

Concentration of Analytical Data as Part of Data Processing in Trace Element Analysis* **

B. Sansoni, R. K. Iyer*** and R. Kurth

Central Department for Chemical Analysis, Kernforschungsanlage Jülich GmbH, D-5170 Jülich-1, Federal Republic of Germany

Konzentrierung analytischer Daten als Teil der Datenverarbeitung in der Spurenelementanalyse

Zusammenfassung. Die statistische Beurteilung von Daten der Spurenelementanalyse war lange Zeit durch das Gesetz der Fehler-Normalverteilung geprägt, welches auf der Annahme einer Gaußschen Normalverteilung beruht. Wird der analytische Fehler der Konzentrationsbestimmung von Spurenelementen vernachlässigbar klein gegenüber den tatsächlichen Schwankungen in der Natur, dann werden schiefe Verteilungen häufig wichtiger als Normalverteilungen. Eine gute Näherung ist oft, allerdings nicht immer, die Lognormalverteilung. Der Spurenanalytiker übersieht häufig auch heute noch, vor Beginn der eigentlichen Datenbeurteilung den Typ der vorliegenden Häufigkeitsverteilung seines Datensatzes zu untersuchen.

Die Datenkonzentrierung erfolgt in der hiesigen Zentralabteilung routinemäßig vor allem mit den Programmen ZCH-2 und ZCH-3/1. Darin wird der Typ der Datenverteilung a) durch Zeichnen der Histogramme, Auftragung als Hazens Gerade in Wahrscheinlichkeitskoordinaten für Normal- und Lognormalverteilung, den Kolmogorov-Smirnov- und Cramer-van-Mises-Anpassungstests sowie Schiefe und Wölbung beurteilt. Bei Quasi-Normalverteilungen können b) Ausreißer durch den t -, Nalimovs t -, Grubbs- und Dixon-Test eliminiert werden. Zentralwerte zur Lokalisierung des Datenkollektives sind c) arithmetisches, geometrisches und harmonisches Mittel sowie Median. Die Streuung um die Mittelwerte wird durch d) Varianz,

Standardabweichung für Einzelwert sowie Mittel, relativen Variationskoeffizienten, geometrische Standardabweichung, Bereich und 80% Interdezil-Bereich charakterisiert. Zusätzlich werden Vertrauensbereiche für das arithmetische und geometrische Mittel und den Median berechnet.

Als typische Probleme der Datenkonzentrierung werden die Abweichungen von der Normalverteilung, ein Vergleich verschiedener Mittelwerte untereinander, die Eliminierung von Ausreißern, die Behandlung von Konzentrationen unterhalb der Nachweisgrenze sowie Homogenität und Heterogenität der Datenkollektive behandelt. Dies wird an Datenreihen aus früheren Untersuchungen eines der Verfasser über Spurenelementkonzentrationen in der Luft, den Bleigehalt in Zahnstein, die Auslaugung toxischer Schwermetalle aus Keramikgeschirr und Gamma-Dosisleistungen aus einem Gebiet der Bundesrepublik mit höherer natürlicher Untergrundradioaktivität erläutert.

Als wesentliches Ergebnis soll die Arbeit dazu beitragen, daß der Spurenanalytiker das „Mysterium der Normalverteilung“ überwindet und die Behandlung schiefer Datenverteilungen sowohl in das Repertoire seiner Routineuntersuchungen als auch in sein analytisches Denken aufnimmt. Um einen größeren Kreis von Spurenanalytikern anzusprechen, ist die Darstellung elementar gehalten.

Summary. Statistical data evaluation in trace element analysis was mainly influenced by the Normal Law of Error, which is based on the assumption of a Gaussian normal distribution. Supposing that the analytical error of the trace element concentration is negligibly small compared with its true variations in nature, skew distributions may become more important than normal ones. Often, but not always, the lognormal distribution is a good approximation of the skew distribution. Careful investigation of the type of distribution before starting data evaluation in trace element analysis is still frequently overlooked today.

