experiment.

## Factorial v.s. OFAT

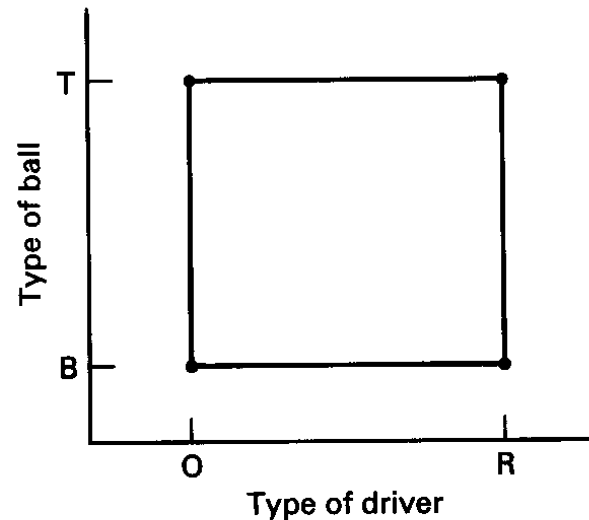
- Factorial design - experimental trials or runs are performed at all possible combinations of factor levels in contrast to OFAT experiments.
- Factorial and fractional factorial experiments are among the most useful multi-factor experiments for engineering and scientific investigations.



- The ability to gain competitive advantage requires extreme care in the design and conduct of experiments. Special attention must be paid to joint effects and estimates of variability that are provided by factorial experiments.
- Full and fractional experiments can be conducted using a variety of statistical designs. The design selected can be chosen according to specific requirements and restrictions of the investigation.

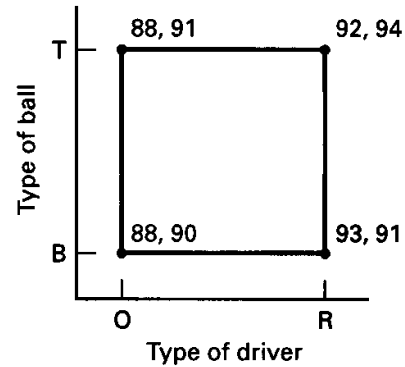
# Factorial Designs

- In a factorial experiment, **all possible combinations** of factor levels are tested
- The golf experiment:
  - Type of driver (over or regular)
  - Type of ball (balata or 3-piece)
  - Walking vs. riding a cart
  - Type of beverage (Beer vs water)
  - Time of round (am or pm)
  - Weather
  - Type of golf spike
  - Etc, etc, etc...

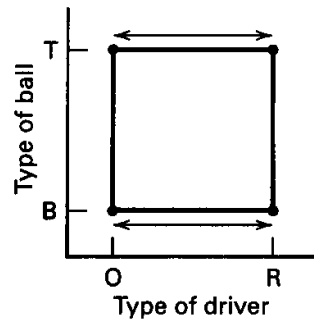


A two-factor factorial experiment involving type of driver and type of ball.

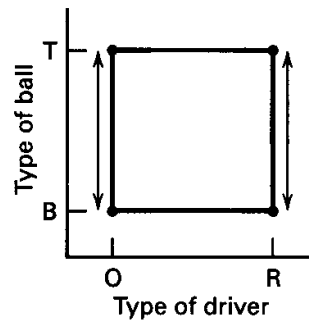
# Factorial Design



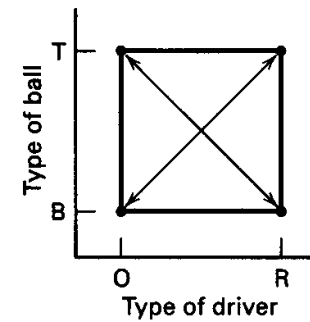
(a) Scores from the golf experiment



(b) Comparison of scores leading to the driver effect



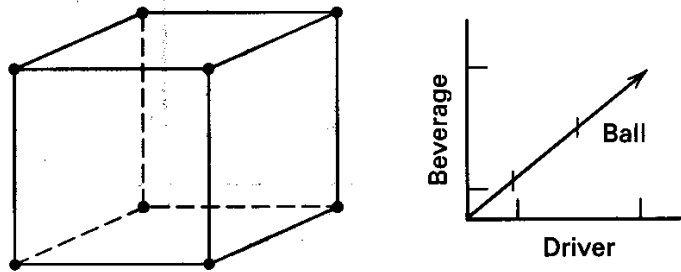
(c) Comparison of scores leading to the ball effect



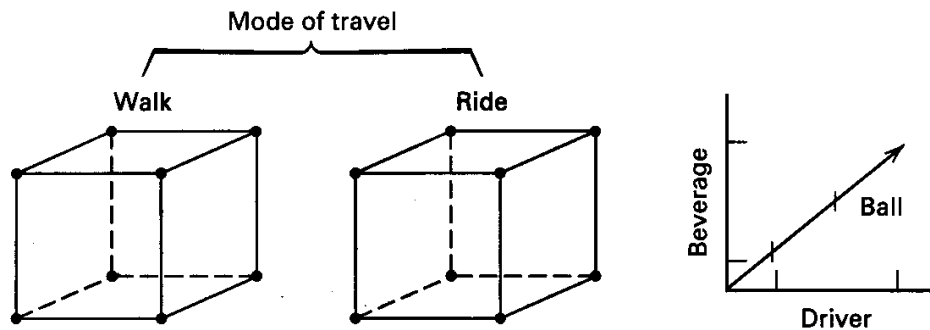
(d) Comparison of scores leading to the ball-driver interaction effect

Figure 1-5 Scores from the golf experiment in Figure 1-4 and calculation of the factor effects.

# Factorial Designs with Several Factors



**Figure 1-6** A three-factor factorial experiment involving type of driver, type of ball, and type of beverage.



**Figure 1-7** A four-factor factorial experiment involving type of driver, type of ball, type of beverage, and mode of travel.

## Erroneous Impressions About Factorial Experiments

- Wasteful and do not compensate the extra effort with additional useful information - this folklore presumes that one knows (not assumes) that factors independently influence the responses (i.e. there are no factor interactions) and that each factor has a linear effect on the response - almost any reasonable type of experimentation will identify optimum levels of the factors
- Information on the factor effects becomes available only after the entire experiment is completed. Takes too long. Actually, factorial experiments can be blocked and conducted sequentially so that data from each block can be analyzed as they are obtained.

## One-factor-at-a-time experiments (OFAT)

- OFAT is a prevalent, but potentially disastrous type of experimentation commonly used by many engineers and scientists in both industry and academia.
- Tests are conducted by systematically changing the levels of one factor while holding the levels of all other factors fixed. The “optimal” level of the first factor is then selected.
- Subsequently, each factor in turn is varied and its “optimal” level selected while the other factors are held fixed.

