INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
INTERNATIONAL CENTRE FOR SCIENCE AND HIGH TECHNOLOGY
© INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS 34100 TRIESTE (ITALY) VIA GRIGNANO, 9 (ADRIATICO PALACE) P.O. BOX 586 TELEPHONE 040-224572 TELEFAX 040-224575 TELEX 460449 APH I

SMR/474 - 11

COLLEGE ON
"THE DESIGN OF REAL-TIME CONTROL SYSTEMS"
1 - 26 October

ANALYSIS AND VISUALIZATION OF RESULTS

P. BARTHOLDI
Observatory of Geneva
CH-12900 Sauverny
Switzerland

Exploitation and visualization of data

Paul Bartholdi
Observatory of Geneva
CH-1290 Sauverny
Switzerland

Preliminary notes - Trieste - october 1990
College on "Design of Real-Time Control Systems"

Contents

bartho@obs.unige.ch bartho@cgeuge54.bitnet 20579::ugobs::bartho
tel +41 22 755 26 11 fax +41 22 755 39 83

1 Introduction

These lectures will be like a bag of partially unrelated tricks about what to do with data just acquired in real time.

We will cover the following topics:

- Self describing data file. Real time data acquisition systems produce a lot of informations, part of which will be archived for later retrival and processing. If the file contains a description preamble, its later use will be much easier.
- Arithmetic, both with integers and floating points. Although it is easier to work with floating point numbers, operations on them are slower, and this may not be acceptable in real time systems.

  Using the lectures on floating points of the previous College, it will be remainded how to avoid some of pitfalls of computer arithmetic.
- Basic statistics, on unordered data, mean, variance, skewness, kurtosis, and how they can be applied to check the quality of data.
- Time series, tests of stability, (basic Fourrier analysis).
- Data visualization, tables, graphics, scaling effects.
- Real time controls, moving averages, abnormalities, alarms, risk of false alarm, risk of missing an error.

2 Self descriptive Data File

Any scientist making experiments, any one making regular observations, keeps a log book of his measurements. This book may also contain notes concerning any change in the procedures and environement. This is very useful to retrieve older values, do extra processing, compare old and new etc.

Today, many data acquisition systems still archive the row data alone, while the user has to keep a manuscript log of all other parameters. This not only gives extra work to the user, with risks of errors, but also preclude the easy use of these informations by later programs.

We shall consider all data files made of three parts:

1. One or more rectangular tables containing the data in a homogeneous form, for example each line concerning an event, each column a different source (This is also the basic structure of Relational Data Bases).
2. A table description, giving all general informations concerning the data and their gathering, dates, instruments, users, operators, weather conditions, number of columns and rows etc. and for each column: its name, limits, mean, min, max, etc.

   This can best be done with *keywords*, each line starting with a main keyword defining the content of the rest of the line.
3. A data manipulation history, where any subsequent operation on the file (statistical analysis, data reduction, corrections etc.) is recorded with date, operator, procedure(s) used, columns, row or file newly created, etc. using similar keywords as before.

For practical reasons, part 2 and 3 above should come either at the begining of the file (they contain information necessary to read part 1) or in a separate file with the same name but a different suffix.

In either case, these two parts should be written in a very standardized format, like blocks of 50 lines of 80 ascii characters each, in such a way that they can be read without any knowledge of the content of the file.

Then data analysis programs can adapt themselves easily.

Finally, a file containing such a good description of its contents is easy to send to other partners around the world without any long description.

The "FITS" format used by astronomers make it very easy to transfer all kind of information from single observation to large set of images, lists of stars etc.

Reference for FITS (Flexible Image Transport System):

Wells, D. C. et al., Astron. Astrophys. Suppl. 44, 363 (1981)

3 Arithmetic

The notes concerning arithmetic from the previous College on Microprocessors will not be repeated here. We will only take a small example concerning the evaluation of a polynome under various circonstances and present graphs that also exhibit visualization of data.

The following graphs show the different effects of rounding errors according to the method used to evaluate a polynomial with know roots.

The three methods are :

$$\begin{aligned} p_1 &= a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \\ p_2 &= a_0 + x(a_1 + x(a_2 + x(\cdots))) \\ p_3 &= (x - x_1)(x - x_2)\cdots(x - x_n) \end{aligned}$$

We will use the following roots:

$x_i = 1,2,3,4,5$ giving coefficients $a_i = -120, 274, -225, 85, -15, 1$

$x_i = 2,4,6,8,10$ giving coefficients $a_i = -3840, 4384, -1800, 340, -30, 1$

and $x_i = 1,2,3,4,5,6,7,8,9,10$ giving coefficients $a_i = 3.629e+06, -1.063e+07, 1.275e+07, -8.41e+06, 3.417e+06, -9.021e+05, 1.578e+05, -1.815e+04, 1320, -55, 1$

None of the coefficients are very large, all the roots are well separated but also in a limited domain. Nevertheless the graphs show how difficult it is to calculate polynomials, and in fact any summation.

The first method is represented with □, the second (Hoerner) with and the last with a continuous line. Some of the effects will appear better if the □ are coloured.

It should be noted that all calculations have been done in single precision floating point arithmetic (All summation should have been done in double precision ...but the demonstration would not have been so clear)

An other little useful, fast and robust algorithm:

you want $D = \sqrt{P^2+Q^2}$, with P and Q possibly very large or small

P^2 and/or Q^2 may overflow or underflow, but not D.

