SOME APPLICATIONS OF STATISTICS IN ANALYTICAL CHEMISTRY

H.Nascu[a], L.Jäntschi[b], T.Hodisan[b]*, C.Cimpoiu[b], G. Cimpan[b]

[a] *Technical University, Department of Chemistry, 11 C.Daicoviciu, 3400 Cluj-Napoca, Romania*

[b] *"Babes-Bolyai" University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos, 3400 Cluj-Napoca, Romania*

e-mail addresses: ljantschi@hotmail.com; lori@lcibar.utcluj.ro; gcimpan@chem.ubbcluj.ro

CONTENTS

SUMMARY

The review presents some considerations on applications of statistics in analytical chemistry such as the statistics of the point, data analysis by regression, correlation and self correlation, dispersional analysis and ANOVA model, validation of statistic hypothesis.

Statistics have undergone an enormous impact from microelectronics, in the form of microcomputers and hand-held calculators. These have brought difficult statistical procedures within the reach of all practising scientists. The availability of the tremendous computing power naturally makes it all the

* corresponding author

more important that the scientist applies statistical methods rationally and correctly.

1. THE STATISTICS OF THE POINT

1.1 General considerations

Any natural phenomenon and especially those that can be characterized by numerical dates are the result of one or more causes of an action. The experiment, this powerful tool of scientific research, becomes efficient through the skill of the scientist to replace such a complex system of causes with a simple system where only one causative circumstance is allowed to change in time.

The chemist and the physicist have a certain advantage in their research. In their fields of science, experiment has reached a high level of perfection. However, even in these sciences, there are wide possibilities of applying research of a permanent statistical character. We could almost say that in all modern research statistical analysis is used. The elaborate methods aiming to eliminate the effect of circumstances that affect the conditions of measurement, although continually improving, have not reached and could not reach perfection. The scientist himself, and his entire apparatus of observation, constitutes a source of errors too; effects like changes in temperature, humidity, pressure and current, vibration, cannot be completely eliminated.

Statistics has to deal principally with numeric dates, generated by multiple causes. Through making an experiment the scientist wants to solve a certain complex of causes, singling out one by removing all the causes excepting one, or more exactly concentrating his attention on studying one of them by reducing the action of the others as much as possible. Statistics, lacking this possibility, is obliged to analyze data which are influenced by other data and, finally, to determine which are the most important causes and which are the results of the observation that can be attributed to the influence of each cause /1/.

We have to mention that any scientist has to take preventive measures. Thus, the quality of the data has to be examined before jumping to conclusions. This is valid for any kind of data and especially for numeric data

with previously verified quality. It is a waste of time trying to apply straight methods of calculation to process primary data that seem to be wrong.

1.2 Localization and scattering indicators

Having a measurement made at more than one time and the results of the measurement divided into two groups of study yields two sets of data. There are two typical fundamental values on which those two sets of data vary statistically:

(1) varying in level, meaning that central value around which the other values gravitate;

(2) varying in the amplitude of spreading observed values around the central value;

The indicators of first type, of level or position of localization are called **means**. Indicators of second type are called **measure scattering indicators**.

There are three forms of mean used more often: *arithmetic mean*, *median* and *mode*. Means like: *geometric* and *harmonic* are rarely used.

Let the row of measurements be $X_1, X_2, \ldots, X_N$

The **arithmetic mean** (or average) is the number M(X) given by:

$$M(X) = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1.1}$$

The **median** is the central value of the variable when its values are arranged in the order of their size or that value that has the property that smaller and bigger values appear in equal frequencies. We note X's median as m(X).

Let:{1,...,N}{1,...,N} be the permutation which arranges the measurements in increasing order:

$$X_{\pi(i)} \le X_{\pi(i+1)}, \quad i = \overline{1, N-1} \tag{1.2}$$

Then, If N is even,

$$m(X) = (X_{\pi(N/2)} + X_{\pi(1+N/2)})/2 \text{ else } m(X) = X_{\pi((N+1)/2)} \tag{1.3}$$

Observations: comparing between the average and the median, the easier operation is calculating the median; considering the fluctuations in selection, the average is more stable, but there are cases when the median is preferable.

Mode is the value of a variable that corresponds to the maximum of an ideal curve that gives the best fit possible to the real repartition. It represents the most frequent value, which is in fact "in vogue". When the repartition of values has a complicated form, there can be more then one mode. Those repartitions are called multimodes. The average and the median are unique for even this kind of repartitions. The mode is usually noted with a $\breve{}$ sign on top of the variable:

$$\breve{X} = \left\{ X_i \mid f_i \geq f_j, \quad j = \overline{1,N}, \quad f_j \text{ frequents of } X_j\text{'s apparition} \right\} \quad (1.4)$$

the frequency of the appearance of $\bar{X}$.

An empiric relation has been established between those three measures, with a sufficient approximation for moderately lop-sided repartitions. It is valid for a large number of cases:

$$\breve{X} \cong M(X) - 3 \cdot (M(X) - m(X)) \quad (1.5)$$

The **geometric mean** is the number given by

$$G(X) = \left(\prod_{i=1}^{N} X_i \right)^{\frac{1}{N}} = 10^{\frac{1}{N}\sum_{i=1}^{N} \lg(X_i)} \quad (1.6)$$

The **harmonic mean** is the number H(X) given by

$$\frac{1}{H(X)} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{X_i} \quad (1.7)$$

It can be seen that when we calculate the average of a set of values of a given measurement it is extremely important what kind of measurement we use for the considered values. An example as illustration is finding the arithmetic average, geometric average, and harmonic average of a large amount of records for a thermometer, expressed in Celsius, Kelvin and Fahrenheit /1/.

There are three indicators of spreading frequently used: the standard deviation, the average deviation, the quartilic deviation (the semiquartilic amplitude).

The standard deviation (the square average) is the number $\sigma(X)$ given by:

$$\sigma(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - M(X)\right)^2} \tag{1.8}$$

The value $\sigma^2(X)$ is called dispersion.

The average deviation is the number am(X) given by:

$$am(X) = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - M(X)\right| \tag{1.9}$$

In the case of symmetrical or moderately lop-sided repartitions, we have:

$$am(X) \cong \frac{4}{5}\sigma(X) \tag{1.10}$$

In the same way as the median, $Q_1(X)$ and $Q_3(X)$ are defined to be medians of the intervals $[X_{min}, m(X)]$ and $[m(X), X_{max}]$. $Q_1(X)$ is called inferior quartile and $Q_3(X)$ is called superior quartile. So $Q_1(X)$, m(X), $Q_3(X)$ divide the area of values observed in four equal groups of frequencies.

The measure Q(X) is given by:

$$Q(X) = \frac{Q_3(X) - Q_1(X)}{2} \tag{1.11}$$

and is called quartilic deviation. The quartilic deviation Q(X) has two advantages as compared to $\sigma^2(X)$ and am(X): it is easy to calculate and it has a clear and simple sense; however in other ways it has drawbacks: it has no simple algebraic proprieties and its behavior through fluctuations of selection is hard to forecast. That is why it is recommended only when other deviations are difficult or impossible to calculate.

The absolute measures of spreading (independent from the measurement units) are obtained dividing a deviation measure by an average. An example

of this kind of measure would be $\sigma(X)/M(X)$. Those proportions allow analogies between measurements of different natures.

There is also the coefficient of variation, defined as:

$$v(X) = 100\sigma(X)/M(X) \tag{1.12}$$

Corrado /2/ proposed two measures of spreading which present some advantages as compared to the standard deviation:

the **coefficient of the average difference $\Delta_1(X)$:**

$$\Delta_1(X) = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1,N}^{k \neq j} |X_k - X_j| \cdot f_k \cdot f_i \tag{1.13}$$

and the **coefficient of average difference with repetition $\Delta_1'(X)$:**

$$\Delta_1'(X) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} |X_k - X_j| \cdot f_k \cdot f_j \tag{1.14}$$

These coefficients are much more difficult to calculate than the deviations, but they have a theoretical attraction: they depend on the variables value difference and not on the spreading of the values around a random point, such as the arithmetic average or the median. They measure the intrinsic spreading of values, being independent of the origin of the calculation or of the level of repartition. In addition, if module function is placed instead of square function in the expression of $\Delta_1'(X)$ and it is noted with $\Delta_2'(X)$ we have:

$$\Delta'_2(X) = 2 \cdot \sigma^2(X) \tag{1.15}$$

For the symmetrical or moderately oblique repartitions we have the relations:

$$am(X) \cong 0,8; \quad Q(X) \cong 0,67 \tag{1.16}$$

For the majority of the repartitions 99% of the values are situated in an interval of $6 \times \sigma(X)$ or $7,5 \times am(X)$ or $9 \times Q(X)$.

The oblique of a repartition is its distance from symmetry. It is of interest in the case when the measures are referring to the same size, being an indicator of the quality of the measurement. There are two valid numeric characterizations for oblique:

$$Ob_{quartilic} = \frac{(Q_3 - m) - (m - Q_1)}{(Q_3 - m) + (m - Q_1)},$$

and respectively, (1.17)

$$Ob_{Pearson}(X) := \frac{M(X) - \overline{X}}{\sigma(X)}$$

The statistical parameters shown above are giving a qualitative measure of the experimental determination, and thus can serve /3,4/ as preliminary values in much more complex processing of the measured data or the experimental values.

1.3 The propagation of errors in calculation

For an analyst the propagation of errors is a very important matter because the errors made can affect the logical reasoning based on them. Let us consider a multiplicative expression like E(A,B)=const A·B where a and B are experimental sizes affected by errors; in this case the absolute error is obtained from the reasoning:

$$\varepsilon(E)=E(A+\delta A,B+\delta B)-E(A,B)=const(B\delta A+A\delta B+\delta A\delta B)\cong const(B\delta A+A\delta B)$$

and the relative error is:

$$\varepsilon_r(E)=\varepsilon(E)/E=\delta A/A+\delta B/B=\varepsilon_r(A)+\varepsilon_r(B) \quad (1.18)$$

so the **final relative error is the sum of the individual relative errors.**

Let us take the expression of a sum like: E(A,B)=A+B; the absolute error is:

$$\varepsilon(E)=E(A+\delta A,B+\delta B)-E(A,B)=(A+\delta A)+(B+\delta B)-A-B=\delta A+\delta B=\varepsilon(A)+\varepsilon(B) \quad (1.19)$$

and so the **final absolute error is the sum of the absolute individual errors**.

For example, if determinations are made of the molar concentration c_M and the volume of a solution V with a maximum precision of 4 decimals, then the number of moles $\nu=c_MV$ can be obtained with a precision $pr(\nu)=pr(c_M)+pr(V)=10^{-4}+10^{-4}=2\cdot10^{-4}$, i.e., not more then three numbers, because there exists the possibility that the error will destroy the value of the 4th exact number /5/. Sometimes in error minimization it is useful to apply neuronal networks /6/.