## One-factor-at-a-time experiments (OFAT)

- OFAT experiments are regarded as easier to implement, more easily understood, and more economical than factorial experiments. Better than trial and error.
- OFAT experiments are believed to provide the optimum combinations of the factor levels.
- Unfortunately, each of these presumptions can generally be shown to be false except under very special circumstances.
- The key reasons why OFAT should not be conducted except under very special circumstances are:
  - *Do not provide adequate information on interactions*
  - *Do not provide efficient estimates of the effects*

## Factorial vs OFAT ( 2-levels only)

### Factorial

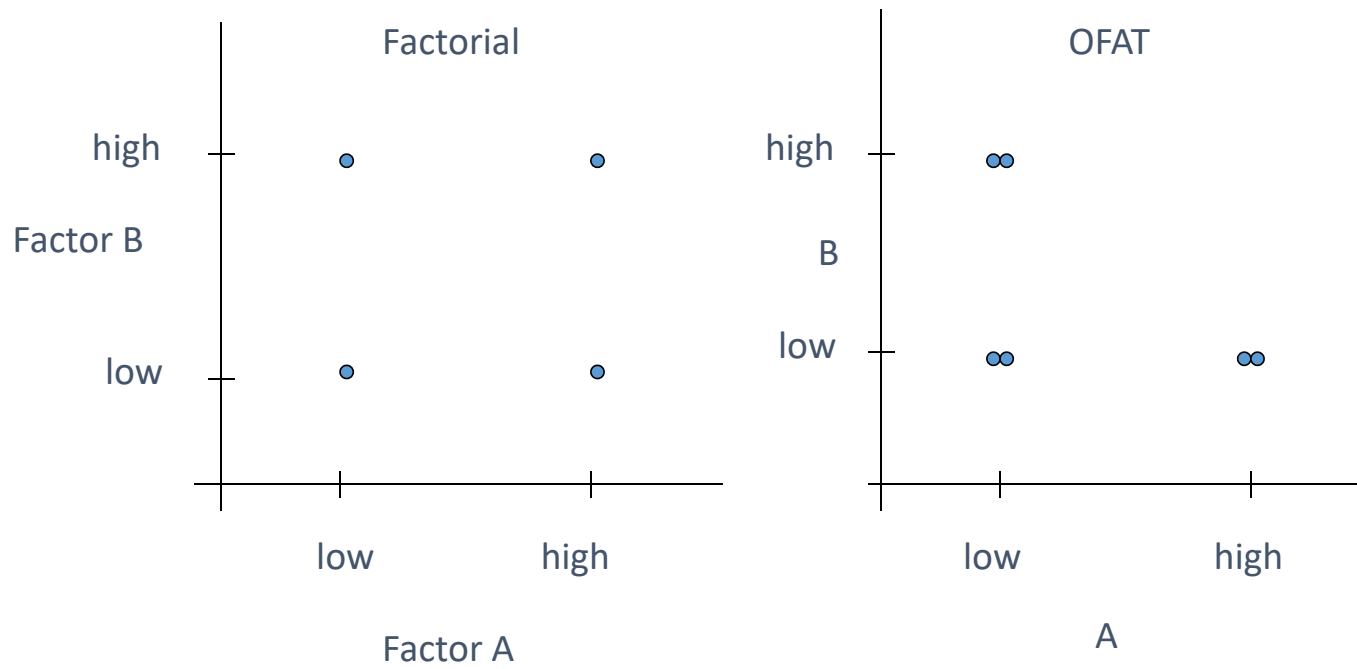
- 2 factors: 4 runs
  - 3 effects
- 3 factors: 8 runs
  - 7 effects
- 5 factors: 32 or 16 runs
  - 31 or 15 effects
- 7 factors: 128 or 64 runs
  - 127 or 63 effects

### OFAT

- 2 factors: 6 runs
  - 2 effects
- 3 factors: 16 runs
  - 3 effects
- 5 factors: 96 runs
  - 5 effects
- 7 factors: 512 runs
  - 7 effects



## Example: Factorial vs OFAT



E.g. Factor A: Reynold's number, Factor B:  $k/D$

Example: Effect of Re and k/D on friction factor f

- Consider a 2-level factorial design ( $2^2$ )
- Reynold's number = Factor A; k/D = Factor B
- Levels for A:  $10^4$  (low)       $10^6$  (high)
- Levels for B: 0.0001 (low)    0.001 (high)
- Responses: (1) = 0.0311,    a = 0.0135,    b = 0.0327,  
ab = 0.0200
- Effect (A) = -0.66, Effect (B) = 0.22, Effect (AB) = 0.17
- % contribution: A = 84.85%, B = 9.48%, AB = 5.67%
- The presence of interactions implies that one cannot satisfactorily describe the effects of each factor using main effects.