Solution :

```
U = abs(P); V = abs(Q);
if U < V then Swap(U,V)
if U = 0 then Return Q
repeat IterCount times  /* 2 for 6 digits precision
                           3   20 digits
                           4   62          */
{ R = U / V ;   R = R * R ;
  R = R / ( 4 + R );
```

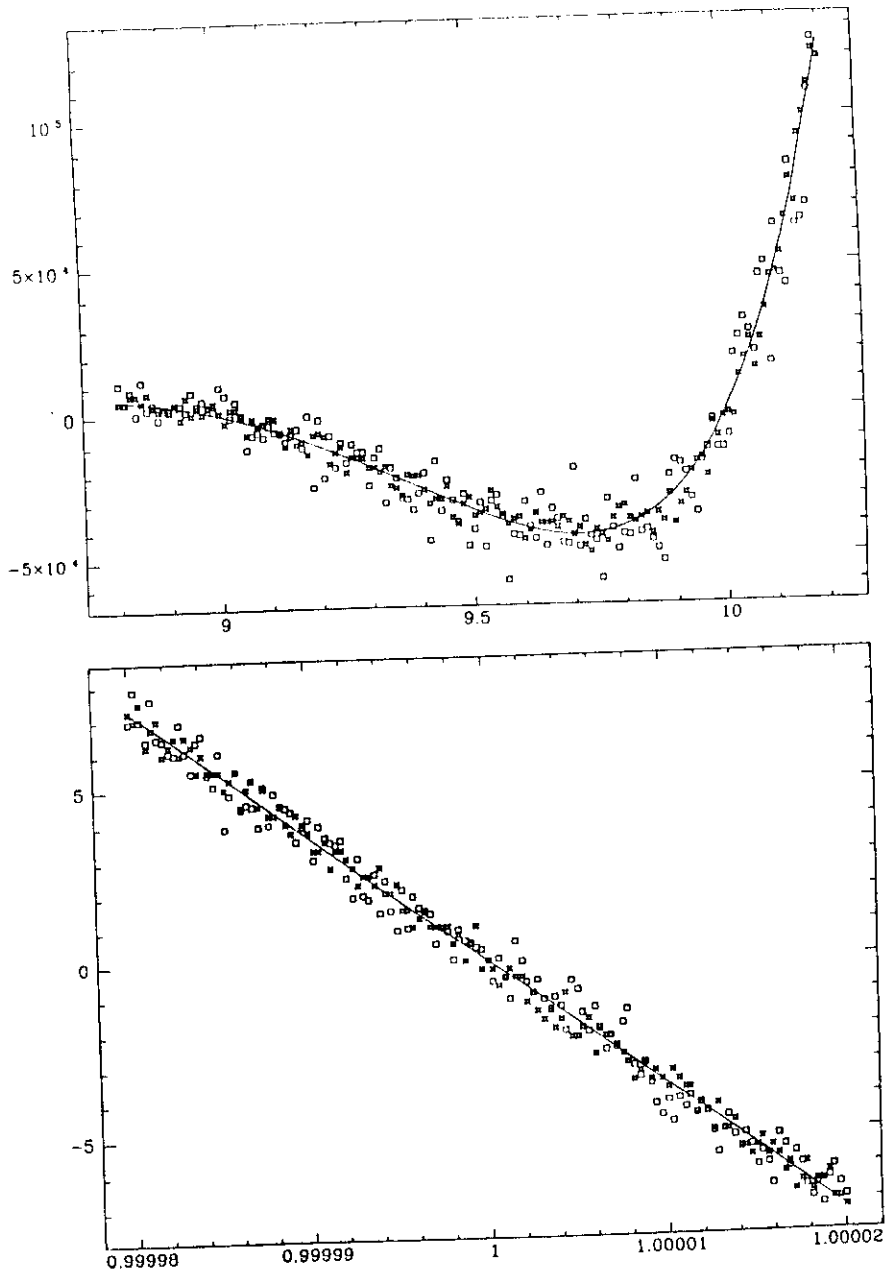



Figure 1: $x_i = \{1,2,3,4,5\}$ with coefficients $a_i = \{-120, 274, -225, 85, -15, 1\}$