In this Department, data concentration is performed on a routine basis by two programmes ZCH-2 and ZCH-3/1. They include a) investigation of the type

* 2nd Contribution to the Principles of Trace Analysis of Elements and Radionuclides. 1st Contribution Lit. [33], see also [34, 35]. Parts of this publication have been presented at the AGF-Seminar, Bonn, October 1979; International Symposium on Management of Environment (plenary lecture), Bombay, February 1980; Annual Symposium on the Anal. Chem. of Pollutants, Dortmund, May 1980; 8th International Microchemical Symposium, Graz, August 1980

** Dedicated to Prof. Dr. E. Blasius, on occasion of his 60th birthday

*** Guest scientist from the Analytical Chemistry Division, Bhabha Atomic Research Centre, Bombay, India

of distribution by drawing the histogrammes, probability plot for normal and lognormal distribution as Hazen's straight lines, Kolmogorov-Smirnov- and Cramer- van Mises goodness-of-fit tests as well as skewness and kurtosis. Outliers of the normal distribution can be eliminated b) by t -, Nalimov's r -, Grubb's- and Dixon's tests. c) As central values for data location, the arithmetic, geometric, and harmonic means and median are calculated. d) The dispersion around the mean is characterized by variance, standard deviation of a single value as well as the mean, relative coefficient of variation, mean deviation from the arithmetic mean, geometric standard deviation, range and 80% inter-decile range. Confidence intervals are given for the arithmetic mean, geometric mean and median.

Typical problems of data concentration are the deviation from the normal distribution, comparison of different mean values, outlier elimination, concentrations below detection limit, homogeneity and heterogeneity of the data sample. They are discussed for examples of data series from one of the author's laboratories, e.g. trace element concentrations in air, lead content in dental calculus, toxic heavy metal leaching from ceramic ware and gamma dose rates from an area of higher natural radioactivity in the Federal Republic of Germany.

As a conclusion, it is emphasized that the trace element analyst should overcome the "Mystery of the Normal and Quasi-normal Distribution" and include skew data distributions and their statistical treatment into his repertoire of routine procedures as well as in his way of thinking.

Key words: Spurenanalyse; Datenkonzentrierung, Datenverarbeitung

1. Introduction

For a decade it has been realized [25, 26, 28, 30c, 30d, 33, 35, 38, 40, 41] that the main topics for future development in trace element analysis might be sampling [34], sample preparation [13] including ashing [23] and dissolution, data processing and evaluation rather than new methods for preconcentration, separation and physical measurement (Table 1).

With respect to data processing and evaluation, training courses have been organized since 1970, together with Professor Dr. G. Gottschalk, in the author's laboratory at GSF-Neuherberg [8, 27, 30a]. They have been concerned mainly with elementary statistics, simple data reduction, variance and regression analysis together with a careful estimation of the reliability of the analytical results and their corresponding analytical methods. In most of these cases, the data

distributions have been of the normal or approximated Gaussian normal type.

In the meantime, large series of trace element concentrations, contents and radioactivities have been analysed [26, 30e, 32, 37, 38, 40, 43]. The concentration of these data into a few characteristic statistics was inevitable, in order to evaluate and compare the large amount of data from environmental, biological, and geochemical materials and from intercomparison runs. After a few initial failures, however, it was recognized, how important it is, before starting the concentration or reduction of the data, to carefully investigate the type of distribution, the homogeneity of the data sample and to homogenize heterogeneous samples. Logarithmic or other skew distributions, which can be approximated by a lognormal distribution, have been observed more often than the Gaussian normal distribution. This has a strong influence on selecting the appropriate procedures for data concentration [29, 30a, 30b, 30e, 30f].

In the following, firstly some fundamentals on statistical data concentration are given, secondly the routine computer programme ZCH-3/1, which is in use since two years at the Central Department for Chemical Analysis at Jülich, is discussed [36]. In the third part, typical problems arising in data concentration are presented in an elementary way, in order to also convince the trace analyst, who is not familiar with statistical methods. Special consideration is given to the type of distribution and its location rather than to its dispersion.

An introduction to statistical data evaluation can be found in [3-5, 7, 9, 10, 15, 16, 18-22, 24, 27]. Data concentration or reduction is treated especially in [1, 10, 42], distribution-free statistics and order statistics in [2, 10, 22, 36].

2. Statistical Background

2.1 Data Processing

In the general scheme of trace element analysis according to Table 1, the physical measurement produces the analytical signal, which is a function of concentration, content or radioactivity of the component to be determined in the material sample under investigation [24, 28]. The signal is usually transformed within the measuring instrument by analog or digital procedures into numerical values.