DESIGN-EASE Plot

$\ln(f)$

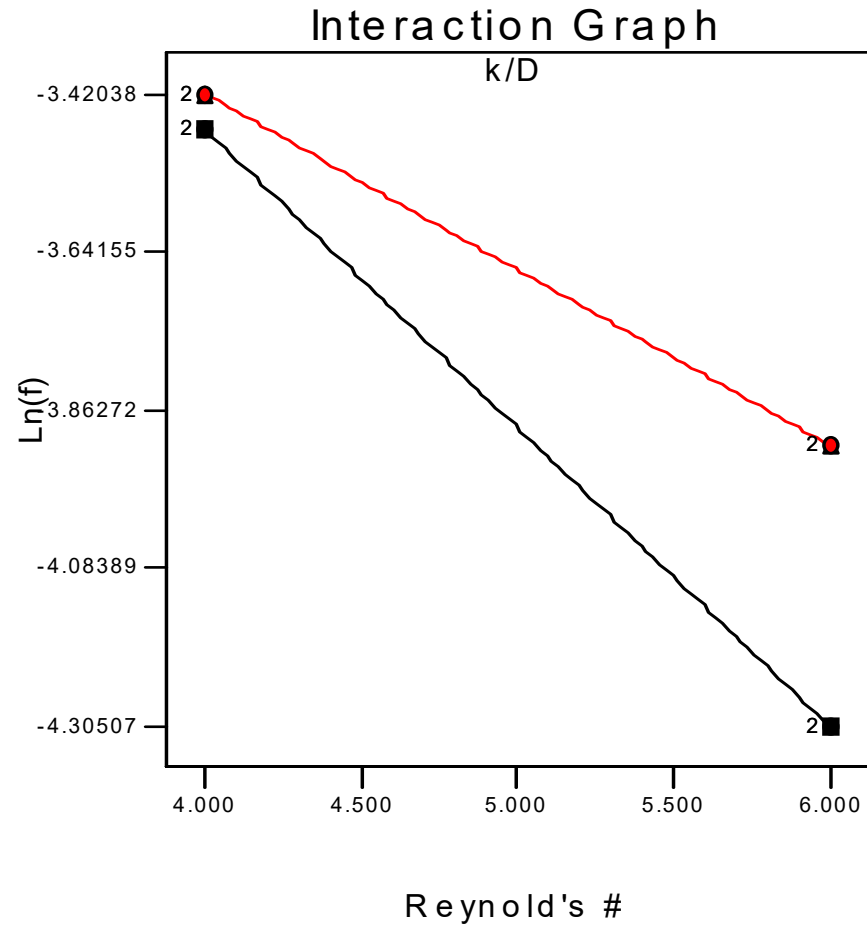
X = A: Reynold's #

Y = B: k/D

● Design Points

■ B - 0.000

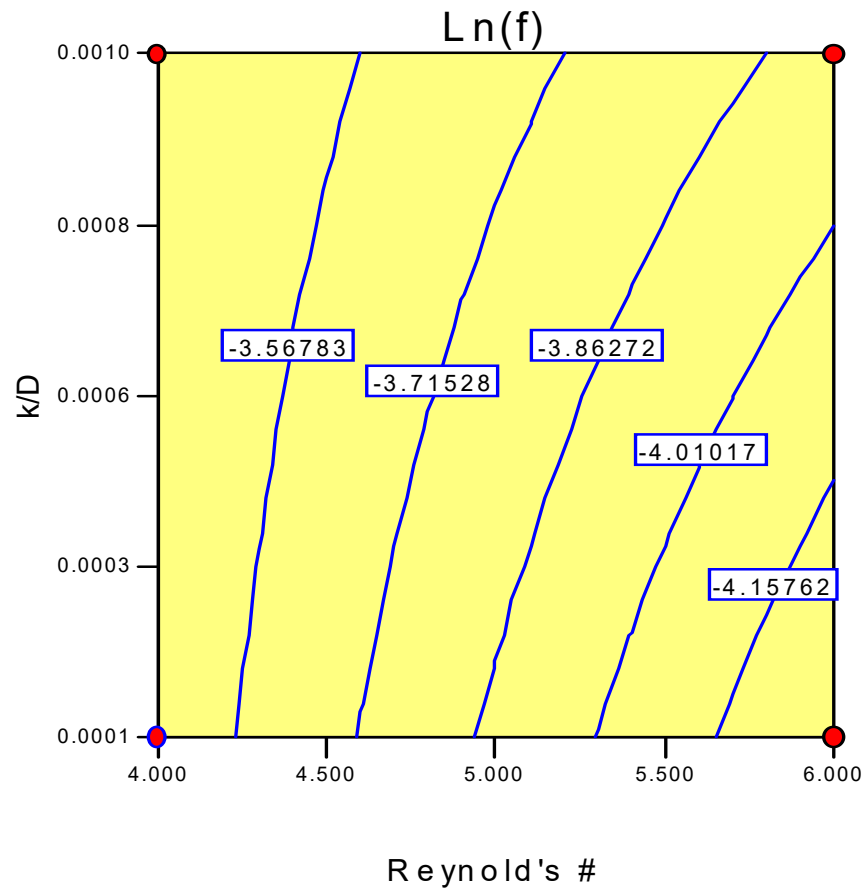
▲ B + 0.001



DESIGN-EASE Plot

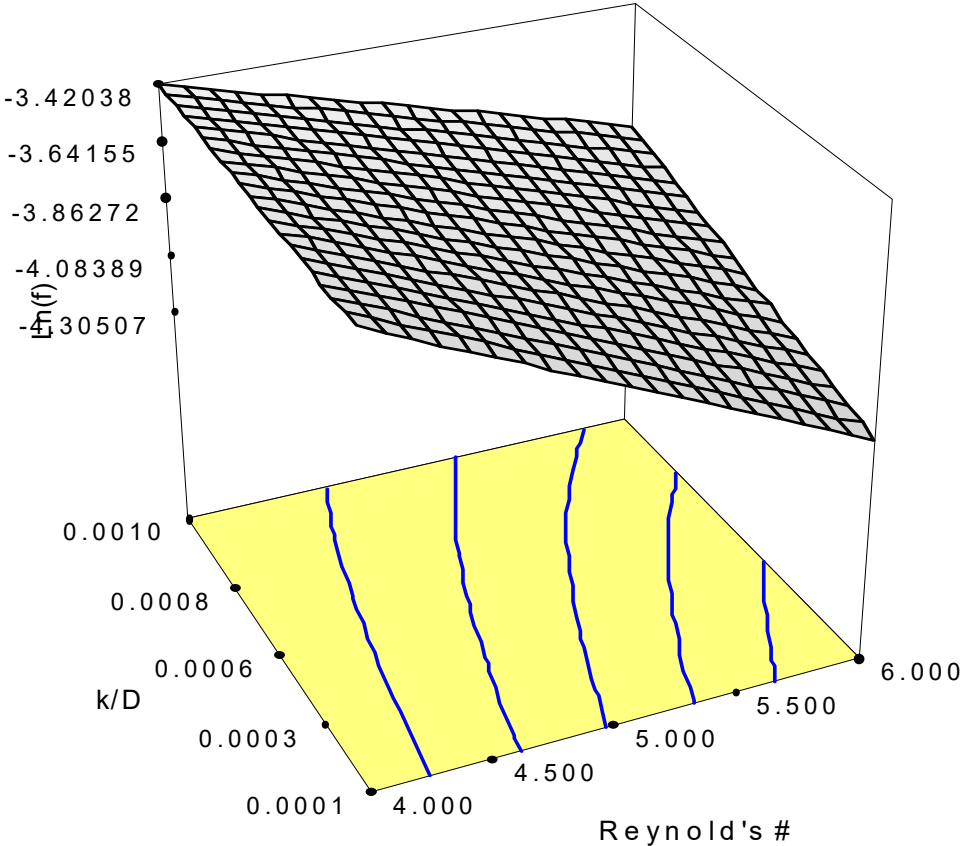
Ln(f)  
X = A: Reynold's #  
Y = B: k/D

● Design Points



DESIGN-EASE Plot

Ln(f)  
X = A: Reynold's #  
Y = B: k/D



With the addition of a few more points

- Augmenting the basic  $2^2$  design with a center point and 5 axial points we get a central composite design (CCD) and a 2nd order model can be fit.
- The nonlinear nature of the relationship between  $Re$ ,  $k/D$  and the friction factor  $f$  can be seen.
- If Nikuradse (1933) had used a factorial design in his pipe friction experiments, he would need far less experimental runs!!
- If the number of factors can be reduced by dimensional analysis, the problem can be made simpler for experimentation.