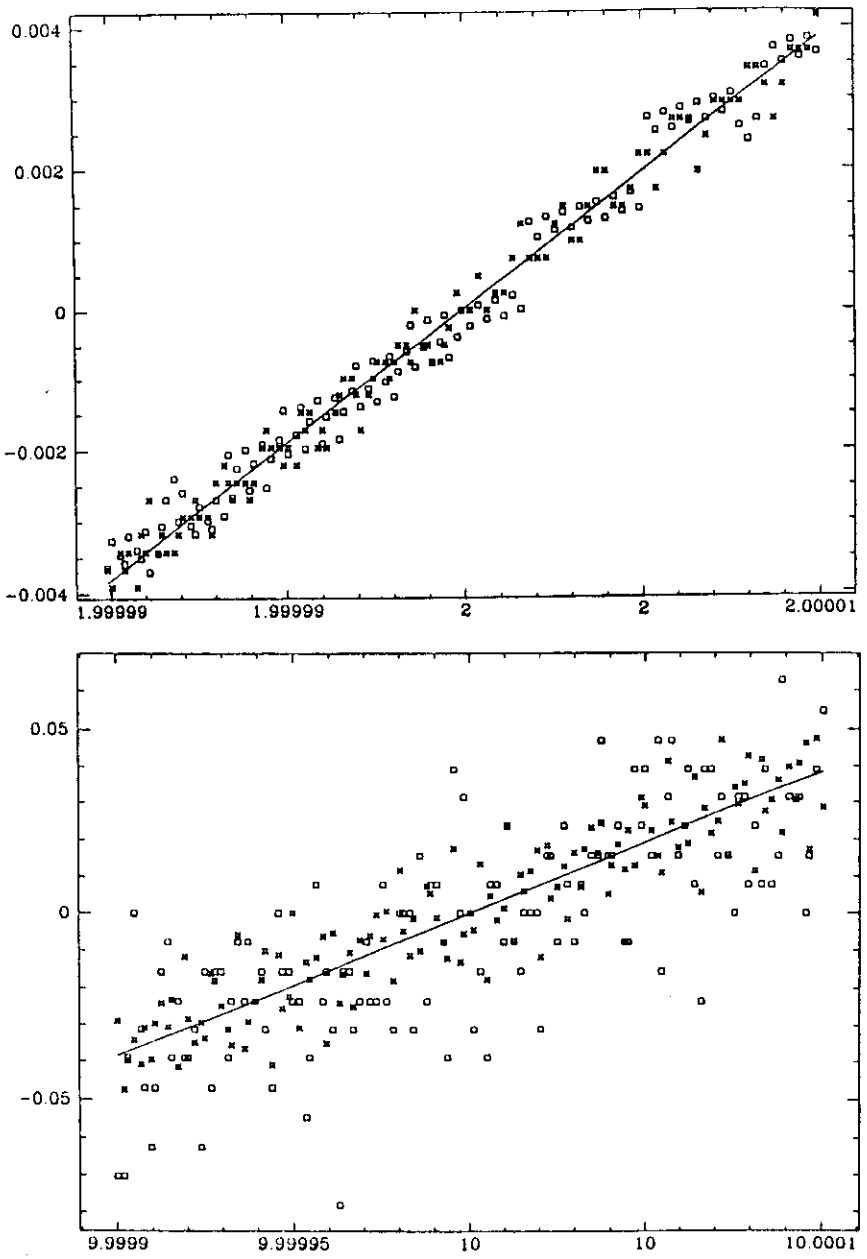


Figure 2: $x_i = \{2, 4, 6, 8, 10\}$ with coefficients $a_i = \{-3840, 4384, -1800, 340, -30, 1\}$

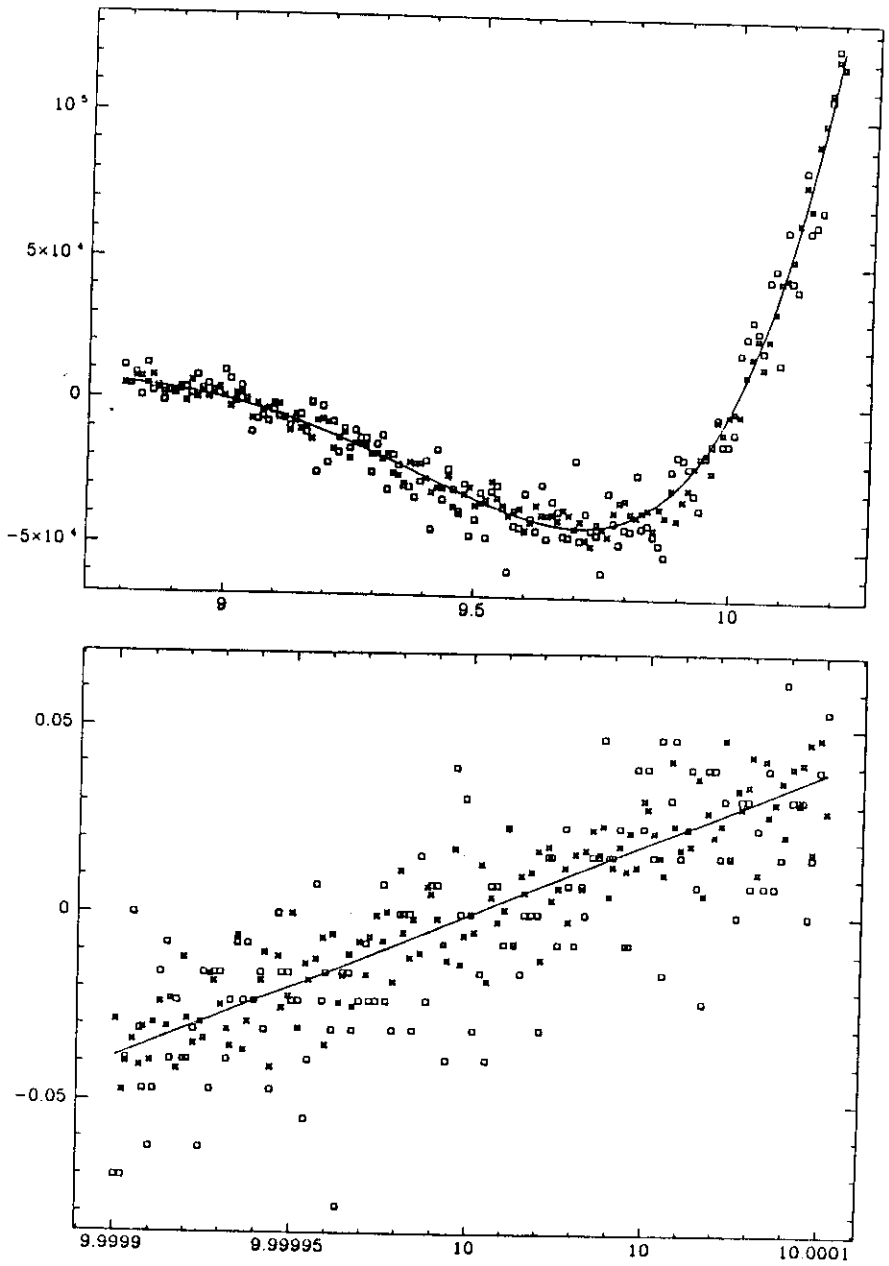


Figure 3: $x_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
with coefficients $a_i = \{3.629e+06, -1.063e+07, 1.275e+07, -8.41e+06, 3.417e+06, -9.021e+05, 1.578e+05, -1.815e+04, 1320, -55, 1\}$