1.4 The criteria for eliminating the doubtful measurements

Consideration of the statement from the previous paragraph is required to make a better preliminary analysis of the measured values. Such an analysis, concluded by eliminating the questionable values (given by casual errors), can be considered a successful analysis /7/. We continue with the presentation of the elimination criteria of questionable data: /8/

"k" Criterion

The $\overline{X}$ and σ parameters are determined without the value of X_d (considered doubtful); the k parameter is determined from:

$$k = \frac{|X_d - \overline{X}|}{\sigma} \text{ where } \sigma = \sqrt{\frac{\Sigma(X_1 - \overline{X})^2}{N-2}} \qquad (1.20)$$

In the case of $k > k_P$ from the table then X_d must not be contained in the final calculation of the average value. When $k<k_P$, then we have no reason to consider it questionable for the given probability P% and n (observation number without the questionable value).

Table 1.1.

Value of parameter k as function of n determinations, n, and deviation probability

n	95%	99%	99,73%	n	95%	99%	99,73%
9	4.42	7.48	11.49	25	3.84	5.14	6.25
10	4.31	6.99	10.26	30	3.80	5.00	5.95
12	4.16	6.38	8.80	40	3.75	4.82	5.56
15	4.03	5.88	7.68	50	3.73	4.70	5.34
20	3.90	5.41	6.73	100	3.76	4.48	5.07

Chauvenet Criterion

According to this criterion, any value from an array of n determinations has to be eliminated if its deviation from the mean has a value such as to make the probability of appearance, for all deviations equal to or greater than this one, not higher than n/2.

We determine h (the precision modulus) and s (the deviation or the standard error of an observation):

$$h = \sqrt{\frac{n-1}{2\sum x_i^2}}\,; s = \frac{1}{h\sqrt{2}}\,; x_i = X_i - \overline{X}\,; \tag{1.21}$$

According to the relation (22) (hx_i) and (x_i/s) are determined. Those values of x_i can be admitted as non-questionable which verify at least one of the relations:

$$(x_i / s) \le (x/s)_{table}\ ;\ (x_i h) \le (xh)_{table} \tag{1.22}$$

where $(x/s)_{table}$ and $(xh)_{table}$ are obtained from Table 1.2.

It should be mentioned that if the value of X_i, which has the biggest deviation from the average, is not eliminated, then none of the other values of the variable X must be eliminated.

"R" Criterion

In the same way, for the application of R criterion we calculated:

$$s = \sqrt{\frac{\sum x_i^2}{n-2}}\ ; \tag{1.23}$$

For each observation we calculated:

$$x_i = X_i - \overline{X} \text{ and then } R_i = \frac{|x_i|}{s} \tag{1.24}$$

In the case $R_i \le R_{table}$ (Table 1.3), then the observation i is admitted in the final calculation of the medium value, not being a questionable value.

"r" Criterion

This criterion is based on the statistical hypothesis of the r value normal distribution, given by:

$$r = \frac{|x_i|}{s \cdot \sqrt{\frac{n-1}{n}}}; \text{ where } s = \sqrt{\frac{\sum x_i^2}{n-1}} \text{ and } x_i = X_i - \overline{X},$$

a normality with (n-2) degrees of freedom. The questionable value x_i, with an r-value lower than r_{table}, is eliminated. Considering the same principle, the r admissible values, for a probability p=95%, are given.

Table 1.2

Value of (xh) and (x/s) for divert value of n

n	(xh)	(x/s)	n	(xh)	(x/s)	n	(xh)	(x/s)	n	(xh)	(x/s)
5	1,16	1,68	9	1,35	1,92	18	1,56	2,20	26	1,66	2,35
6	1,22	1,73	10	1,39	1,96	20	1,58	2,24	30	1,69	2,39
7	1,27	1,79	12	1,44	2,03	22	1,61	2,28	40	1,77	2,50
8	1,32	1,86	14	1,49	2,10	24	1,63	2,31	50	1,82	2,58

Table 1.3

Value of parameter R for n determinations and divert probability

n	95%	99%	n	95%	99%	n	95%	99%	n	95%	99%
4	7,71	16,27	7	3,98	5,88	10	3,54	4,75	14	3,36	4,28
5	5,08	9,00	8	3,77	5,33	11	3,48	4,58	16	3,32	4,17
6	4,34	6,85	9	3,63	4,98	12	3,42	4,45	18	3,30	4,08

Table 1.4

Value of parameter r for n determinations and P=95%

n	95%	n	95%	n	95%
3	1,397	7	1,640	11	1,649
4	1,559	8	1,644	21	1,649
5	1,611	9	1,647	22	1,648
6	1,731	10	1,648	40	1,648

"Q" Criterion

This criterion, even simpler, is statistically correct, and gives better results than the previous did. The row of measurements is ordered, i.e., the permutation $\pi:\{1,...,N\}\rightarrow\{1,...,N\}$ is found so that the order is:

$$X_{\pi(i)} \leq X_{\pi(i+1)}, \quad i = \overline{1,N-1}$$

We determined the value:

$$Q_1 = \frac{X_{\pi(2)} - X_{\pi(1)}}{X_{\pi(N)} - X_{\pi(1)}}$$; if $Q_1 > Q_{P,table}$, $X_{\pi(1)}$ is eliminated from the set adjusting the indices in the row, and the procedure is repeated until $Q_1 < Q_{P,table}$. Then we determined the value:

$$Q_N = \frac{X_{\pi(N)} - X_{\pi(N-1)}}{X_{\pi(N)} - X_{\pi(1)}}$$; $Q_N > Q_{P,table}$, $X_{\pi(N)}$ is eliminated from the row, following the above procedure until $Q_N \leq Q_{P,table}$

$Q_{P,table}$ is an admitted value of Q_K with the p probability, corresponding to the value $X_{\pi^{-1}(K)}$ as an admissible value.

Table 1.5

Value of parameter Q for divert n value and divert probability

n	p=95%	p=99%	p=99,5%	n	p=95%	p=99%	p=99,5%
3	0,941	0,988	0,994	11	0,392	0,502	0,542
4	0,765	0,889	0,926	12	0,376	0,482	0,522
5	0,642	0,780	0,821	13	0,361	0,465	0,503
6	0,560	0,698	0,740	14	0,349	0,450	0,488
7	0,507	0,637	0,680	15	0,338	0,438	0,475
8	0,468	0,590	0,634	16	0,329	0,426	0,463
9	0,437	0,555	0,598	18	0,313	0,407	0,442
10	0,412	0,527	0,568	20	0,300	0,391	0,425

Table 1.6

Value of parameter t for divert n value and divert probability

n	95%	99%	99,5%	n	95%	99%	99,5%	n	95%	99%	99,5%
3	6,31	31,82	63,66	10	1,86	2,90	3,36	17	1,75	2,60	2,95
4	2,92	6,97	9,93	II	1,83	2,82	3,25	18	1,75	2,58	2,92
5	2,35	4,54	5,84	12	1,81	2,76	3,17	19	1,74	2,57	2,90
6	2,13	3,75	4,60	13	1,80	2,72	3,11	20	1,73	2,55	2,88
7	2,02	3,37	4,03	14	1,78	2,68	3,06	30	1,70	2,47	2,76
8	1,94	3,14	3,71	15	1,77	2,65	3,01	40	1,68	2,42	2,70
9	1,90	3,00	3,50	16	1,76	2,62	2,98	60	1,67	2,39	2,66

"t" Criterion

This criterion is also called the Student criterion because it uses the t variable from the student distribution. We determine the mean of (n-1) from n values:

$$M_i(X) = \frac{n \cdot M(X) - X_i}{n-1} \tag{1.25}$$

that is, more exactly, the average of all the others, less X_i. We also determined the standard error of a measuring as the one from the example:

$$s_i = \sqrt{\frac{\sum \left(M_i(X) - X_i\right)^2}{n-2}} \tag{1.26}$$

Finally, we can determine the t_i measure, corresponding to each i measurement:

$$t_i = \frac{\left|M_i(X) - X_i\right|}{s_i \sqrt{\frac{n}{n-1}}} \tag{1.27}$$

If $t_i < t_{table}$ then the i measurement is accepted as being non-questionable.

In conclusion, all these tests are efficient, and have been proven to give good results, but the best one is that which eliminates a questionable value with the highest probability: *t criterion*. Their use depends on the analyst's

choicee in function of the problem complexity which is considered. Some papers have approached the same problem by diverse ways /9-12/

Example 1

A standard sample of pooled human blood serum contains 42.0g of albumin per litter /13/. Five laboratories (A-E) each do six determinations (on the same day) of the albumin concentration, with the following results (g/l throughout):

Labs	det1	det2	det3	det4	det5	det6
A	42.5	41.6	42.1	41.9	41.1	42.2
B	39.8	43.6	42.1	40.1	43.9	41.9
C	43.5	42.8	43.8	43.1	42.7	43.3
D	35.0	43.0	37.1	40.5	36.8	42.2
E	42.2	41.6	42.0	41.8	42.6	39.0

The following averages are obtained for each laboratory

L	M(L)	m(L)	G(L)	H(L)	σ(L)	σ^2(L)	am(L)
A	41.90	42.00	41.897	41.895	0.4939	0.2440	0.3667
B	41.90	42.00	41.871	41.841	1.7076	2.9160	1.3000
C	43.20	43.20	43.198	43.196	0.4195	0.1760	0.3333
D	39.10	38.80	38.987	38.874	3.2520	10.576	2.8000
E	41.53	41.90	41.516	41.498	1.2879	1.6586	0.8444

The mode is calculated with the determinations from all laboratories, because if we consider the determinations of each laboratory the number of determinations will be insufficient to decide the mode. The calculated frequencies are shown in the next table:

freq.	det1	det2	det3	det4	det5	det6
A	1	2	2	2	1	3
B	1	1	2	1	1	2
C	1	1	1	1	1	1
D	1	1	1	1	1	3
E	3	2	1	1	1	1

The mode is reached in the first determination of laboratory E and also the sixth determination of laboratories **A** and D: 42.20. So, $\tilde{L} = 42.20$. For the closer data sets the values of the parameters defined by the relations (1.11)-(1.18) are obtained:

L	m(L)	Q1(L)	Q3(L)	Q(L)	$\Delta_1(L)$	$\Delta_1'(L)$	$Ob_{quartil}(L)$	$Ob_{Pearson}(L)$
A	42.0	41.6	42.2	0.30	0.600	0.500	-0.333	-0.607
B	42.0	40.1	43.6	1.75	2.080	1.733	-0.085	-0.175
C	43.2	42.8	43.5	0.35	0.520	0.433	-0.070	+2.383
D	38.8	36.8	42.2	2.70	3.973	3.311	+0.259	-0.953
E	41.9	41.6	42.2	0.30	1.333	1.111	+0.000	-0.520

Example 2

Let us consider the measurements taken with a voltmeter that indicates the potential with a precision of four decimals /10/:

E(V)= 1.1125; 1.1124; 1.1124; 1.1125; 1.1125; 1.1124

We observe that if the modes 1.1124 and 1.1125 are values from the series, and so real measurements of the phenomenon, which is an advantage, there is still the disadvantage that a single mode (which could be considered the most profitable value)is not available.