DESIGN-EXPERT Plot

Log10(f)

X = A: RE

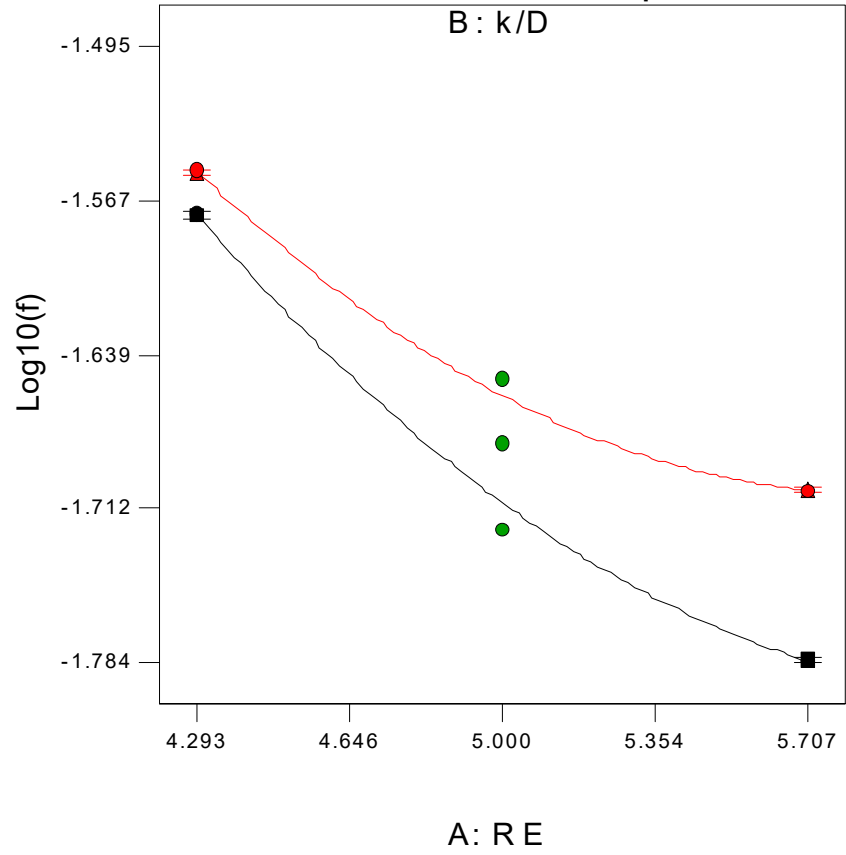
Y = B: k/D

● Design Points

■ B- 0.000

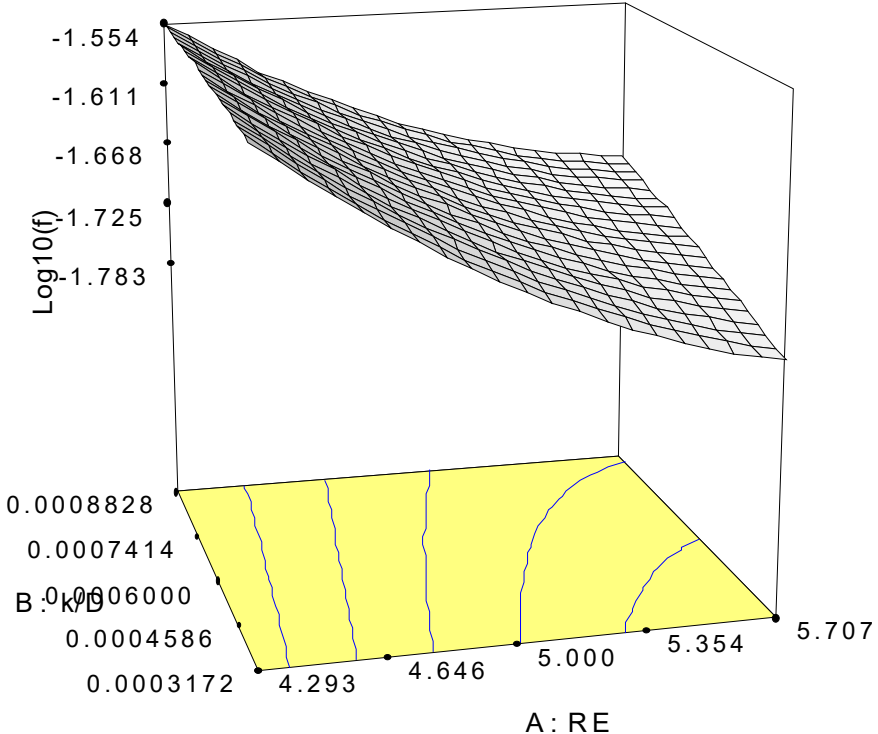
▲ B+ 0.001

### Interaction Graph



DESIGN-EXPERT Plot

Log10(f)  
X = A : RE  
Y = B : k/D





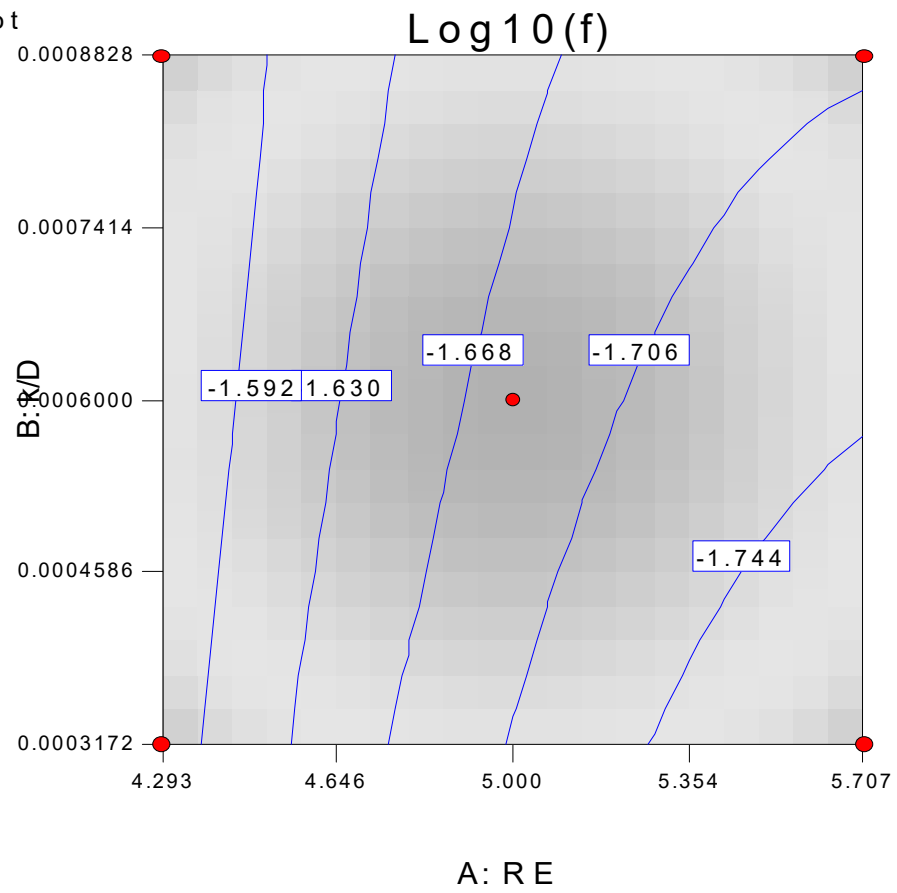
DESIGN-EXPERT Plot

Log 10(f)

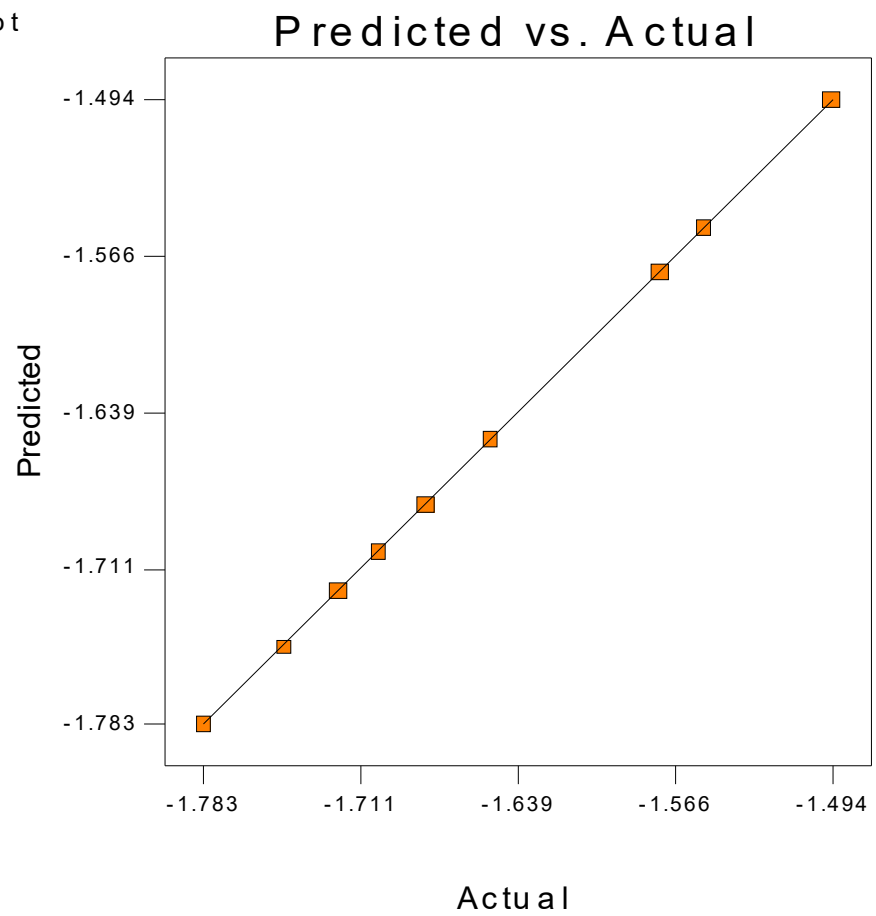
● Design Points

X = A: RE

Y = B: k/D



DESIGN-EXPERT Plot  
Log10(f)



## FACTORIAL ( $2^k$ ) DESIGNS

- Experiments involving several factors (  $k$  = # of factors) where it is necessary to study the joint effect of these factors on a specific response.
- Each of the factors are set at two levels (a “low” level and a “high” level) which may be qualitative (machine A/machine B, fan on/fan off) or quantitative (temperature  $80^{\circ}$ /temperature  $90^{\circ}$ , line speed 4000 per hour/line speed 5000 per hour).

## FACTORIAL ( $2^k$ ) DESIGNS

- Factors are assumed to be fixed (fixed effects model)
- Designs are completely randomized (experimental trials are run in a random order, etc.)
- The usual normality assumptions are satisfied.

## FACTORIAL ( $2^k$ ) DESIGNS

- Particularly useful in the early stages of experimental work when you are likely to have many factors being investigated and you want to minimize the number of treatment combinations (sample size) but, at the same time, study all  $k$  factors in a complete factorial arrangement (the experiment collects data at all possible combinations of factor levels).

## FACTORIAL ( $2^k$ ) DESIGNS

- As  $k$  gets large, the sample size will increase exponentially. If experiment is replicated, the # runs again increases.

$k$	# of runs
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024

## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- Two factors set at two levels (normally referred to as low and high) would result in the following design where each level of factor A is paired with each level of factor B.

Generalized Settings			
RUN	Factor A	Factor B	RESPONSE
1	low	low	$y_1$
2	high	low	$y_2$
3	low	high	$y_3$
4	high	high	$y_4$

Orthogonal Settings			
RUN	Factor A	Factor B	RESPONSE
1	-1	-1	$y_1$
2	+1	-1	$y_2$
3	-1	+1	$y_3$
4	+1	+1	$y_4$

## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- Estimating main effects associated with changing the level of each factor from low to high. This is the estimated effect on the response variable associated with changing factor A or B from their low to high values.