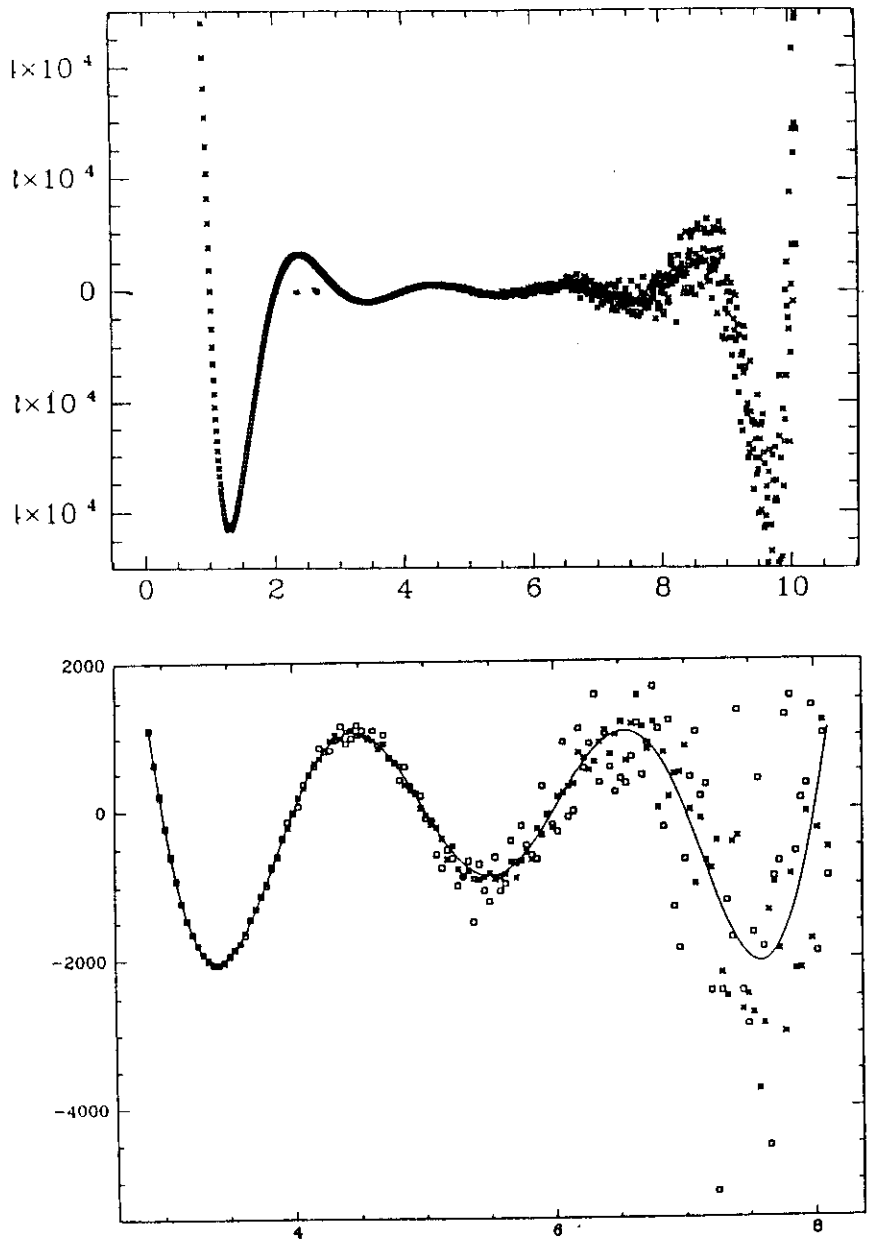


Figure 4: An general view of the last polynome

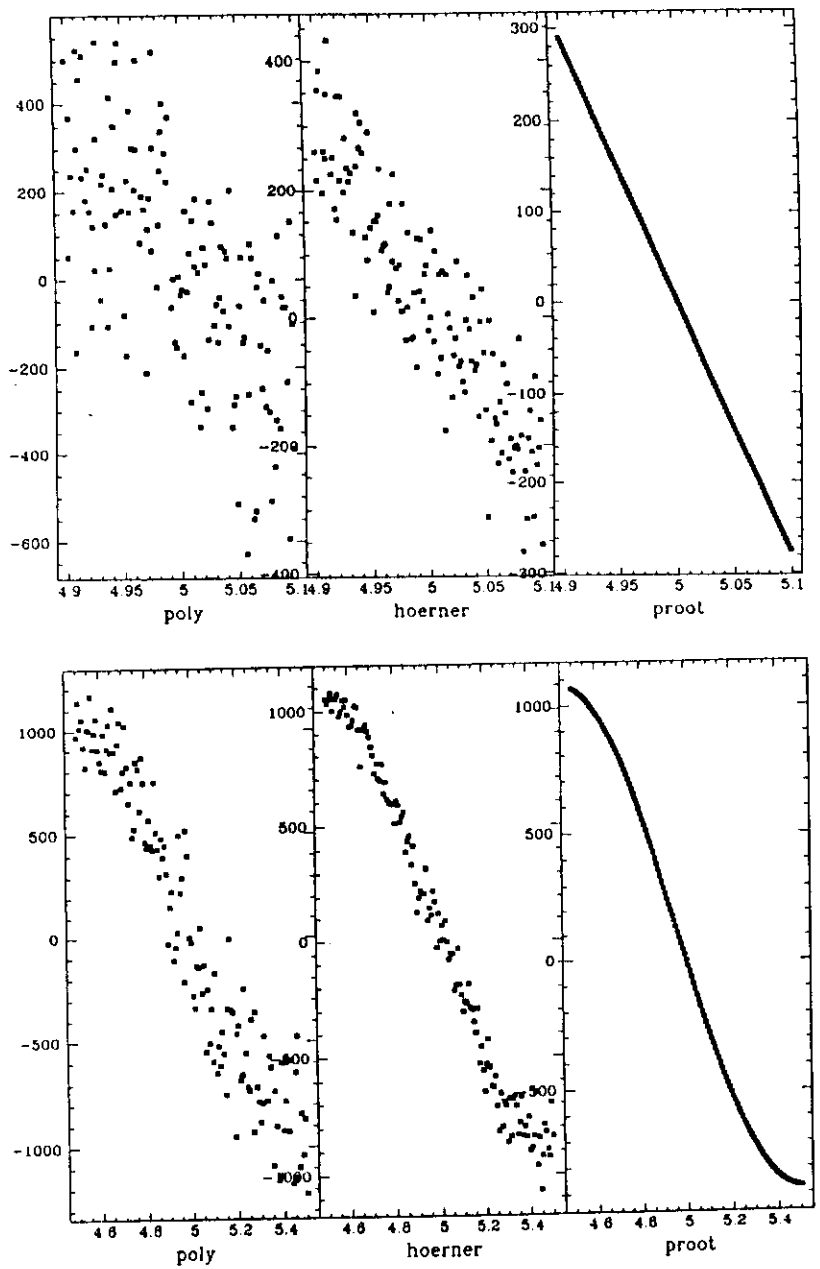


Figure 5: Individual evaluation around $x_i = 10$

4 Basic statistic

Having a set of n data, $\{d_i\}$, we can define the following descriptive parameters:

The *mean* $\overline{d} = \sum d_i/n$.

The *median*, such that there is un equal number of data smaller and larger than the median. (Sort the d_i, then the median $= d_{n/2}$.

The *variance* $\sigma^2 = \sum (d_i - \overline{d})^2/(n-1)$ This is the second order momentum. It is a measure of the dispersion of the data around the mean.

The *skewness* $A = \sum (d_i - \overline{d})^3/(n-1)$ is the third order momentum. Compared to σ^2, it is sensitive to any asymetry around the mean.

The *kurtosis* $K = \sum (d_i - \overline{d})^4/(n-1)$ is the fourth order momentum. Compared to σ^2, it is sensitive to the flatness of the distribution around the mean.

The exact relation between the last three parameters depends on the distribution od the data.

It is usually easier to calculate running sums of the various powers of the data, and then to combine them to get the previous parameters. The individual data doesn't need to stay in memory.

If we have $S_j = \sum d_i^j/n$, then, by simple calculation:

$$\sigma^2 = S_2 - S_1^2/n$$

$$A = S_3 - 3S_1S_2 + 2S_1^3/n$$

$$K = S_4 - 4S_1S_3 + 6S_1^2S_2 - 3S_1^4/n$$

4.1 Algorithms to evaluate the *mean* in real time

We can consider that all parameters above, except the median, are some sort of means. So a good algorithm for the mean can be used for the others too.