The arithmetical mean and the median, even though they have the most appropriate value (most frequent), an incontestable advantage, can never be obtained as a real measured value, which confers a disadvantage to the arithmetical mean.

2. DATA ANALYSIS BY REGRESSION

2.1 Regression as an investigation tool of series trends

The regression analysis can be applied if the characteristics Y, X_1, X_2, ..., X_p show a close relationship, almost functional, when they are simultaneously studied for a specific type of samples, chemical species or analyzed materials.

Regression analysis in this case is synonymous with empirical modeling, curve fitting or forecasting.

The regression analysis allows the creation of a regression equation, that is a function that permits the calculation of concentration for one of the species, on the basis of the other data, and with measurable errors.

The confidence in the established equation increases with the increasing of the number of points (in the multidimensional space).

The relationship between the factors affecting the analytical signal is a statistical one, due to the random errors that always can appear. The Y values can be obtained by using interpolation procedures of the distribution $Y(X_1,X_2,...,X_p)$, performing a proximity between the functional relationship (ideal) and the statistic one (real). This kind of analysis gives a useful mathematical model, which apparently has no physical model as support.

However, the best results are obtained when a concordance is established between the considered physical model and the mathematical one. For example, the Lambert-Beer law assures the validity of linear equations for spectrometric absorption methods.

2.2 Regression models

According to the mathematical model this model can be divided into linear and non-linear /11,12/. Taking into account the number of independent variables, there are monovariable models ($Y=Y(X)$) and multivariable models /13/ ($Y=Y(X_1,X_2,...,X_n)$), which may be regarded as similar if we suppose that $X=(X_1,X_2,...,X_n)$.

The unidimensional linear regression is frequently used in analytical practice and it considers the following model for the phenomenon:

$$y = \hat{y} + \varepsilon ; \; \hat{y} = b_0 + b_1 \cdot x \qquad (2.1)$$

where:

x, y are the characteristics measured by the analyst;

$\hat{y}$ is the characteristic estimated for y using the model;

b_0 and b_1 are coefficients estimated with the model;

and ε is the error of estimation.

Even in the case of linear regression, there is a broad meaning of the linear dependency concept, which can be extended to linear dependency. According to this concept, a regression equation is linear if the functional

dependency between the considered variables can be linearised /14/. Therefore, the following regression equations:

$$\begin{aligned} &y=a\cdot\log(x)+b; \\ &y=a\cdot\log(\log(x))+b; \\ &y=a\,(1/x)+b; \\ &y=a\cdot e^{x}+b \end{aligned} \tag{2.2}$$

are linear dependencies and can be associated with the linear model:

$$y=a\cdot z+b \tag{2.3}$$

where the new independent variable z is obtained by substitutions:

$$z=\log(x);\ z=\log(\log(x));\ z=1/x \text{ or } z=e^{x} \tag{2.4}$$

Another extension of the linear regression model can be obtained when the error factor influences both variables involved in the regression. In this case, the formulae for the validation of regression parameters have another form /15/.

A problem that appears in the case of regression, in general, is the parameters' estimation. This problem is solved differently by many authors.

A well known estimation model for parameters is based /16/ on the **minimization of risk,** defined as the mean of square loops function proposed by Kolmogorov, best known under the name of the **least squares method**. The expression which must be minimized is, in this case, given by:

$$K(X,Y,B)=\sum\left(\hat{y}-y\right)^{2}=\sum\left(b_{0}+b_{1}\cdot x-y\right)^{2} \tag{2.5}$$

where X,Y,B are the column vectors of the independent variable X, of the dependent variable Y and of coefficients B.

Other papers /17,18,19,20/ have described different approaches of the estimation model based on the least square method. It has to be mentioned here that other estimation models based on the loops function and the sum of residues has been developed.

Briefly, these functions are:

R. Fisher, 1912, **maximum verisimilitude method**:

$$F(X,Y,B) = \sum\left(1 - e^{-(\hat{y}-y)^2/2}\right) = \sum\left(1 - e^{-(b_0+b_1\cdot x-y)^2/2}\right) \quad (2.6)$$

J. Newman, A. Waad, **minimax method**:

$$NW(X,Y,B) = \sum|\hat{y} - y| \quad (2.7)$$

Bayes, 1750, **maximum aposteriori probability**:

$$BY(X,Y,B) = \sum\begin{cases}0, & \hat{y} - y < D(\hat{Y}-Y)/2\\ 1, & \hat{y} - y \geq D(\hat{Y}-Y)/2\end{cases} \quad (2.8)$$

By substitution in the multidimensional case of linear regression it follows:

$x^T=(x^0,x^1,...,x^p)$, $x^0=1$; $X=(x_1,x_2,...,x_N)$; $Y=(y_1,y_2,...,y_N)$; $\hat{Y} = (\hat{y}_1,\hat{y}_2,...,\hat{y}_N)$; $B^T=(b^0,b^1,...,b^p)$ and $\hat{y}$ is given by:

$$\hat{y} = \sum_{i=0}^{p} b^i \cdot x^i \quad (2.9)$$

Minimizing square of errors (using (2.5)), the loop equation is K(X,Y,B)=min, resulting in:

$$K(X,Y,B) = \sum(\hat{y} - y)^2 = \sum_{j=1}^{N}\left(\sum_{i=0}^{n} b^i x_j^i - y_j\right)^2 = \min \quad (2.10)$$

The solution is found using linear algebra by the equation system:

$$\frac{\partial}{\partial b^k}\sum_{j=1}^{N}\left(\sum_{i=0}^{p} b^i x_j^i - y_j\right)^2 = 0, \quad k = \overline{0,p} \quad (2.11)$$

which, after rearrangement of the sum is:

$$\sum_{i=0}^{p} b^i\left(\sum_{j=1}^{N} x_j^k x_j^i\right) = \sum_{j=1}^{N} x_j^k y_j \quad k = \overline{0,p} \quad (2.12)$$

and the solution is given by:

$$B=CZ^{-1} \tag{2.13}$$

where:

$$Z = (z_k^i)_{\substack{0\le i\le p \\ 0<k<p}} = (\sum_{j=1}^{N} x_j^k x_j^i)_{\substack{0\le i\le p \\ 0<k<p}} \text{ and}$$

$$C^T = (c^k)_{0\le k\le p} = (\sum_{j=1}^{N} x_j^k y_j)_{0\le k\le p} \tag{2.14}$$

In spectral analysis /21/, for p samples, each with r compounds and the constituents determined having q channels (for example different wavelengths in the same spectra) which linearly depend on concentrations, the following equation can be written:

$$R=CS^T+E \tag{2.15}$$

where:

R is the matrix for signals obtained on q channels for each of the p samples (dimension p×q);

C is the concentration matrix for r compounds (dimension p×r);

S is the sensitivity matrix (dimension q×r);

E is the error matrix (dimension p×q).

Due to the fact that computerized data acquisition is now usual in chemical laboratories, such methods based on linear algebra and multilinear statistics are being routinely applied to multicomponent quantitative analysis /25-32/. The method was applied for the establishment of functional dependencies between the retention factors in chromatography (log k) and the molecular parameters of the separate components /18,33/ or for the description of the influence of different factors on the chromatographic retention /34,35/ or the partition coefficient /36/, (log K). The method was also related to the association criteria of some solvents, regarding the alkaloids separation and the multidimensional relationship between their structure and properties /37,38/. The method will probably be successful, due to the diversification and the improvement of the detector instrumentation,

which means an improvement of resolution and sensitivity and a decrease of noise. The automated data processing has a considerable contribution to make, too.

After establishing the coefficients and the errors that affect the data, using the regression equation, the reverse way is followed and the regression equations are transferred to calibration equations (the multidimensional correspondent of the bidimensional calibration curve). Between these two aspects, there are elements that should be clarified in practice /39/. This method of data analysis was described in many recent books /13,40/ and papers /41-44/. The multivariate calibration with non-linear equations is also used /45/.

As an auxiliary technical method, *Artificial Neural Networks*, by W.J. Welsh *et al.* /46/ is used as a preliminary investigation of input experimental data. Authors use a computer program: *Brainmaker professional* (California Scientific Software) to fit the neural networks model /47/. The terms of "expert systems" are also used for this area of application /43/.

The equations and regression models have a widespread application together with the development of analytical instrumentation. The calibration curves are habitual in this field and many analysts use the calibration through the regression curves /18,25,31-33,37,48,49/.

Otherwise, the principal component analysis has recently been applied in the multilinear regression analysis /50/. This method is close to factorial analysis (see *Factorial Analysis and ANOVA*), but is related to multilinear regression.

The method is in fact a repeated linear regression for a number of times equal to the number of considered main components. The coefficients of the considered component are determined for each iteration, having as input data:

X_K: the principal characteristic K;

Y_K: the residue obtained from the iteration for the principal component K-1

and as output data:

Y_{K+1}: the residue obtained from the regression Y_K with X_K;

B_K: the vector of coefficients for principal component K.

The **principal component analysis** (PCA) is preferred to multilinear regression (PLS-partial least squares) for theoretical and practical reasons. One of the theoretical reasons is that the B_K vectors (K=1,2,...) are orthogonal in the multidimensional space of principal components. From the practical points of view:

- the number of principal components it is not fixed from the beginning and can be modified without affecting the already calculated principal components;
- it is easier to interpret each component by its projection in the corresponding plane;
- the correlation between rows of data are not influenced by applying repeated linear correlation instead of a multiple linear regression /39, 51-55/.

In case of optimization, when the number of data sets exceeds the number of coefficients /3,4/, the optimization model leads to a regression equation system. In this case the sum of errors generated by each equation is minimized for obtaining a determinate equations system, the coefficients of which are deduced on the basis of the same algebraic principle enunciated in multilinear regression. The obtained regression equation is used on to give the quantitative interpretations of the studied phenomenon through optimized parameters.

Software development has produced an explosion on the market of specialized programs for statistics. The majority of these programs have implemented routines for calculus of different regression types. Some of these programs are presented in the next paragraph "Useful Programs".