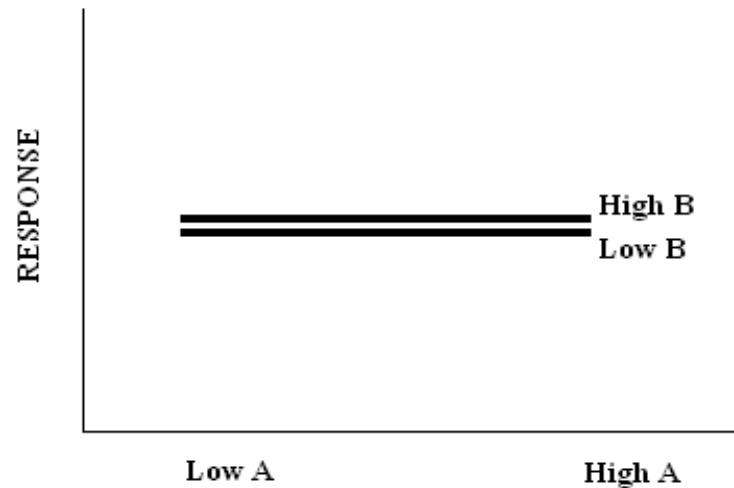
$$\textit{Factor A Effect} = \frac{(y_2 + y_4)}{2} - \frac{(y_1 + y_3)}{2}$$

$$\textit{Factor B Effect} = \frac{(y_3 + y_4)}{2} - \frac{(y_1 + y_2)}{2}$$



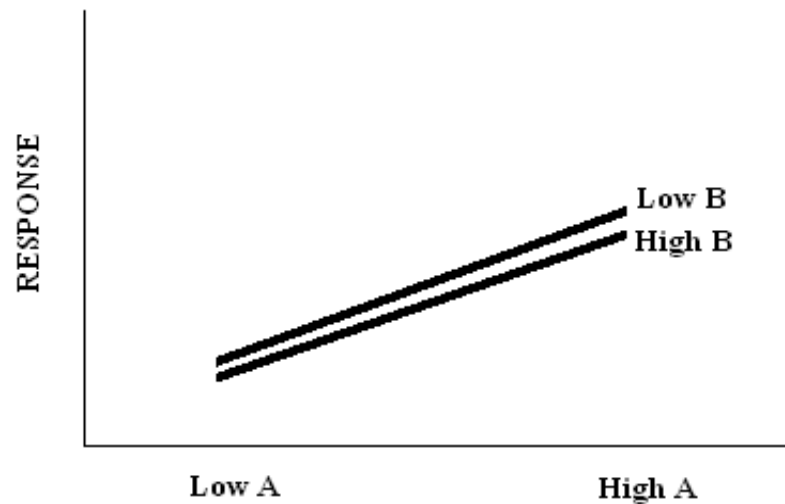
## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ ): GRAPHICAL OUTPUT

- Neither factor A nor Factor B have an effect on the response variable.



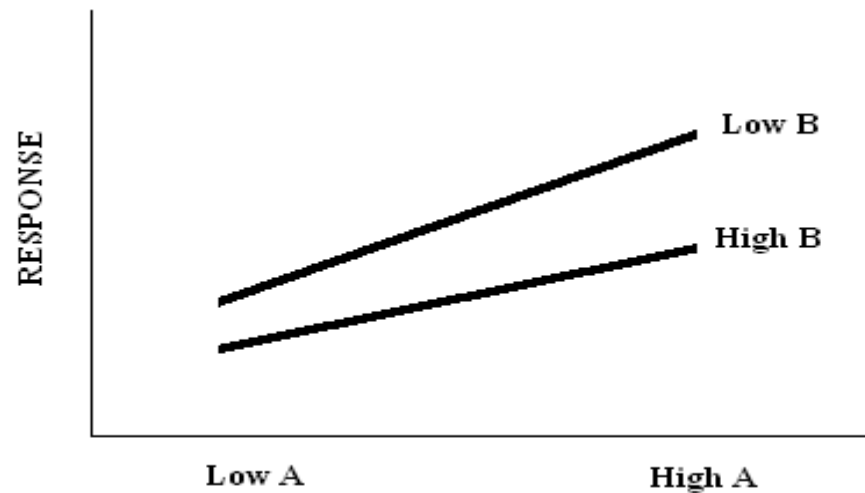
## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ ): GRAPHICAL OUTPUT

- Factor A has an effect on the response variable, but Factor B does not.



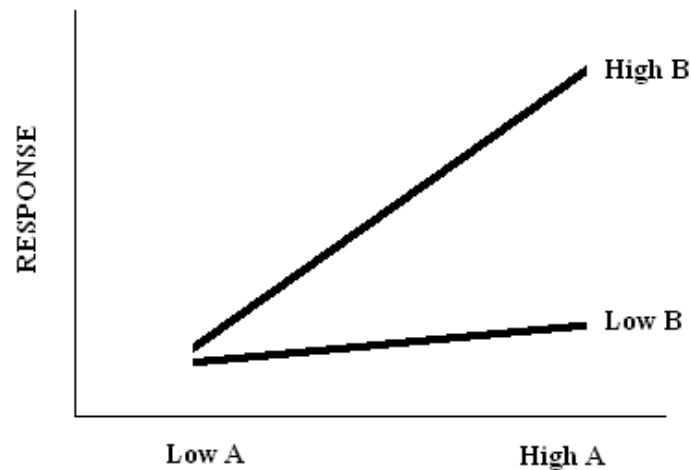
## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ ): GRAPHICAL OUTPUT

- Factor A and Factor B have an effect on the response variable.



## FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ ): GRAPHICAL OUTPUT

- Factor B has an effect on the response variable, but only if factor A is set at the “High” level. This is **called interaction** and it basically means that the effect one factor has on a response is dependent on the level you set other factors at. Interactions can be major problems in a DOE if you fail to account for the interaction when designing your experiment.



EXAMPLE:  
FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- A microbiologist is interested in the effect of two different culture mediums [medium 1 (low) and medium 2 (high)] and two different times [10 hours (low) and 20 hours (high)] on the growth rate of a particular CFU [Bugs].

EXAMPLE:  
FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- Since two factors are of interest,  $k = 2$ , and we would need the following four runs resulting in

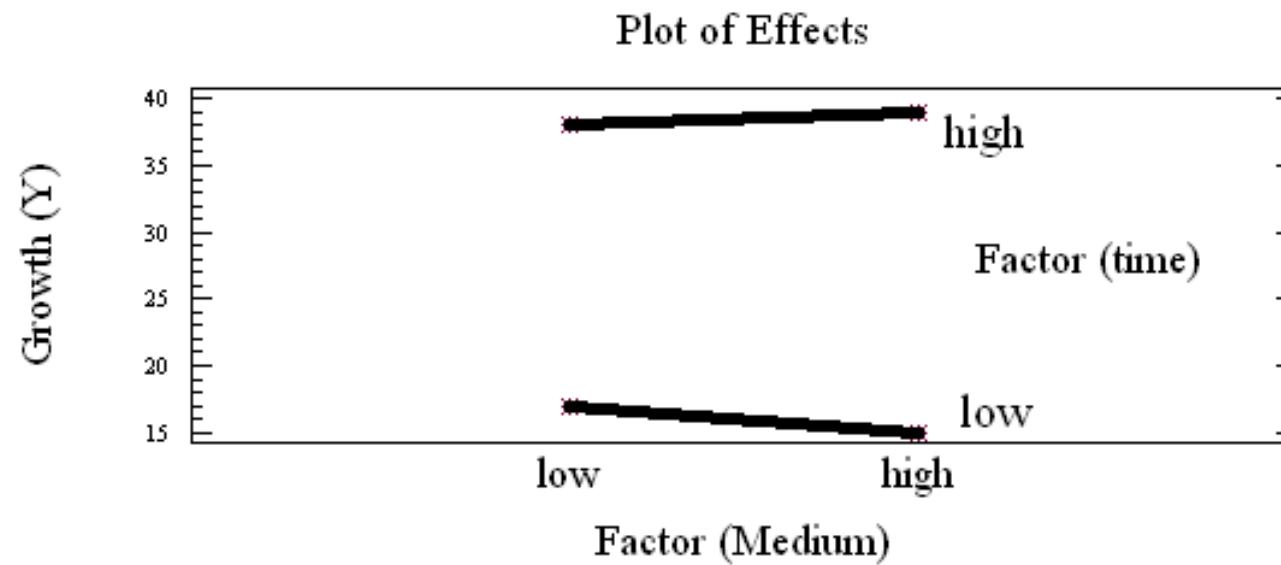
Generalized Settings			
RUN	Medium	Time	Growth Rate
1	low	low	17
2	high	low	15
3	low	high	38
4	high	high	39