These fundamental analytical raw data have to be transformed in several ways. 1) The results form the data sample. 2) They are converted into concentrations, contents or activity values. 3) Repeated measurements have to be carried out under exactly the same conditions in order to increase their precision. These data have to be concentrated into a mean value and its dispersion is determined. 4) The accuracy of this mean

Table 1. General approach of trace element analysis [25, 26, 41]

Definition of the question to be answered
1. Sampling
2. Sample preparation
3. Ashing
4. Dissolution
5. Preconcentration
6. Separation
7. Preparation of measuring sample
8. Physical measurement of the analytical signal
– integral
– spectral
– calibration
9. Data processing
10. Data evaluation
11. Quality control
Answering the question

value is estimated according to different calibration methods. 3) and 4) together represent the total error of the analytical data. Large data series of concentrations, contents or radioactivities are obtained within biological, medical, geochemical and environmental systems as well as in intercomparison runs. 5) They are concentrated into a few statistics, which characterize the whole data sample and are utilized for evaluation and comparison with similar series of measurements. The total analytical error of the data is assumed to be negligibly small compared to the natural variations, as a first approximation.

From this it becomes obvious, that two different types of data concentration are relevant, which differ in the type of distribution involved and, therefore, in the procedures. In 3) usually a normal distribution can be assumed, where the arithmetic mean \bar{x} , standard deviation s or the relative variation coefficient V_r characterize the whole data sample [4, 5, 7–9, 14–16]. With respect to 5), when the Gaussian normal distribution is not valid, data concentration becomes more complex. This shall be demonstrated in the following instance.

2.2 Data Sample

As an example for a series of different and large data samples, air dust data from the northern border of Munich, measured by one of the authors together with his coworkers from GSF between 1971 and 1976 shall be considered.

Air dust has been collected as daily 24 h samples on millipore filters EAWP 047 (2.7 cm diameter, 1 μ m porosity) at an altitude of 5 m above ground level at GSF, Neuherberg, about 110 m east of

Bundesstraße B13 in front of the Department for Radiation Protection [32].

The daily air dust samples have been analysed for the air dust by weighing, for the ash content by oxygen plasma ashing and for the element contents of lead, cadmium, zinc and calcium by atomic absorption spectrometry [43]. In addition, several meteorological parameters such as wind direction, wind speed and precipitation have been measured. The absolute amounts have been related to the volume of air, to the dust weight and to the ash weight of the dust. This results in daily concentrations of the components in air, contents in air dust and ash of dust. They amount together with the meteorological parameters to about 17,000 data per year. Table 2 presents a part of these data in form of a table. It is obvious, that this enormous amount of data cannot be assessed by simply reading the computer outputs.

The graphical representation of the individual values in a diagram, according to Fig. 1, allows a much better qualitative survey, from which e.g. seasonal variations and other general trends can be easily estimated in a qualitative manner. It is remarkable, that Fig. 1 concentrates as many as about 6,100 individual measurements in one single diagram.

2.3 Data Concentration

For a quantitative treatment, the huge amount of data, even within the graphical representation, has to be concentrated before further evaluation can be made. Instead of data “concentration”, sometimes the term data “reduction” is used. The authors, however, prefer “concentration”, because it is more characteristic of this operation than “reduction”.

The usual way is to compare the annual or monthly arithmetic means \bar{x} , together with the corresponding standard deviation s or the relative variation coefficient V_r , either as a table (Table 3) or as a block diagram (Fig. 2). The mean values in the table allow a quantitative survey over the large data material, whereas from the diagram a general qualitative trend is more easily seen. It was already mentioned that \bar{x} , s and V_r are appropriate only in the case of a Gaussian normal distribution. If this is not fulfilled, Table 3 as well as Fig. 1 will be incorrect!

In general, each analytical data sample, which belongs to a statistically homogeneous population, can be concentrated into a few statistics, which characterize the main properties of the sample without presenting every single value. The sample, regardless of its type of distribution, can be characterized by the

- location of the centre of gravity of the data (central or mean value),
- deviation from this central value (dispersion) and
- shape of sample distribution.

The

d) type of distribution has to be estimated in order to allow correct estimate of a) and b). Furthermore,

Table 2. Numerical representation of data samples as computer output. Air dust monitoring data from the northern border of Munich (GSF, Neuherberg), 1975. The different data samples include the concentration of dust, air dust ash, lead, cadmium, zinc, calcium in air dust; wind direction, wind speed, precipitation. Metal determinations by atomic absorption spectrometry according to [32] and [37]

Table with multiple columns: Datum, Konzentration in der Luft, Gehalt der Luft, meteorologische Daten, Konzentration in der Luft, Gehalt der Luft, meteorologische Daten, Konzentration in der Luft, Gehalt der Luft, meteorologische Daten, Konzentration in der Luft, Gehalt der Luft, meteorologische Daten. Each column contains dense numerical data for various parameters over time.