2.3 Useful Programs

GraFit; Data Analysis and Graphics Program; Erithacus Software Ltd.
Slide; Slide Write Plus for Windows; Advanced Graphics Software Inc.;
MathCad; MathSoft Inc.; Collabra Software Inc.;
Excell; Microsoft Corporation; Soft Art Dictionary and Program;
Statistica; Statistica for Windows; StatSoft Inc.;
Surfer for Windows; Software Package; Golden Software;
Statistix; Analytical Software; Tallahassee;
Brainmaker Professional; California Scientific Software;

3. CORRELATION AND SELF CORRELATION

3.1 Correlation. Indicators of linear correlation

Correlation supposes the existence of at least two numerical series, ordered by a certain common criterion, usually temporal. For example the

series obtained by observing the concentration and the potential measured with an electrode concurrently with the titrant volume added in a titrant reaction constitutes two numerical series ordered temporally.

We note the numerical series X(t) and Y(t) where $t= t_1, t_2, \ldots, t_n$ are the suitable times of observation. The correlation study can be made either by graphic or analytic means. If the graphic route is more easy for an analyst, the analytic one provides more advantages, both through the development of calculation technique and through the automatic analysis (processing) of data /1/.

Graphically, we have two possibilities of investigating the correlation between the two series of time: the first, by representing on the same graph the dependencies X=X(t) and Y=Y(t) the connection is observed that can be established by the representation in a XOY system of Y=Y(x) dependence or X=X(y) dependence. In the second case the figure obtained is called field of correlation /56/. The functional dependence Y=Y(x) or X=X(y) can be inferred from the study of this field of correlation.

Analytically, the following indicators of the correlation are defined:

$$\nu(X,Y) = M(X \cdot Y) = \frac{1}{n}\sum_{i=1}^{n} X_i \cdot Y_i \tag{3.1}$$

$$\mu(X,Y) = \nu(X,Y) - M(X) \cdot M(Y) \tag{3.2}$$

where ν is the second-degree moment and so it is the mean of $X_i \cdot Y_i$ parameters and μ is the second degree moment or covariance or correlation of the two data through the numeric series considered previously.

We can also derive the *correlation coefficient* given by:

$$r(X,Y) = \frac{\mu(X,Y)}{\sqrt{D^2(X)D^2(Y)}} = \frac{\mu(X,Y)}{\sigma(X)\sigma(Y)} \tag{3.3}$$

Among all indicators, the correlation coefficient is the one that is most often used for the analytic characterization of the correlation between two parameters.

The higher the $\mu(X,Y)$ correlation, the stronger the functional dependence between X and Y, and r also becomes higher. When r = 1 the correlation reaches the maximum, and X and Y change in directly proportionality. The

smaller the $\mu(X,Y)$ parameter, the stronger the functional dependence between X and Y is stronger, but inversely, when X increases Y decreases. When r = -1, the correlation is also at its minimum value, X and Y vary in inverse proportionality.

We must mention that the correlation expression as it is given by the connection (1)-(3) between X and Y is a quantitative dependence. Because of this many authors call the (1)-(3) correlation **linear correlation**.

For this reason, the smaller the $\mu(X,Y)$, the weaker the linear dependence, and the correlation coefficient is smaller in absolute value. When r=0, we can say that there is no linear correlation between X and Y.

Although a way was proposed to investigate the non-linear correlation Y=Y(X), it is not as accessible as linear correlation; therefore when correlation is non-linear, regression analysis is preferred instead of correlation analysis.

Some authors are considering r^2 as a statistical index of correlation between series, and present results in this form /57-60/.

These authors use the r^2 parameter because by squaring a fractional number, the first digits of this number become more significant. The r^2 parameter allows one to observe, with the same precision of four digits, the changes of the digits of r up to 8^{th} position (see also the error propagation) /61/.

3.2 The ranks correlation. Parameters

The ranks correlation is used especially when the series of inputs do not have rigorous values, being affected by a large amount of experimental errors or by the resulting calculations. An important cause is also when the errors that are acting on inputs have a systematic character. Usually in this case a great deal of importance can be attached to the resulting values, even though the absolute individual values from the numerical series are no longer expressing the phenomena observed; still the order connections that are established between these values are under the influence of the error factor. So, in this case, the only useful parameter is the position of a measurement in their ordered row, the parameter which is in fact the one that is used.

We introduce here the notion of **rank**: the rank is the measurement position in the ordered row of measurements. Let the row be $X_1, X_2, \ldots, X_N$ and let it be permutated:

$$\pi:\{1,...,N\}\rightarrow\{1,...,N\}:\ X_{\pi(i)} \le X_{\pi(i+1)},\quad i=\overline{1,N-1} \tag{3.4}$$

so as to put measurements in increasing order (see also the first paragraph). The rank of X_i is therefore $X_{\pi(i)}$.

There are the series $X_1, X_2, \ldots, X_N$ and $Y_1, Y_2, \ldots, Y_N$ and π_1 and π_2 the permutations that put the X and Y values in the order:

$$X_{\pi_1(i)} \le X_{\pi_1(i+1)},\quad i=\overline{1,N-1};\ Y_{\pi_2(i)} \le Y_{\pi_2(i+1)},\quad i=\overline{1,N-1} \tag{3.5}$$

$$\text{There is } d_k = \pi_1(k)-\pi_2(k) \text{ where } k=\overline{1,N} \text{ and also } d=\Sigma|d_k|. \tag{3.6}$$

If d=0 then the considered series are on the same order and there is a perfect correspondence of ranks.

The correlation coefficient of ranks, the so called C.E. Spearman coefficient, is obtained by effecting the simple calculations of $r(\pi_1,\pi_2)$, and taking into account that

$$M(\pi_1)=M(\pi_2)=\frac{n+1}{2}:\theta(X,Y)=r(\pi_1,\pi_2)=1-\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.7}$$

Notes:

(i) when $\Sigma\delta_i^2=0$, then $\Sigma|d_i|=0$ and $\theta=1$ (the maximum rank correlation)

(ii) when $\Sigma\delta_i^2$=maxim then $\Sigma|d_i|$=maxim and $\theta = -1$ (the maximum reversed rank correlation)

(iii) $-1 \le \theta \le 1$

In order to define Kendal's coefficient we introduce the functions K_1 and K_2 given by the relations:

$$K_1(i)=\left|\left\{k \mid \pi_2(k)<\pi_2(i),\ \pi_1(k)<\pi_1(i),\ k<i\right\}\right| \tag{3.8}$$

meaning that the rank number Y is smaller than the rank i of Y in the rank series of X to rank i;

$$K_2(i)=\left|\left\{k \mid \pi_2(k)>\pi_2(i),\ \pi_1(k)<\pi_1(i),\ k<i\right\}\right| \tag{3.9}$$

meaning that the rank number of Y is bigger than the rank if of Y in the rank series of X to rank i;

The measures $P_i=1-\pi_2(i)+K_1(i)$, $Q_i=N-\pi_2(i)-K_2(i)$ are calculated and these values $S_i=P_i+Q_i$ (3.10)

The relation gives Kendal's coefficient:

$$\kappa = \frac{2 \cdot \sum_{i=1}^{N} S_i}{n(n-1)} \tag{3.11}$$

Observations:

(i) k = 1 when both series are in the same order $\pi_1=\pi_2$

(ii) k = -1 when both series are in the contrary order $\pi_1 \circ \pi_2 = 1_N$

The rank correlation is successfully used at Genetic Programming /22/.

3.3 Self-correlation. The ranks self-correlation

Stage of correlation. The "stage of correlation" is the time from which the moments of time should be counted so that the correlation should be maximum as against the other possibilities of counting. In the case where the correlation is maximum and attains unity, it is said that the series are in synchronism. In this case, in calculating the correlation, only the N-φ data sets are taking part, where φ is the stage presumed as being the correlation stage.

Self-correlation. The term "Self-correlation" is used when every individual numerical series is studied. This is, therefore, different from correlation because the series x and Y represent the same measurement, possibly accomplished in another temporal order or with time distance. More often, the self-correlation is used with time distance. **Self-correlation of series at first rank** is the name given to correlation between the initial series and the initial series moved by one term to the right, so that the correlation coefficient suitable is:

$$r_1(X) = \frac{\sum_{k=2}^{N} X_k \cdot X_{k-1}}{\frac{N-1}{N} \cdot \sum_{k=1}^{N} X_k^2} \tag{3.12}$$

and for self-correlation at rank j, the correlation coefficient is:

$$r_j(X) = \frac{\sum_{k=j+1}^{N} X_k \cdot X_{k-j}}{\frac{N-j}{N} \cdot \sum_{k=1}^{N} X_k^2} \tag{3.13}$$

It can be noticed that the more rank j increases, the number of terms involved in correlation wanes and, because of this, when the numerical series are studied by the help of self-correlation it is preferred that the number of terms be as big as possible.

Periodicity. For phenomena that are repeated in time, it is important to find out the time span after which the system crosses again through the vicinity of initial state, repeating almost the same system. The time sequence after which the system changes this state is called **period**. The period is larger here: the first j rank to which on j. k ranks the function of self-correlation $r_{kj}(X)$ attains local maximum values is called **period of self-correlation**:

$$P=\min\{j|\ r_{k\,j}(X)>r_i(X)\ \forall k,\ i\neq k\ j\} \tag{3.14}$$

Rank self-correlation. The series X is replaced with the series π of ranks obtained by ordering X series. Coefficients of self-correlation which are obtained are:

$$\rho_1(X) = \frac{\sum_{k=2}^{N} \pi(k) \cdot \pi(k-1)}{\frac{(N^2-1)(2 \cdot N+1)}{6}};$$

$$\rho_j(X) = \frac{\sum_{k=2}^{N} \pi(k) \cdot \pi(k-j)}{\frac{(N-j)(N+1)(2 \cdot N+1)}{6}} \tag{3.15}$$

Example 3

Multiplex dsDNA fragment size distribution uses intercalation dyes and capillary array. Studying electrophoresis by ionic effects on the stability and electrophoretic mobility of DNA-dye complexes, Clark and Mathies /63/ obtained the effect of DNA/dye ratio on electrophoretic mobility. The sample consisted of pBR322 MspI complexes with TOTO at the indicated bp/dye ratios and the standard consisted of ΦX174 HaeIII with 1 buTOTIN/25bp. Size was estimated against a second order polynomial fit to the first eight Φ X174 HaeIII peaks. The TO control presents the results when both samples were strained on-column by adding 0.1 μM TO to the 80mM taps-NPe_4, 1mM H_2EDTA, 0.8% HEC, pH 8.4, electrophoresis buffer. All DNA concentrations were 25pg/μL in 1/100×TAE.