EXAMPLE:

FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- Estimates for the medium and time effects are
- Medium effect =  $[(15+39)/2] - [(17 + 38)/2] = -0.5$
- Time effect =  $[(38+39)/2] - [(17 + 15)/2] = 22.5$

EXAMPLE:  
FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )





## EXAMPLE: FACTORIAL ( $2^k$ ) DESIGNS ( $k = 2$ )

- A statistical analysis using the appropriate statistical model would result in the following information. Factor A (medium) and Factor B (time)

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
FACTOR A	0.25	1	0.25	0.11	0.7952
FACTOR B	506.25	1	506.25	225.00	0.0424
Residual	2.25	1	2.25		
Total (corrected)	508.75	3			

All F-ratios are based on the residual mean square error.

## EXAMPLE: CONCLUSIONS

- In statistical language, one would conclude that factor A (medium) is not statistically significant at a 5% level of significance since the p-value is greater than 5% (0.05), but factor B (time) is statistically significant at a 5 % level of significance since this p-value is less than 5%.

## EXAMPLE: CONCLUSIONS

- In layman terms, this means that we have no evidence that would allow us to conclude that the medium used has an effect on the growth rate, although it may well have an effect (our conclusion was incorrect).

## EXAMPLE: CONCLUSIONS

- Additionally, we have evidence that would allow us to conclude that time does have an effect on the growth rate, although it may well not have an effect (our conclusion was incorrect).

## EXAMPLE: CONCLUSIONS

- In general we control the likelihood of reaching these incorrect conclusions by the selection of the level of significance for the test and the amount of data collected (sample size).

## $2^k$ DESIGNS ( $k \geq 2$ )

- As the number of factors increase, the number of runs needed to complete a complete factorial experiment will increase dramatically. The following  $2^k$  design layout depict the number of runs needed for values of  $k$  from 2 to 5. For example, when  $k = 5$ , it will take  $2^5 = 32$  experimental runs for the complete factorial experiment.

## Interactions for $2^k$ Designs ( $k = 3$ )

- Interactions between various factors can be estimated for different designs above by multiplying the appropriate columns together and then subtracting the average response for the lows from the average response for the highs.

## Interactions for $2^k$ Designs ( $k = 3$ )

<b>a</b>	<b>b</b>	<b>c</b>	<b>ab</b>	<b>ac</b>	<b>bc</b>	<b>abc</b>
-1	-1	-1	1	1	1	-1
+1	-1	-1	-1	-1	1	1
-1	+1	-1	-1	1	-1	1
+1	+1	-1	1	-1	-1	-1
-1	-1	+1	1	-1	-1	1
+1	-1	+1	-1	1	-1	-1
-1	+1	+1	-1	-1	1	-1
+1	+1	+1	1	1	1	1



## $2^k$ DESIGNS ( $k \geq 2$ )

- Once the effect for all factors and interactions are determined, you are able to develop a prediction model to estimate the response for specific values of the factors. In general, we will do this with statistical software, but for these designs, you can do it by hand calculations if you wish.

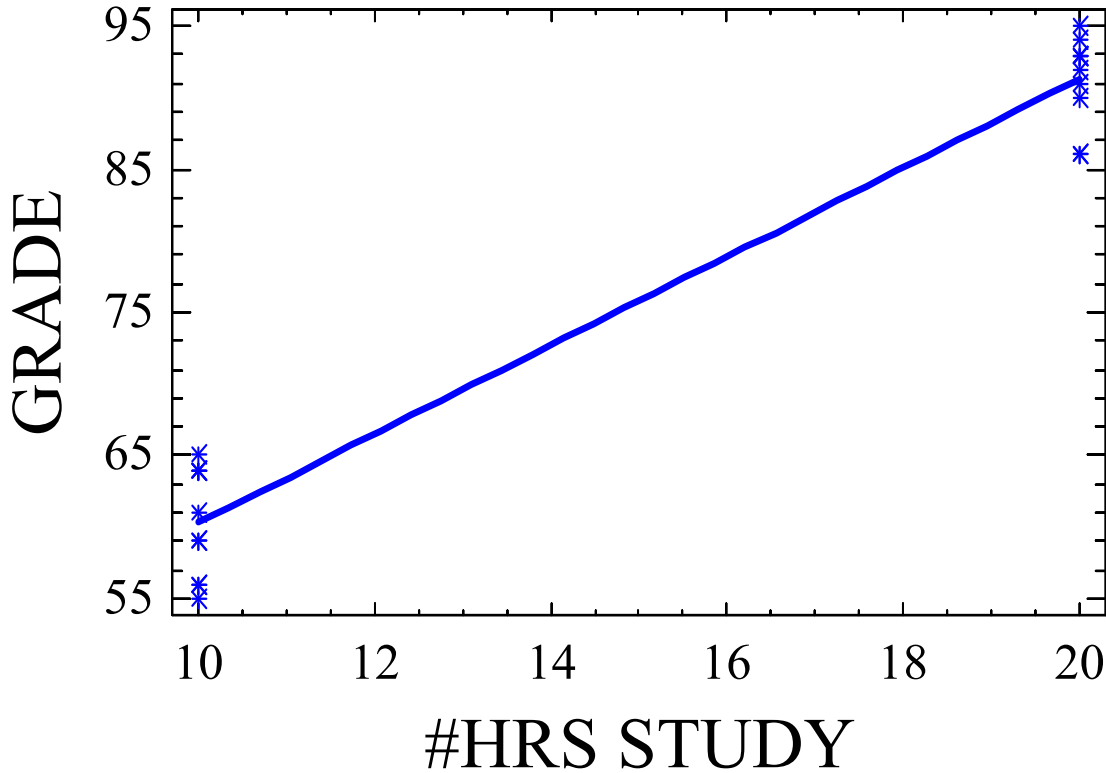
## $2^k$ DESIGNS ( $k \geq 2$ )