Here is list of successive methods with decreasing truncation errors and risk for overflow:

$$\begin{aligned}
M_1 &= \sum d_i/n \\
M_2 &= d_1 + \sum (d_i - d_1)/(n-1) \\
M_3 &= M_n, \quad M_i = [(i-1)/i]\, M_{i-1} + (1/i)d_i, \quad M_0 = 0 \\
M_4 &= M_n, \quad M_i = M_{i-1} + (d_i - M_{i-1})/i, \quad M_0 = 0 \\
M_5 &= M_n, \quad M_i = (d_i/i) + M_{i-1} - (M_{i-1}/i), \quad M_0 = 0 \\
M_{4+} &= M_n, \quad M_i = M_{i-1} + (d_i - M_{i-1} + i/2)/i, \quad M_0 = 0 \\
M_{5+} &= M_n, \quad M_i = (d_i + i/2)/i + M_{i-1} - (M_{i-1} + i/2)/i, \quad M_0 = 0
\end{aligned}$$

The following algorithm is as good as the previous, but specialy adapted to fixed point calculations, and great care has been given to avoid arithmetic overflow.

SHR(*A*, *B*) denotes a shift of the variable *A* to the right *B* bits.
SHL( ) similarly denotes a shift left.

I. Each data point is incorporated into a series of sums as shown in Figure 3. *N* and MAX are initialized at 0.
   A. A new data point is introduced
      $X \leftarrow X_i$; $I \leftarrow 0$; $N \leftarrow N + 1$; $NN \leftarrow N$
   B. As mentioned in Figure 3, the bits of *NN* serve as a guide to the merging process. *X* is an intermediate variable, ultimately placed in the appropriate register of the array *D*( ).
      If *NN* is odd,
         $D(I) \leftarrow X$
         If $I >$ MAX then MAX $\leftarrow I$
         go to *A*.
      If *NN* is even,
         $X \leftarrow$ SHR(*D*(*I*) + *X*, 1)
         $I \leftarrow I + 1$
         $NN \leftarrow$ SHR(*NN*, 1)
         go to *B*.