The table is:

Table 3.1

no (t)	actual size/bp (X)	TO control (Y1)	100bp/ TOTO (Y2)	50bp/ TOTO (Y3)	25bp/ TOTO (Y4)	10bp/ TOTO (Y5)	5bp/ TOTO (Y6)
1	67	69	67	65	65	66	67
2	76	77	74	75	73	73	74
3	90	86	84	85	85	84	84
4	110	104	103	104	103	103	103
5	123	115	114	115	114	114	114
6	147	139	138	138	138	139	139
7	160	149	149	148	149	150	150
8	180	171	171	170	170	170	170
9	190	185	185	184	184	184	184
10	201	192	193	191	191	191	191
11	217	211	211	209	209	208	209
12	238	233	233	231	231	230	231
13	242	240	240	238	237	237	237
14	307	308	308	306	306	305	306
15	404	414	414	412	410	409	410
16	527	527	528	526	525	521	524
17	622	618	616	615	615	603	591

Parameters given by (3.1)-(3.11) are:

Table 3.2

para-meter	X, Y1	X, Y2	X, Y3	X,Y4	X, Y5	X, Y6
ν	75050.17	74965.58	74662.11	74556.11	73941.17	73672.11
μ	23243.81	23294.21	23206.71	23195.20	22823.22	22594.66
r	0.99950	0.99954	0.99958	0.99967	0.99950	0.99895
θ	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
κ	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

and parameters given by (3.12), (3.13), (3.15) are (see Table 3.3):

The series are not periodic because rank $j>1$ local maxim does not exist.

4. THE DISPERSIONAL ANALYSIS AND ANOVA MODEL

Dispersional analysis is related to a group of statistical methods for studying experiments which produce data depending on different factors, about the suspected influence of which we are not sure whether or not it must be taken into account. The aim of these methods is to establish the main factors that will be considered. The analysis does not always produce a function (curve or surface), but only the coordinates for determination of the proper function on the basis of another experiment. Sometimes, the analysis result shows only the factors that must be strictly controlled (kept constant),so that the data obtained in two different laboratories should be similar.

R. A. Fischer introduced the dispersion concept, and used it in the determination of qualitative and quantitative factors. Box and Wilson /64/ have completed the approach, applying the dispersion for the determination of different factor effects.

The method can be found under different names: "dispersional analysis" /65/, "experiments projection" /66/, "response surface analysis" /67/ or "factorial analysis" /68/. The models which consider only part of the variables as random, while another part are kept constant, are well known as ANOVA /69/. However, this definition is not always respected.

The main idea of these methods is that the dispersion of an experimental result or of a variable (i.e. partition coefficient) is increasing when the

Table 3.3

coefficient	X	Y1	Y2	Y3	Y4	Y5	Y6
r_1	0.884015	0.883578	0.884446	0.883565	0.883003	0.887215	0.892874
r_2	0.775345	0.772642	0.772847	0.771398	0.770916	0.776278	0.780539
r_3	0.702429	0.691392	0.691440	0.689576	0.689808	0.695223	0.699495
coefficient	X	Y1	Y2	Y3	Y4	Y5	Y6
r_4	0.666409	0.651090	0.651051	0.649039	0.648977	0.654629	0.658978
r_5	0.660153	0.641974	0.641775	0.639914	0.640481	0.645611	0.650084
r_6	0.645215	0.625574	0.625232	0.623771	0.624155	0.629592	0.633519
r_7	0.634883	0.614179	0.613573	0.612855	0.613040	0.619091	0.622273
ρ_1	1	1	1	1	1	1	1
ρ_2	1	1	1	1	1	1	1
ρ_3	1	1	1	1	1	1	1
ρ_4	1	1	1	1	1	1	1
ρ_5	1	1	1	1	1	1	1
ρ_6	1	1	1	1	1	1	1
ρ_7	1	1	1	1	1	1	1

measured value is deviating from the mean, in the presence of the considered factor. Two dispersions can be obtained by measuring the dispersion in the presence and in the absence of the considered factor. The statistic comparison of these two dispersions (performed by statistic tests) can establish if the factor has a significant influence or not. If the measured value is a random variable with respect to the considered influence factors, the measured value usually tends to the mean value. Information regarding the influence of significant factors (or insignificant factors) together or apart can be obtained by estimating the dispersion of a andom variable in the simultaneous presence of factors and in their absence. This means a saving in time and materials and it avoids the study of some unimportant factors.

The F test is used for the comparison of dispersion obtained in the presence and in the absence of factors (see validation of statistical hypothesis).

Let us assume that Y is a random variable, depending on factors X^1, X^2, . . ., X^p, and having the mean value b^0. Usually, the matrix of factor values has a finite number of elements, named classes. One of these values is called the factor level.

Different hypotheses can be initially admitted in order to study the factors. The most simple and frequently used hypothesis is the one which satisfies the condition:

$$y = b^0 + \sum_{i=1}^{p} b^i \cdot X^i + e \qquad (4.1)$$

where e is the error.

Considering the variable $Y = y - b^0$ the following equation can be written:

$$Y = \sum_{i=1}^{p} b^i \cdot X^i + e \qquad (4.2)$$

where b^i are the considered statistic parameters, and X^i can be 0 or 1 according to the null hypothesis or positive factor effect.

If there are more determinations then Y is the vector of observations and the equation (2) can be written:

$$Y=X\cdot\beta+E \qquad (4.3)$$

where X is the matrix having only 0 and 1 as elements, β - the vector of estimated elements (factors), and E is the vector of residual variations, that means the vector of measurement error contribution to the observations. The last ones have a normal distribution and they are independent.

There are other hypotheses (models) less used, for example non-linear like a quadratic or polynomial equation. Even if the model is very complicated, the coefficients still remain of first degree and the problem is not complicated from a mathematical point of view.

Finally, the H_0 hypothesis is tested: $b_1 = b_2 = \ldots = b_p = 0$. This hypothesis is verified only when there is no effect of the considered factors that must be taking into account. The F test can be applied in two ways (see the statistical hypothesis verification):

a) In the case where we have reason to assume that $\sigma_1 > \sigma_2$, the F test verified only one confidence limit, the superior one:

$$F_{exp} = \frac{s_1^2}{s_2^2} > F_{\nu_1,\nu_2,\alpha} \tag{4.4}$$

where:

- s_1^2 and s_2^2 are the simple dispersions (the largest and the smallest) from which we estimated the theoretical dispersions σ_1^2 and σ_2^2;
- ν_1 and ν_2 is the number of degrees of freedom;
- F_{exp} is the calculated value from experimental data;
- $F_{\nu_1,\nu_2,\alpha}$ is the theoretical value (tabulated) of F distribution for chosen ν_1, ν_2 and incertitude α.

If we obtained $F_{exp} < F_{\nu_1,\nu_2,\alpha}$ then the two dispersions would be significantly different.

b). If there are no reasons to proceed at point a then the bilateral F test will be applied. In these cases, the hypothesis $\sigma_1 > \sigma_2$ is rejected if either of the following inequalities is satisfied:

$$\frac{s_1^2}{s_2^2} > F_{\nu_1,\nu_2,\frac{\alpha}{2}} \quad \text{or} \quad \frac{s_1}{s_2} < F_{\nu_1,\nu_2,\left(1-\frac{\alpha}{2}\right)} \tag{4.5}$$

Actually the ratio between the biggest and the lowest dispersion is compared with the tabulated $F_{\nu_1,\nu_2,\left(1-\frac{\alpha}{2}\right)}$ value.

If $F_{exp} \leq F_{\nu_1,\nu_2,\left(1-\frac{\alpha}{2}\right)}$ then the dispersions are considered equal and the considered factor has an insignificant effect. Usually, two levels are used; one inferior and the other superior for the considered factor reported to the start level (zero). The factor can be pH, temperature, the analyte nature /70,71/ etc. If there are n factors, 2^n experiments are necessary. For example, if n = 4 then 16 experiments are necessary. The experiment can be repeated for minimum three times in order to minimized the random errors. If n = 2 the experiment should be performed according to the following scheme /76/:

No experiment	Factor			Response
	x_0	x_1	x_2	
1	+1	+1	+1	y_1
2	+1	-1	+1	y_2
3	+1	+1	-1	y_3
4	+1	-1	+1	y_4

where the inferior level is +1 and the superior level is −1. Obviously, the order of experiments is not 1, 2, 3, 4 but the order given by random numbers. The scheme is equilibrated if all experiments are performed. A non-equilibrated scheme can be used as well, but the methods need special instructions /73/. The identification techniques of significant parameters can be dealt with by neuronal networks /74/. Also, the ANOVA technique can be used together with a verification method of linearity and a statistic test (F) /75/

5. VALIDATION OF STATISTIC HYPOTHESIS

The classical distributions are often models for physical and chemical phenomena. The advantage of using the distribution theory in the physical - chemical sciences is that the studied phenomenon causes the beginning of an analytical study in which the distributions are deeply studied /76/.

The most widely known distributions are:

beta	cauchy	chi squared	exponential	F	gamma
log normal	logistic	normal	student's t	uniform	weibull

Their analytical expressions and representations in different particular cases are presented in the appendix. An application of the distribution functions is also of use in checking the statistical hypotheses (see t, F, z, χ^2 tests etc.).

Sometimes it is useful to use a modified form of a distribution function /77/ in order to model the studied phenomenon. Also, some authors use the distribution in a technique known as "pattern–recognition" where, besides distributions, they use the neural networks /78/. A distribution model is often combined with an overlapping technique /79,80/.

5.1 The Fischer F Statistics

F1: Indicates the significance level of the regression equation. The F estimator is calculated with the relation:

$$F = \frac{SS_{reg}}{SS_e / (n-k)} \tag{5.1}$$

where SS_{reg} is the **sum of squares for errors** attributed to regression:

$$SS_{reg} = \sum_i (y_{i,calc} - \bar{y})^2 \tag{5.2}$$

and SS_e is the **sum** of s**quares for residuals**:

$$SS_e = \sum_{i=1}^{n} (y_{i,calc} - y_i)^2 \tag{5.3}$$

The calculated F value is compared (with eq. (4.1)) with the tabulated F value (depending on degree of freedom), corresponding to the needed significance level. If $F_{calc} > F_{tab}$, the result must be considered as significant. As a function of significance level, the quality of regression equations is

considered (significance level / regression quality) >99% / excellent; =99% / very good; 97.5 - 98% / good; 95% / satisfying; 90% / weak; <90% / unsatisfying /81/.

Generally, high values of F parameter indicate a good regression equation /81/.

F2: The F Test is also used /82/ to test whether one equation fits a set of data significantly better than does a second equation /75,83/. Before using this test it is necessary to fit the same data set using two different equations. List the results and note the reduced (chi-squared) values. By comparing these values it is possible to calculate the probability that the fits are the same /85,86/. A low probability value indicates that one of the two equations (that giving the lower reduced values) fits the data significantly better than the other. To be significant the probability should be lower than 0.1, and preferably lower than 0.05.

Suppose that we have a function:

FTEST(array1, array2)

where:

Array1 is the first array or range of data.

Array2 is the second array or range of data.

and

FTEST the return value of the F-Test.