- For example, if there are no significant interactions present, you can estimate a response by the following formula. (for quantitative factors only)

$$Y = (\text{average of all responses}) + \sum \left[ \left( \frac{\text{factorEFFECT}}{2} \right) * (\text{factorLEVEL}) \right]$$
$$= \bar{Y} + \left( \frac{\Delta_A}{2} \right) * A + \left( \frac{\Delta_B}{2} \right) * B$$

# ONE FACTOR EXAMPLE

## Plot of Fitted Model



## ONE FACTOR EXAMPLE

- The output shows the results of fitting a general linear model to describe the relationship between GRADE and #HRS STUDY. The equation of the fitted general model is
- $\text{GRADE} = 29.3 + 3.1 * (\text{\#HRS STUDY})$
- The fitted orthogonal model is
- $\text{GRADE} = 75 + 15 * (\text{SCALED \# HRS})$

## Two Level Screening Designs

- Suppose that your brainstorming session resulted in 7 factors that various people think “might” have an effect on a response. A full factorial design would require  $2^7 = 128$  experimental runs without replication. The purpose of screening designs is to reduce (identify) the number of factors down to the “major” role players with a minimal number of experimental runs. One way to do this is to use the  $2^3$  full factorial design and use interaction columns for factors.

Note that

\* Any factor d effect is now confounded with the a\*b interaction

\* Any factor e effect is now confounded with the a\*c interaction

\* etc.

\* What is the d\*e interaction confounded with?????????

a	b	c	d = ab	e = ac	f = bc	g = abc
-1	-1	-1	1	1	1	-1
+1	-1	-1	-1	-1	1	1
-1	+1	-1	-1	1	-1	1
+1	+1	-1	1	-1	-1	-1
-1	-1	+1	1	-1	-1	1
+1	-1	+1	-1	1	-1	-1
-1	+1	+1	-1	-1	1	-1
+1	+1	+1	1	1	1	1

# Problems that Interactions Cause!

- Interactions – If interactions exist and you fail to account for this, you may reach erroneous conclusions. Suppose that you plan an experiment with four runs and three factors resulting in the following data:

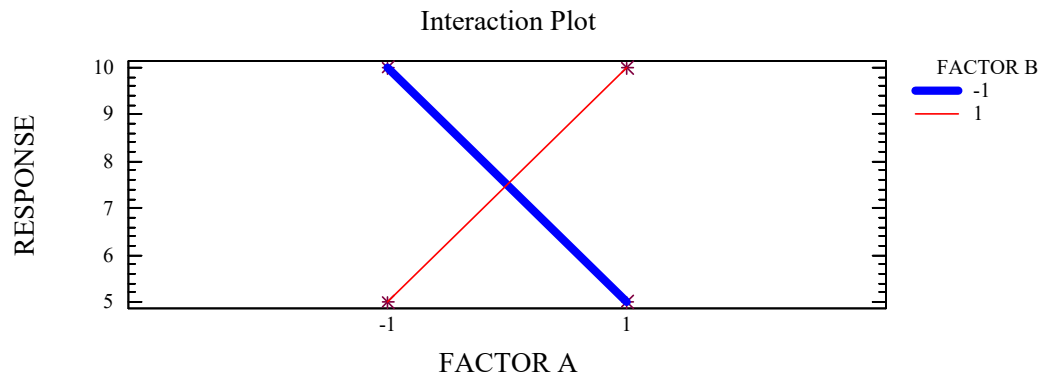
Run	Factor A	Factor B	Results
1	+1	+1	10
2	+1	-1	5
3	-1	+1	5
4	-1	-1	10

# Problems that Interactions Cause!

- Factor A Effect = 0
- Factor B Effect = 0
- In this example, if you were assuming that “smaller is better” then it appears to make no difference where you set factors A and B. If you were to set factor A at the low value and factor B at the low value, your response variable would be larger than desired. In this case there is a factor A interaction with factor B.



# Problems that Interactions Cause!



# Resolution of a Design

- Resolution III Designs – No main effects are aliased with any other main effect BUT some (or all) main effects are aliased with two way interactions
- Resolution IV Designs – No main effects are aliased with any other main effect OR two factor interaction, BUT two factor interactions may be aliased with other two factor interactions
- Resolution V Designs – No main effect OR two factor interaction is aliased with any other main effect or two factor interaction, BUT two factor interactions are aliased with three factor interactions.

## Common Screening Designs

- Fractional Factorial Designs – the total number of experimental runs must be a power of 2 (4, 8, 16, 32, 64, ...). If you believe first order interactions are small compared to main effects, then you could choose a resolution III design. Just remember that if you have major interactions, it can mess up your screening experiment.

## Common Screening Designs

- Plackett-Burman Designs – Two level, resolution III designs used to study up to  $n-1$  factors in  $n$  experimental runs, where  $n$  is a multiple of 4 ( # of runs will be 4, 8, 12, 16, ...). Since  $n$  may be quite large, you can study a large number of factors with moderately small sample sizes. ( $n = 100$  means you can study 99 factors with 100 runs)

## Other Design Issues

- May want to collect data at center points to estimate non-linear responses
- More than two levels of a factor – no problem (multi-level factorial)
- What do you do if you want to build a non-linear model to “optimize” the response. (hit a target, maximize, or minimize) – called response surface modeling

## Response Surface Designs – Box-Behnken

RUN	F1	F2	F3	Y <sub>100</sub>
1	10	45	60	11825
2	30	45	40	8781
3	20	30	40	8413
4	10	30	50	9216
5	20	45	50	9288
6	30	60	50	8261
7	20	45	50	9329
8	30	45	60	10855
9	20	45	50	9205
10	20	60	40	8538
11	10	45	40	9718
12	30	30	50	11308
13	20	60	60	10316
14	10	60	50	12056
15	20	30	60	10378

## Response Surface Designs – Box-Behnken

Regression coeffs. for Var\_3

-----

constant	=	2312.5
A:Factor_A	=	36.575
B:Factor_B	=	200.067
C:Factor_C	=	3.85
AA	=	9.09875
AB	=	-9.81167
AC	=	-0.0825
BB	=	0.117222
BC	=	-0.311667
CC	=	1.10875

## Response Surface Designs – Box-Behnken

### Contours of Estimated Response Surface