II. The "Wrap Up": The mean is computed from the intermediate sums. *NN* again serves as a guide. The "ones" in *N* indicate which registers in *D*( ) must be incorporated into the final sum.
   A. If MAX = 0 then *D*(0) is the mean . . . (trivial case)
      $I \leftarrow 0$; $K \leftarrow 0$; MEAN $\leftarrow 0$
   B. Starting with the register whose value encompasses the least number of data points, *D*(0), the contents of the pertinent registers of *D*( ) are weighted and added into the final sum.
      $NN \leftarrow$ SHR(*N*, *I*)
      If *NN* is odd,
         MEAN $\leftarrow$ SHR(MEAN, *K*) . . . (weighting)
         MEAN $\leftarrow$ SHR(SHL(*D*(*I*), 1) + MEAN, 1) . . . (weighting and adding)
         $K \leftarrow 0$
         go to *C*.
      If *NN* is even,
         $K \leftarrow K + 1$
         go to *C*.
   C. $I \leftarrow I + 1$
      If $I <$ MAX then go to *B*
      If $I =$ MAX then MEAN $\cdot 2^{MAX}/N$ is the computed mean

Figure 6: M_6 fixed point algorithm for the mean

5 Time series

We will first look at parameters that give indication of any trend or abnormal data in a serie.

The first test was introduced by the german optician Abbe.

The idea is to compare the variance obtained in the previous section (that do not depends on the order of the data) to the variance obtained by squaring the differences between successive data:

If $s^2 = \sum (d_i - d_{i-1})^2/2n$, then $r = s^2/\sigma^2$ is very sensitive to abnormal behaviour of the data. If they are random and uncorrelated, then $r = 1$. If the data tend to oscillate, then $1 < r < 2$. If slow changes are perturbing the data, then $0 < r < 1$. r is roughly normal, with variance $1/\sqrt{n}$. This test is very insensitive to the distribution of the data.

We can, as in the previous section, keep a running sum of the successive squared difference S_d, and combine this with the other S_j to get s^2 and r, without having to keep all the data in memory.

The second test consist in using the skewness defined before. Any drop of the "signal" will appear as a negative asymetry, any "spike" as a positive one. This test is quite sensitive to the theoretical distribution of the data and the value obtained must be compared to either a theoretical one, or to previously obtained ones.

The third test consist in comparing the observed and theoretical variance (or previously obtained ones). For example, if the d_i consist of counts of random events, then the d_i have a poissonian distribution, with variance $\sigma^2_{\text{theoretical}} = \overline{d}$. We can define $p = \sigma^2_{\text{observed}}/\sigma^2_{\text{theoretical}}$ If $p < 1$, then something is shurely wrong, as the dispersion is smaller than can be expected. If $p > 1$, then some extra "noise" is present in the data. This test is also insensitive to the distribution of the data. p follow roughly a Fisher distribution.

5.1 Frequency domain

Time series are best analyzed in the frequency domain, using Fourrier Analysis. This is out of the scope of these lectures, but a very important point should be noted, concerning the sampling of variable signals.

The Nyquist theorem says that there should be at least two samples for the shortest period, corresponding to the highest frequency, present in the signal and the noise. Otherwise, all kind of strange effects (frequency fording) will take place. If necessary, a low pass filter should be placed in front of the sampler. In practice, 3-5 samples/period are suggested.

6 Data presentation

Some principles:

- The author of a graph or table should understand the application well enough to summarize the important data and avoid unnecessary details.
- Graphs and tables can portray a complex set of relationships simply and clearly. The reader should use his vision system more than his cognitive skills. Structures and relationships are more important than details.
- They should be accurate, not only numerocally but also in quality. The reader should get the true meaning of the data, and not missinterpret them (see the effect of linear or logarithmic scale for example).
- They should be attractive, but for the data, not for themselves. Start with simple design, verify that the effects seen are significant, and use eraser for superfluous ink ...

The examples bellow show how tables and graphs can be improved, and some effects of scaling.

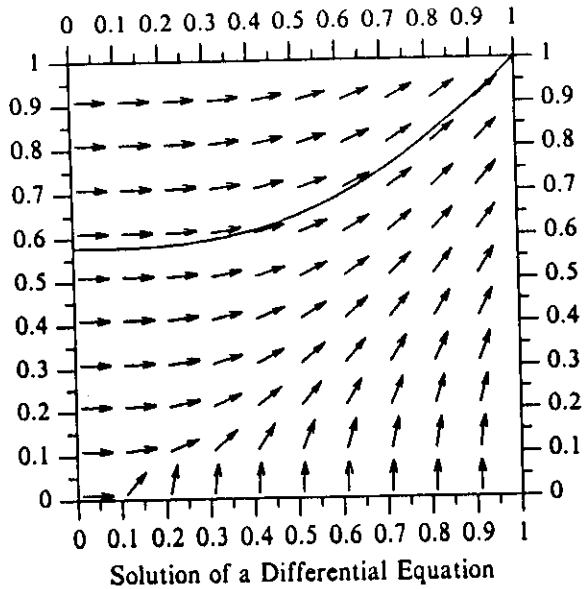
Continent	Area	%Earth	Pop.	%Total
Asia	16,999,000	29.7	2,897,000,000	59.8
Africa	11,688,000	20.4	551,000,000	11.4
North America	9,366,000	16.3	400,000,000	8.3
South America	6,881,000	12.0	271,000,000	5.6
Antarctica	5,100,000	8.9	0	0
Europe	4,017,000	7.0		
Australia	2,966,000	5.2		