In this case the F-test yields the one-tailed probability that the variances in array1 and array2 are not significantly different. Use this function to determine if two samples have different variances. Let us use:

FTEST({6,7,9,15,21},{20,28,31,38,40}) equals 0.648318

Example 4.

Suppose we have a series of numbers giving an amount:

Time	1	2	3	4	5	6
Amount	110	80	60	45	35	28

The precedent data were used, and fitted to a single exponential decay equation, and to an equation for a single exponential decay including a background offset.

Single exponential decay:

$$y = A \cdot e^{-kt} \qquad (5.4)$$

Single exponential decay including offset:

$$y = A \cdot e^{-kt} + \text{offset} \quad (5.5)$$

When fitted to a single exponential decay, the reduced chi-squared is 2.92; fitted to a single exponential decay including a background offset the reduced chi-squared is 0.09189. The second value is lower, but the difference between the two equations is not great, as is shown below:

Parameter	According to Equation (5.4)	According to Equation (5.55)
Reduced χvalue	2.82	0.09189
Number of parameters	2	3
Degrees of freedom	3	2

and: *F statistic* = 30.6889 *Probability* = 0.0634443

The probability that the two fits are equally appropriate is 0.06, which is low. We can therefore be reasonably confident that it is more appropriate to fit the data using the equation that includes a background offset (providing there is a theoretical or experimental justification for using this equation).

5.2 "t"–Student Statistics

Student's t-Test is used /87/ to compare two sets of data and to test the hypothesis that a difference in their means is significant. The data are tested with two underlying assumptions:

- They represent two independent normal distributions
- The values of their variances are equal

The t-Test is used in two ways: Independent and Paired.

Independent t-Test

The Independent t-Test can be used when two groups are thought to have the same overall variance but different means. It can provide support for a statement about how a given population varies from some ideal measure, for example how a treated group compares with an independent control group. In this case the t-test can be performed on data sets with an unequal number of points.

Paired t-Test

The Paired t-Test takes a paired approach, assuming that the variance for any point in one population is the same for the equivalent point in the second population. This test can be used to support conclusions about a treatment by comparing experimental results on a sample-by-sample basis. For example, to compare results for a single group before and after a treatment. This approach can help to evaluate two data sets whose means do not appear to be significantly different using the Independent t-Test. This test is only performed if the two data sets have an equal number of points.

In general, a t-test yields the probability associated with a set of data. The t-Test is used to determine whether two samples are likely to have come from the same two underlying populations that have the same mean.

As an extension, the t-test can be used to study the slope and the intercept /88/. In the case of regression the t estimator also indicates the signification level of the coefficients b_j; it is calculated by following equation /81/:

$$t_j = \frac{|b_j|}{\sigma_{bj}} \tag{5.6}$$

where σ_{bj} is the standard error of the b_j regression coefficients.

The validation or invalidation of the x_{ij} variable contribution at global correlation is realized by comparison of the calculated t_j values with a tabulated value (for an imperative significance level, which is a function of the degree of freedom for regression).

Multicollinearity means the existence of some g functional relationship between z variables from a total of m considered predictor variables, $2 \le z \le m$:

$$x_{ij} = g(x_{i,j+1},\ldots,x_{i,j+z-1}) \qquad i = 1,2,\ldots,n; \quad j = 1,2,\ldots,m \tag{5.7}$$

Actually, the predictor variables are orthogonal if any two of them are orthogonal, namely when the regression equations:

$$x_{ij} = \alpha x_{ik} + \beta, \qquad i = 1,2,\ldots,n; \qquad j,k = 1,2,\ldots,m; \qquad j \neq k \tag{5.8}$$

are characterized by $r^2 \leq 0.40$. The physical significance of a regression equation is compromised by the presence of multicollinearity. Therefore it is better to obtain a regression equation of good quality with a smaller number of predictor orthogonal values. In this case a good way is orthogonalizing the given set of variables /81/.

With Microsoft Excel, the t-tests are calculated as follows:

TTEST(array1, array2, tails, type)

where:

Array1 is the first data set.

Array2 is the second data set.

Tails specifies the number of distribution tails. If tails = 1, T-Test uses the one-tailed distribution. If tails = 2, T-Test uses the two-tailed distribution.

Type is the kind of t-test to perform.

A t-test of types 1, 2, 3 is performed in the following cases:

1 Paired

2 Two-sample equal variance (homoscedastic)

3 Two-sample unequal variance (heteroscedastic)

Example (calculated with Microsoft Excel):

TTEST({3,4,5,8,9,1,2,4,5},{6,19,3,2,14,4,5,17,1},2,1) equals 0.196016

If the t-test is performed with SlideWrite, it will perform independent and paired t-tests on your data. The paired test is not performed if the columns do not have the same number of points. SlideWrite reports the following results:

T-Statistic Measures the significance of the difference of the means.

p-Value The actual probability that the absolute value of the t-Statistic takes on its value or larger by chance.

Hypothesis SlideWrite compares the p-Value with the alpha Level and reports its conclusion:

When Hypothesis reports:

p-Value < alpha Level then there is *Difference*

p-Value ≥ alpha Level then there is *No Difference*

5.3 Z-Test

The z-test generates a standard score for x with respect to the data set, arrays, and returns the two-tailed probability for the normal distribution. You can use this function to assess the likelihood that a particular observation is drawn from a particular population. The Z-Test may be calculated as follows:

$$Ztest(array,x) = 1 - dnorm(\frac{\mu - x}{\sigma / \sqrt{n}}) \tag{5.9}$$

Example (calculate with Microsoft Excel):

ZTEST({3,6,7,8,6,5,4,2,1,9},4) equals 0.090574

where: {3,6,7,8,6,5,4,2,1,9} is the array of data against which to test x; and 4 is the value to test (x);

Instead of σ the population the known standard deviation s value (the sample standard deviation) may be used.

5.4 χ^2 - test

The previous tests described have, in general, been concerned with testing whether the mean of several observations differs significantly from the value proposed by the null hypothesis. The data used have been on a continuous scale. In contrast, χ^2 is concerned with frequencies (i.e. the number of times a given event occurs /13/. The χ^2 - test is a test for independence. We can use the χ^2-test to determine if results based on a hypothesis are verified by an experiment /89/. The most frequent application of the χ^2-test is the quality of fit test which compares an observed distribution with a theoretical distribution. Suppose a function, named CHITEST with the following syntax:

CHITEST(series_actual, series_expected)

and where:

actual_series is the series of data that contains observations to test against expected values;

expected_series is the series of data that contains the ratio of the product of row totals and column totals to the grand total.

The function returns the value from the chi-squared (χ^2) distribution for the statistic and the appropriate degrees of freedom. The χ^2 test first calculates a χ^2 statistic and then sums the differences of actual values from

the expected values. The equation for this function is CHITEST=p($X>\chi^2$), where:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{\left(A_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \qquad (5.10)$$

and with:

Aij = actual frequency in the i-th row, j-th column

Eij = expected frequency in the i-th row, j-th column

r = number or rows

c = number of columns

Observations: if CHITEST returns the probability for a χ^2 statistic and degrees of freedom df, then df is given by df := (r - 1)(c - 1).

Example 5

Thin layer chromatography and liquid chromatography of 200 compounds was performed. The results are grouped in 10 classes and the class width is $0.1R_f$ units. The observed frequencies are:

n	1	2	3	4	5	6	7	8	9	10
TLC	17	22	25	16	15	21	12	23	29	20
lq	18	20	23	19	18	21	14	17	22	18

Now we test with χ^2 statistics if the observed frequencies correspond to a rectangular distribution. Using (5.10), the χ^2 statistic for the data above is 17.384 with 9 degrees of freedom.

CHITEST(A,E) equals 0.083018

5.5 Other tests

Chromatographic Resolution Statistic is a response function among the various response functions used to numerically assess the quality of a separation //90-92/. There is a deeper concern for peak distributions /93/. The next tests can be used for the verification of peak characteristic /94/ of any distribution (see appendix):

Kurtosis Test

Yields the kurtosis of a data set. Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. Positive kurtosis indicates a relatively peaked distribution. Negative kurtosis indicates a relatively flat distribution.

Kurtosis is defined as:

$$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{X_i - M(X)}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (5.11)$$

where: s is the sample standard deviation.

Example (calculated with Microsoft Excel):

KURT(3,4,5,2,3,4,5,6,4,7) returns -0.1518

Skew Test

Yields the skewness of a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean.

Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values. Negative skewness indicates a distribution with an asymmetric tail extending towards more negative values.

The equation for skewness is defined as:

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{X_i - M(X)}{s}\right)^3 \quad (5.12)$$

Example (calculated with Microsoft Excel):

SKEW(3,4,5,2,3,4,5,6,4,7) equals 0.359543

A less known test, Wilk's lambda is combined with the analysis of main components to calculate the discrimination power of variables /95/. The statistical tests are described in detail by many books and papers /51/.

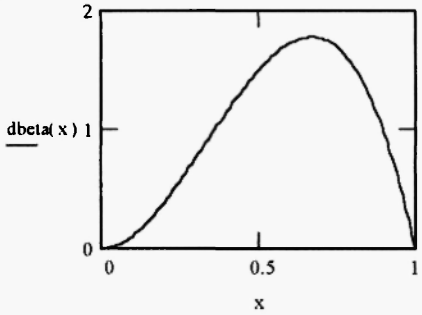
Appendix 1. Statistical distributions and their analytical forms.

$$dbeta(x) := \frac{\Gamma(s1 + s2)}{\Gamma(s1)\cdot\Gamma(s2)}\cdot x^{s1-1}\cdot(1-x)^{s2-1}$$

$s1 := 3$

$s2 := 2$

$x := 0, 0.01 .. 1$

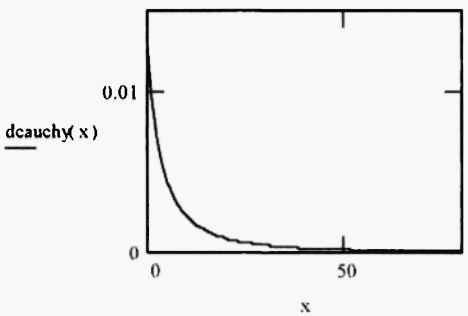


$$dcauchy(x) := \left[\pi\cdot s\cdot\left(1+\frac{x-l}{s}\right)^2\right]^{-1}$$

$s := 2$

$l := 5$

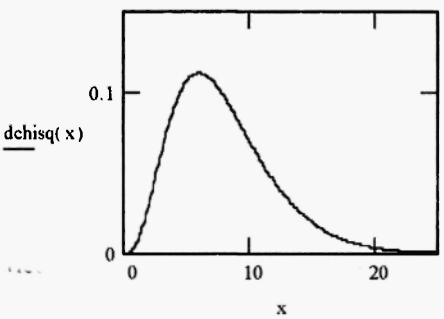
$x := 0, 0.5 .. 80$



$$dchisq(x) := \frac{\exp\left(-\frac{x}{2}\right)}{2\cdot\Gamma\left(\frac{d}{2}\right)}\cdot\left(\frac{x}{2}\right)^{\frac{d}{2}-1}$$

$d := 8$

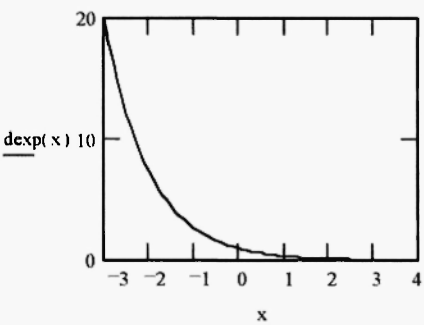
$x := 0, 0.1 .. 25$



$$dexp(x) := r\cdot\exp(-r\cdot x)$$

$r := 1$

$x := -3, -2.9 .. 4$

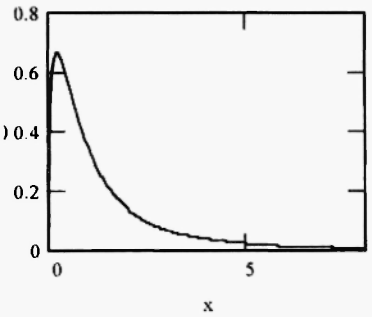


$$df(x) := \frac{d1^{\frac{d1}{2}} \cdot d2^{\frac{d2}{2}} \cdot \Gamma\left(\frac{d1+d2}{2}\right)}{\Gamma\left(\frac{d1}{2}\right)\cdot\Gamma\left(\frac{d2}{2}\right)} \cdot \frac{x^{\frac{d1-2}{2}}}{(d2+d1\cdot x)^{\frac{d1+d2}{2}}}$$

$d1 := 3$

$d2 := 4$

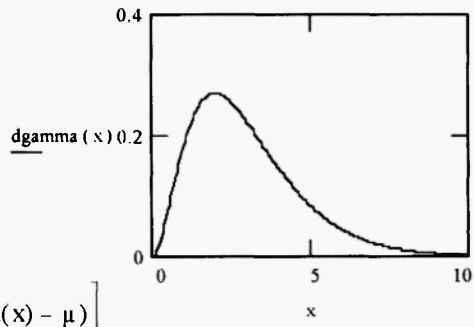
$x := 0, 0.05 .. 8$



$$dgamma(x) := \frac{x^{s-1}\cdot\exp((-x))}{\Gamma(s)}$$

$s := 3$

$x := 0, 0.05 .. 10$

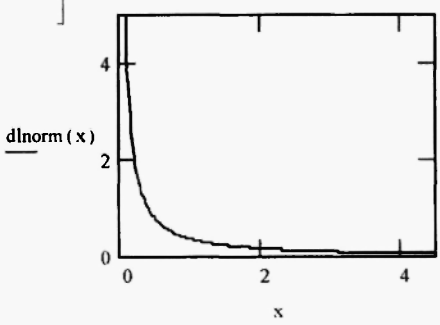


$$dlnorm(x) := \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma\cdot x}\cdot\exp\left[-\frac{1}{2\cdot\sigma^2}\cdot(\ln(x)-\mu)\right]$$

$\sigma := 2$

$\mu := 5$

$x := 0.1, 0.12 .. 4.5$

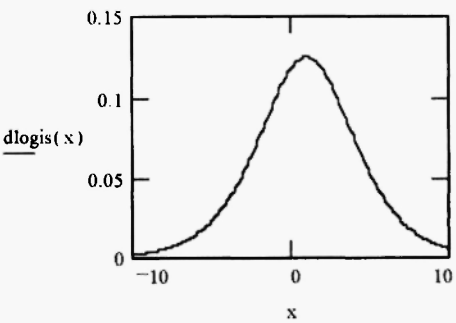


$$dlogis(x) := \frac{\exp\left(-\frac{x-l}{s}\right)}{s\cdot\left(1+\exp\left(-\frac{x-l}{s}\right)\right)^2}$$

$s := 2$

$l := 1$

$x := -10, -9.9 .. 10$

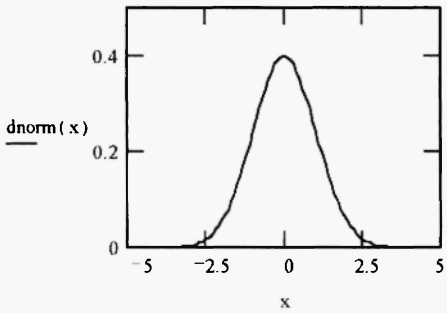


$$dnorm(x) := \frac{1}{\sqrt{2\cdot\pi}\cdot d}\cdot\exp\left[-\frac{(x-m)^2}{2\cdot d^2}\right]$$

$m := 0$

$d := 1$

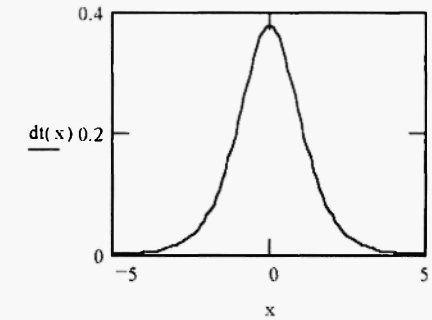
$x := -5, -4.9 .. 5$



$$dt(x) := \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\cdot\sqrt{\pi\cdot d}}\cdot\left(1+\frac{x^2}{d}\right)^{-\frac{d+1}{2}}$$

$d := 5$

$x := -5, -4.9 .. 5$

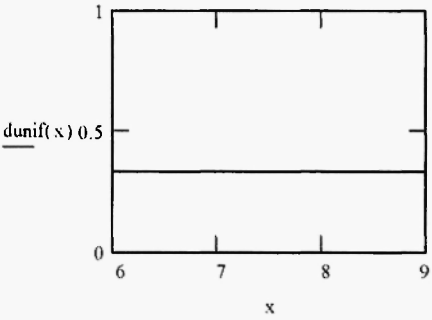


$$dunif(x) := \frac{1}{b-a}$$

$a := 6$

$b := 9$

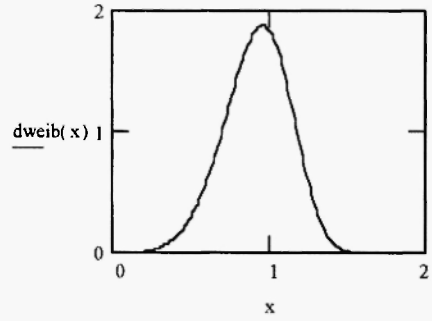
$x := 6, 6.1 .. 9$



$$dweib(x) := s\cdot x^{s-1}\cdot\exp\left(-x^s\right)$$

$s := 5$

$x := 0, 0.01 .. 2$



ENDNOTES

[1] G.U.Yule, M.G.Kendall, *Introduction in statistical theory*, 5th edition, New York, 1968.
[2] G. Corrado, *New statistical index for dispersion*, Rome, 1967.
[3] C. Cimpoiu, L. Jäntschi, T.Hodi°an, *J. Planar Cromatogr.*, **11**(3), 122(1998).
[4] C. Cimpoiu, L. Jäntschi, T.Hodi°an, J. Liq. Cromatogr., Rel. tehn, (1998), in press.
[5] A. Glück, *Mathematical metods in chemical industry*, Techn. Ed., Bucharest, 1971.
[6] G.Barko, J. Hlavay, *Talanta*, **44**, 2237(1997).
[7] S. Ellison, W. Wegschneider, A. Williams, *Anal. Chem.*, 607A(1997).
[8] C. Liteanu, S. Gocan, *Stud. Univ. Babe°-Bolyai, Chem.*, **1969**(1), 29.
[9] O.L.Davies, P.L.Goldsmith, *Statistical metods in research and production*, Longmans, London, 1982.
[10] G.Niac, O.Horovitz, *Indrumator pentru lucrari de laborator*, Tech.Univ.Publ.House, Cluj-Napoca, 1982.
[11] A.J.Thiel, A.G.Frutos, C.E.Jordan, R.M.Corn, L.M.Smith, *Anal. Chem.*, **69**, 4948(1997).
[12] G.Stubauer, T.Seppi, P.Kukas, D.Obendorf, *Anal. Chem* , **69**, 4469(1997).
[13] V.M.Morris, J.G.Huges, P.J.Marriott, *J. Chromatogr*, **755**(2), 235(1996).
[14] J.A.Backwell, R.W.Stringham, J.D.Weckwerth, *Anal. Chem.*, **69**, 409(1997).
[15] C.Sârbu, L.Jäntschi, *Rev.Rom.Chim* , Bucharest, **49**(1), 19(1998).
[16] Moritz H., *Advanced physical Geodesy*, Herbert Wichman Verlag, 1980.
[17] B. Arne, *Theory of errors on generalized matrix inverses*, Elvister, Amsterdam-London-New York, 1973.
[18] N.V.Brandin, *Osnovî eksperimentalnoi kosmiceskoi balistiki*, Ma°inostroenie, Moscow, 1974.
[19] K.Liubov, *Matematiceskie osnovî kibernetiki*, Vis°aia °kola, Kiev, 1977.
[20] Tiron M., Errors theory and least squares method, Tech. Ed., Bucharest, 1972.
[21] D.Lorber, K.Faber, R.Kowalski, *Anal. Chem.*, 69, 1620 (1997).
[22] R.Gilbert, R.Goodacre, A.M.Woodward, D.B.Kell, *Anal. Chem* ,**69**(21), 4381(1997).