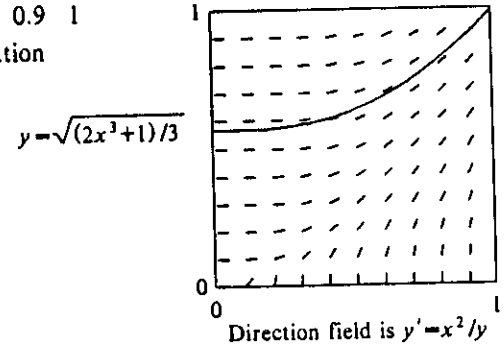
Continent	Area		Population	
	Mill. Sq. Mi.	%	Mill.	%
Asia	16.999	29.7	2,897	59.8
Africa	11.688	20.4	551	11.4
North America	9.366	16.3	400	8.3
South America	6.881	12.0	271	5.6
Antarctica	5.100	8.9	0	0
Europe	4.017	7.0	702	14.5
Australia	2.966	5.2	16	0.3

Continent	Area		Population	
	10^6 Sq. Mi.	% of Total	Millions	% of Total
Asia	16.999	29.7	2,897	59.8
Africa	11.688	20.4	551	11.4
North America	9.366	16.3	400	8.3
South America	6.881	12.0	271	5.6
Antarctica	5.100	8.9	0	0
Europe	4.017	7.0	702	14.5
Australia	2.966	5.2	16	0.3

CONTINENT	LAND AREA		POPULATION	
	Millions of Square Miles	Percent	Millions	Percent
Asia	16.999	29.7	2,897	59.8
Africa	11.688	20.4	551	11.4
North America	9.366	16.3	400	8.3
South America	6.881	12.0	271	5.6
Antarctica	5.100	8.9	0	0
Europe	4.017	7.0	702	14.5
Australia	2.966	5.2	16	0.3

Figure 7: Variations for a table





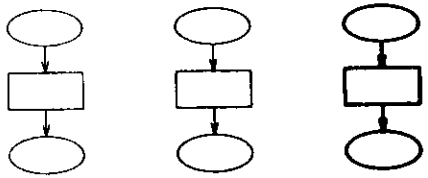
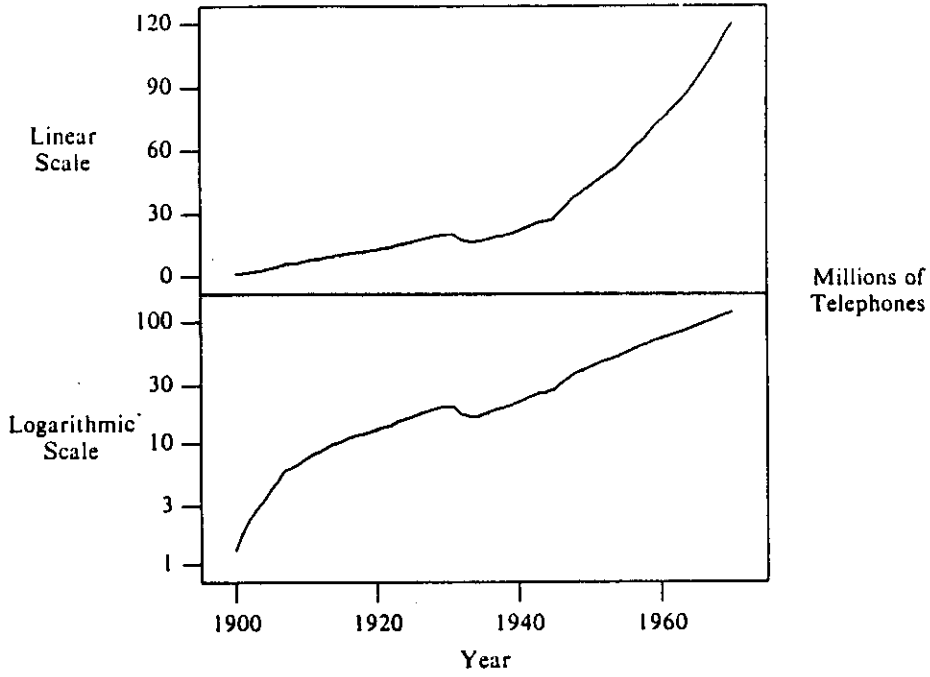
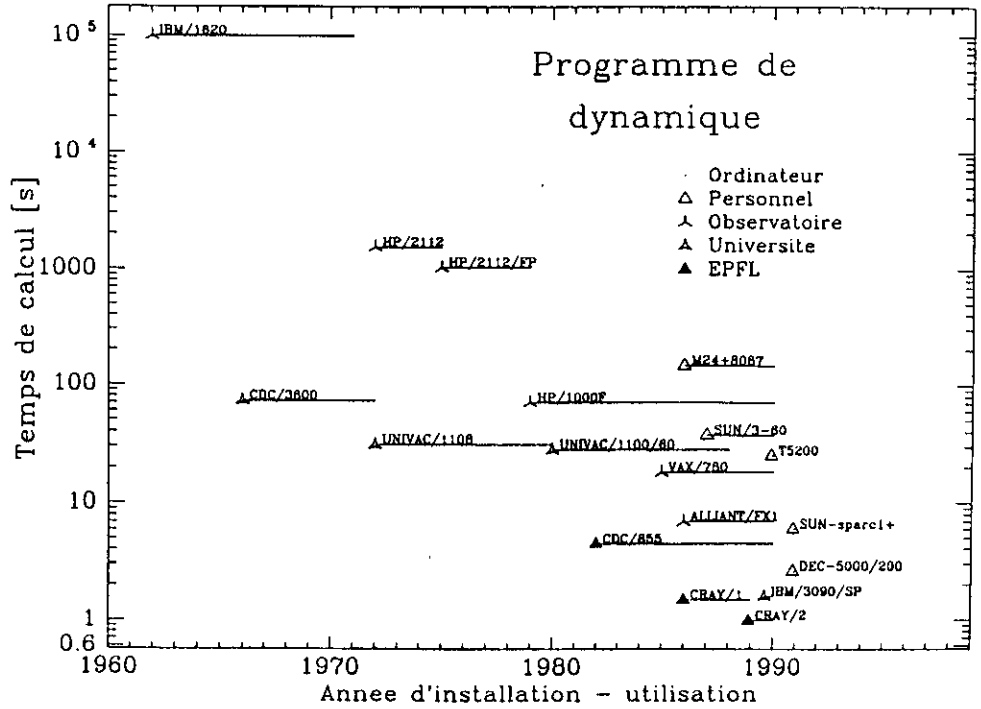