BIBLIOGRAPHY

1. G.U. Yule, M.G. Kendall, *Introduction in Statistical Theory*, 5th edition, New York, 1968.
2. G. Corrado, *New statistical index for dispersion*, Rome, 1967.
3. C. Cimpoiu, L. Jäntschi, T. Hodişan, *J. Planar Chromatogr.*, **11**(3), 91 (1998).
4. C. Cimpoiu, L. Jäntschi, T. Hodişan, *J. Liq. Cromatogr.* (1998), in press.
5. A. Glück, *Mathematical Methods in Chemical Industry*, Techn. Ed., Bucharest, 1971.
6. G. Barko, J. Hlavay, *Talanta*, **44**, 2237 (1997).
7. S. Ellison, W. Wegschneider, A. Williams, *Anal. Chem.*, **69**, 607A(1997).
8. C. Liteanu, S. Gocan, *Stud. Univ. Babeş-Bolyai, Chem.*, **1969**(1), 29.
9. O.L. Davies, P.L. Goldsmith, *Statistical Methods in Research and Production*, Longmans, London, 1982.
10. D.A. Skoog, D.M. West, *Fundamentals of Analytical Chemistry*, 4th edition, Holt Saunders, New York, 1982.
11. J. Topping, *Errors of Observation and their Treatment*, Chapman & Hall, London, 1962.
12. L. Tovissi, V. Voda, *Metode de analiza statistica*, Ed. Stiinþ. Encicl., Bucharest, 1982.
13. J.C. Miller and J.M. Miller, *Statistics for Analytical Chemistry*, 2nd ed., Wiley, New York, 1987.
14. G. Niac, O. Horovitz, *Indrumator pentru lucrari de laborator*, Tech.Univ.Publ.House, Cluj-Napoca, 1982.
15. A.J. Thiel, A.G. Frutos, C.E. Jordan, R.M. Corn, L.M. Smith, *Anal. Chem.*, **69**, 4948 (1997).
16. G. Stubauer, T. Seppi, P. Lukas, D. Obendorf, *Anal. Chem.*, **69**, 4469(1997).
17. V.M. Morris, J.G. Hughes, P.J. Marriott, *J. Chromatogr*, **755**(2), 235(1996).
18. J.A. Blackwell, R.W. Stringham, J.D. Weckwerth, *Anal. Chem.*, **69**, 409(1997).
19. C. Sârbu, L. Jäntschi, *Rev. Rom. Chim., Bucharest*, **49**(1), 19(1998).
20. H. Moritz, *Advanced Physical Geodesy*, Herbert Wichman Verlag, 1980.

21. B. Arne, *Theory of Errors on Generalized Matrix Inverses*, Elsevier, Amsterdam-London-New York, 1973.
22. N.V. Brandin, *Osnovî eksperimentalnoi kosmicheskoi balistiki,* Maşhinostroenie, Moscow, 1974.
23. K. Lyubov, *Matematicheskie osnovî kibernetiki,* Visşhaia Şhkola, Kiev, 1977.
24. M. Tiron, *Errors Theory and Least Squares Method,* Tech. Ed., Bucharest, 1972.
25. D. Lorber, K. Faber, R. Kowalski, *Anal. Chem.*, 69, 1620 (1997).
26. W. Lindberg, J.A. Persson, S. Wold, *Anal. Chem.*, **55**, 643(1983) .
27. M. Otto, W .Wegscheider, *Anal. Chem.*, **57**, 63(1985).
28. B.R. Kowalski, M.B. Seasholtz, *J. Chemometrics*, **5**, 129(1991).
29. K.S. Booksh, B.R. Kowalski, *Anal. Chem.*, 66, 782A(1994).
30. P.P. Wenzell, D.T. Andrews, B.R. Kowalski, *Anal. Chem.*, **69**, 2299 (1997).
31. L. Xu, I. Schechter, *Anal. Chem.*, **69**, 3722(1997).
32. M.P. Nelson, J.F. Aust, J.A. Dobrowolski, P.G. Verly, M.L. Myrick, *Anal. Chem.*, **70**(1), 73(1998).
33. P.J. Jackson, M.R. Schure, T.P. Weber, P.W. Carr, *Anal. Chem.*, **69**, 416(1997).
34. M. Otto, U. Hoerchner, W. Wegschneider, *J. Chromatography*, **485**, 453(1989).
35. U. Haldna, J. Pentchuk, M. Righezza, J.R. Chretien, *J. Chromatog.*, **670**, 51(1994).
36. R.J. Bartelt, *Anal. Chem.*, **69**(3), 364(1997).
37. J.C. Dore, J. Pothier, J. Galand, C. Viel, *Analysis*, **23**, 342(1995).
38. P.L. Bonate, *LC-GC*, **10**(7), 531(1992).
39. H. Markus, T. Naes, *Multivariate Calibrations*, Wiley, New York, 1989.
40. I.T. Jalliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
41. S.T. Balke, *J. Appl. Polym. Sci.*, **43**, 5(1989).
42. R. Kaliszan, *Chemom. Intell. Lab. Syst.*, **24**(2), 89(1994).
43. J.C. Berridge, P. Jones, A.S. Roberts-McIntosh, *J. Pharm. Biomed. Anal.*, 9(8), 597(1991).
44. C.H. Spiegelman, M.J. McShane, G.L. Cote, M.J. Goetz, M. Motamedi, Q.L. Yule, *Anal. Chem.*, **70**, 35(1998).
45. K.T. Kinnear, H.G. Monbouquette, *Anal. Chem.*, 69(9), 1771(1997).

46. E.R. Collantes, R. Duta, W.J. Welsh, W.L. Zielinski, J. Brower, *Anal. Chem.*, **69**(7), 1392(1997).
47. J. Zupan, J. Gasteiger, *Neural Network for Chemists*, VCH Publishers, New York, 1993.
48. T Hodişan, M. Curtui, I. Haiduc, *J. Radioanal. Nucl. Chem.*, in press.
49. T. Hodişan, M. Curtui, S. Cobzac, C. Cimpoiu, I. Haiduc, *J. Radioanal. Nucl. Chem.*, in press.
50. K.S. Johnston, S.S. Yeo, K.S. Booksh, *Anal. Chem.*, **69**(10), 1844(1997).
51. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, *Chemometrics: A Textbook*, Elsevier, Amsterdam, 1988.
52. R.G. Brereton, Chemometrics: *Applications of Mathematics and Statistics to the Laboratory*, Ellis Horwood, Chichester, 1990.
53. J. Einax, H. Zwanziger, S. Geiss, *Chemometrics in Environmental Analysis*, Wiley, Chichester, 1997.
54. M. Meloun, J. Mlitky, M. Forina, *Chemometrics for Analytical Chemistry, vol I: PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester, 1994.
55. ***, Infometrix. *Chemom. Appl. Overview*, **14**(4), 1993.
56. I. Todoran, *Rãspunsuri posibile, corelaþie °i prognozã*, Dacia Ed., Cluj-Napoca, 1989.
57. Z. Gong, Z. Zhang, X. Yang, *Analyst*, **122**, 283(1997).
58. S. Fournout, R. Rouquet, S.L. Salhi, R. Seyer, V. Valverde, J.M. Masson, P. Jouin, B. Pau, M. Nicolas, V. Hanin, *Anal. Chem.*, **69**(9), 1746(1997).
59. H. Xie, R.M. Moore, *Anal. Chem.*, **69**(9), 1753(1997).
60. D.A. Holman, G.D. Christian, J.K. Ruzicka, *Anal. Chem.*, **69**(9), 1763(1997).
61. D.E. Pivonka, T.R. Simpson, *Anal. Chem.*, **69**, 3851(1997).
62. R. Gilbert, R. Goodacre, A.M. Woodward, D.B. Kell, *Anal. Chem.*, **69**(21), 4381(1997).
63. S.M. Clark, R.A. Mathies, *Anal. Chem.*, **69**, 1355 (1997).
64. G.E.P. Box, B. Wilson, *J. R. Statist. Soc.*, **B13**, 1(1951).
65 I. Vaduva, *Analiza Dispersionalã*, Ed. Techn., Bucharest, 1970.
66. W.G. Cochran, G. Cox, *Experimental Designs*, 2nd ed., Wiley, New York, 1957.
67. L. Mutihac, V. David, *Chemometrie*, Univ. of Bucharest Publ. House, Bucharest, 1997.

68. I. Marinescu, *Analiza Factorialã*, Sci. and Encycl. Publ. House, Bucharest, 1984.
69. V. Vodă; *Metode de analizã statisticã*, Sci. Publ. House, Bucharest, 1982.
70. I.A. Stuart, R.O. Ansell, J. Maclachan, P.A. Bather, W.P. Gardiner, *Analyst*, **122**, 303(1997).
71. G. Maio, C. von Holst, B.W. Wenclawiak, R. Darskus, *Anal. Chem.*, **69**(4), 601(1997).
72. D.L. Duewer, L.A. Currie, D.J. Reeder, S.D. Leigh, J.J. Filliben, H.K.Liu, J.Mudd, *Anal. Chem.*, 69, 1882(1997).
73. D.C. Hoaglin, F. Mosteller, J.W. Tukey, *Fundamentals of Exploratory Analysis of Variance*, Wiley, New York, 1991.
74. J.M. Sutter, P. Jurs, *Anal. Chem.*, 69, 856(1997).
75. J.A.M. Pulgarin, A.A. Molina, P.F. Lopez, *Analyst*, **122**, 247(1997).
76. G. Niac, Al. Popescu, I. Bolocan, P. Capotă, *Granulometric Analysis of Işalnita Power-Station Ashes and Elementary Composition of Classes,* Workshop - Chemometrie, Timişoara (Romania), 25-26 Sept 1997, SCAR.
77. J.V. Selinger, S.Y. Rabbany, *Anal. Chem.*, 69, 170(1997).
78. A.S. Bangalore, G.W. Small, R.J. Combs, R.B. Knapp, R.T. Kroutil, C.A. Traynor, J.D. Ko, *Anal. Chem.*, **69**, 118(1997).
79. J.M. Davis, *Anal. Chem.*, 69, 3796(1997).
80. F. Deridi, A. Bassi, A. Cavazzini, M.C. Pietrogrande, *Anal. Chem.*, **70**(4), 766(1998).
81. M.V. Diudea, O. Ivanciuc, *Molecular Topology*, Comprex Publ. House, Cluj-Napoca, 1995.
82. G.A. Eiceman, D. Preston, G. Tiano, J. Rodriguez, J.E. Parmeter, *Talanta*, **45**, 57(1997).
83. V.I. Slaveykova, M. Hoenig, *Analyst*, **122**, 247(1997).
84. D. Dons, M.H. Ramsey, I. Thornton, *Analyst*, **122**, 421(1997).
85. S.K. Poole, C.F. Poole, *Analyst*, **122**, 267(1997).
86. J.D. Krass, B. Jastorff, H.G. Genieser, *Anal. Chem.*, 69(13), 2575(1997).
87. S.Sopok, *J. Chromatogr.*, 739(1-2), 163(1992).
88. J. Medina-Escriche, A. Seviliano-Cabeza, M.A. Martin-Penella, *Analyst*, **110**, 807(1985).
89. S. Canepari, V. Carunchio, P. Castellano, A. Messina, *Talanta*, **44**, 2059(1997).

90. T.D. Schlabach, J.L. Excoffier, *J. Chromatogr.*, **439**, 173(1988).
91. J.L. Glajch, L.R. Snyder, *Computer Assisted Method Development for High-Performance Liquid Chromatography*, Elsevier, New York, 1990.
92. J.C. Berridge, *Techniques for the Automated Optimization of HPLC Separation*, Wiley, New York, 1986
93. J. Li, *Anal. Chem.*, **69**, 4452(1997).
94. C. Liteanu, I. Râcă, *Anal. Chem.*, **51**, 12(1979).
95. M.I. Guerrero, C.H. Pagliai, A.M. Camean, A.M. Troncoso, A.G. Gonzalez, *Talanta*, **45**, 379(1997).