Figure 8: Various graphics





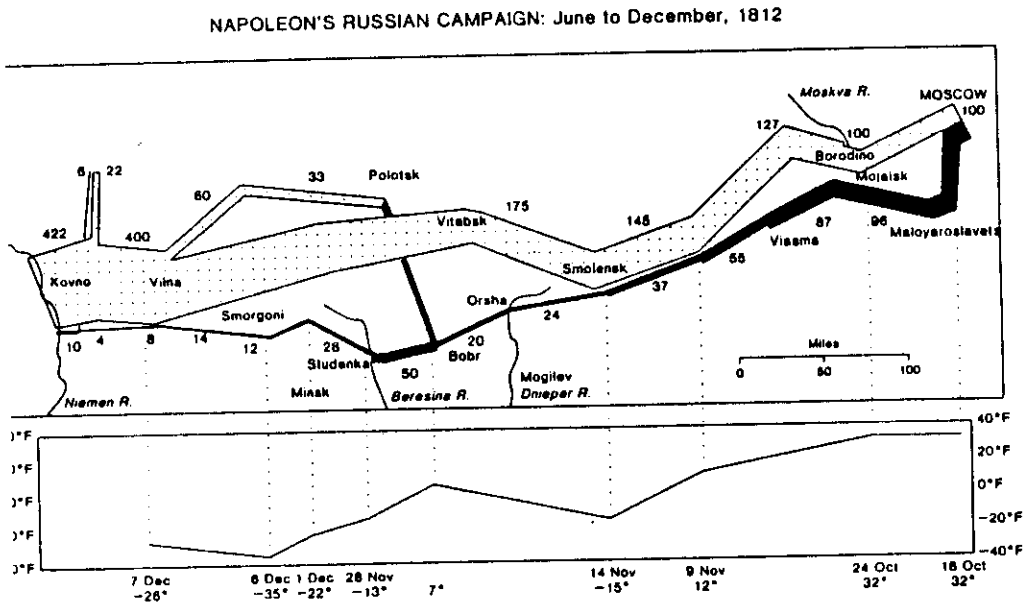


Figure 10: Napoleon goes to Moscow ... and comes back

7 Real Time Controls

7.1 Data Filtering

When data are noisy, it is difficult to see possible trends, changes or get the form of a signal.

What we want in fact is often to reduce a large set of noisy data into a much smaller one that is easy to interprete.

The statistical parameters seen earlier are one way to extract these basic informations. It assumes a non variable signal, whose mean is sometimes the only thing we are interested in.

A second way is to model the data according to a known relation. For example, if we record the radioactive emission of a short leaving element, then we know that the counts should follow a decaying exponential, and we could fit such a function to the data, using nonlinear least squares methods, and get good estimates of the initial amount and decay rate.

This method is very effective because it uses external information (exponential decay in our example) and reduce all the data to a strict minimum of parameters.

The information content in a set of data is always limited. If too many parameters are present, they will be less significant.

An unreferenced theorem says:

> From a given set of not infinitely precise data, one can extract an infinite number of meaningless parameters ...

The third method assumes no apriori functional knowledge. The data are filtered to reduce, possibly eliminate, the noise.

Effective technics assumes that we know the spectral distribution of the noise and signal, but this can be estimated from the data themselves, or a sample of them.

A very simple, but usually quite effective technic consists in replacing the original data with (weighted) average of the surrounding ones.

$$d_i \longrightarrow f_i = \sum_{j=i-k}^{j=i+k} w_j d_j$$

with $\sum w_j = 1$.

The w_j can be all equal, or approximate a gaussian using binomial values. The later choice gives a better smothing.

A variant use a recursive form that is easier and faster to implement:

$$d_i \longrightarrow f_i = \alpha d_i + \beta f_{i-1} \quad (1)$$

$$\text{or} = \alpha d_i + \beta f_{i-1} + \gamma f_{i-2} \quad (2)$$

with $\alpha + \beta = 1$ or $\alpha + \beta + \gamma = 1$.

In the first form, the effect of an erroneous value decays exponentially.

The smoothing effect is smal if $\alpha \sim 1$, large if $\alpha \ll 1$.

7.2 Alarms, risks

Ideally, one would like to ring an alarm when the signal overpass some limit, but only when it is significant, and never when it is just due to noise.

This is of course impossible, and we have to to accept some tradeof between the two risks. If the limit is to low, many falase alarm will ring (first risk), if it is set to hich, many real alarms will be lost (second risk).

A good estimate of the distribution of the signal, in particular its variance, will permit to set the limit at a value minizing the total (weighted) risk, using the usual statistical tests.