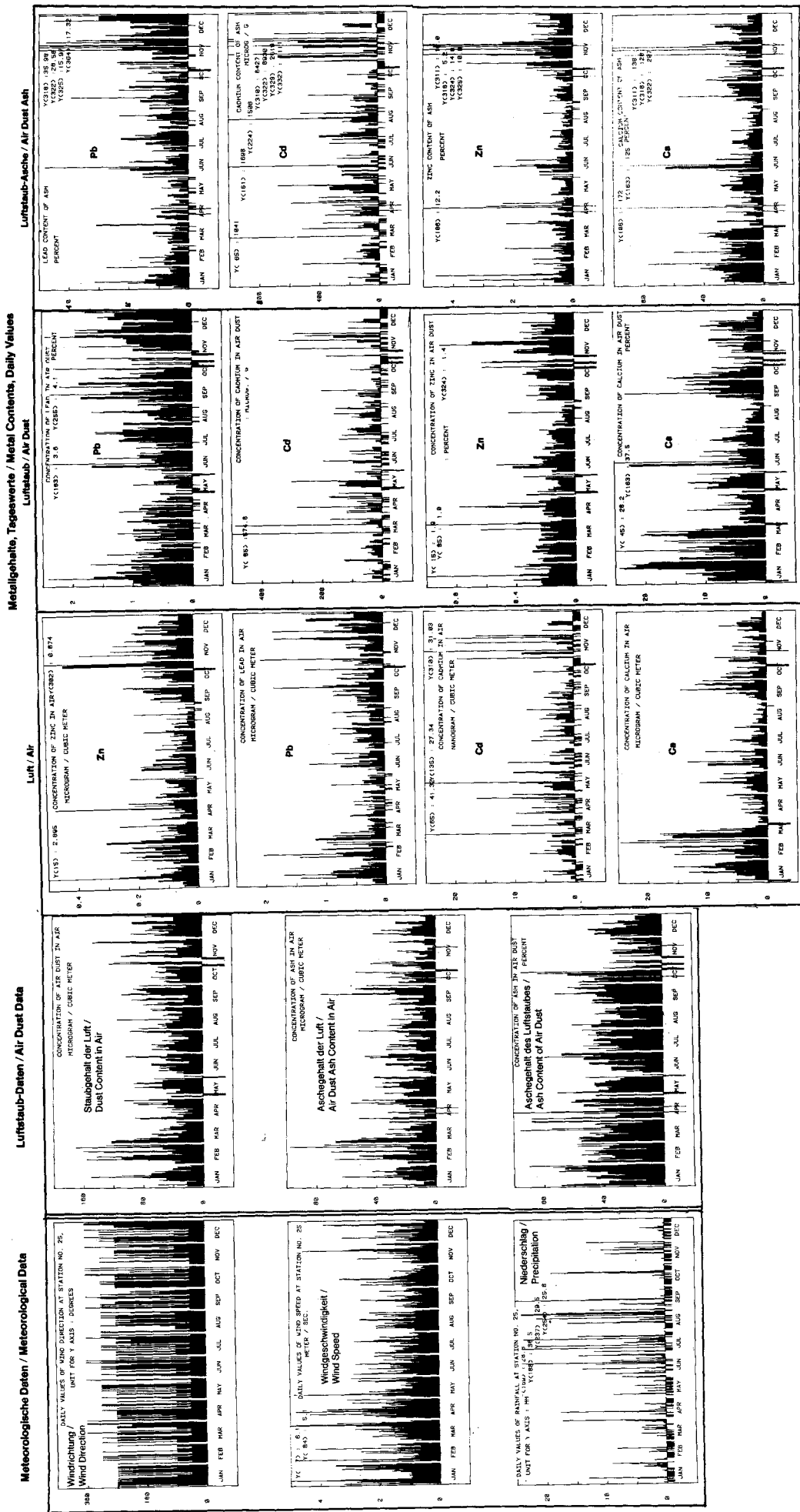


Fig. 1. Graphic representation of the data sample as a diagram. Daily air dust data from Table 2 [30f, 32, 37]. About 6100 different individual data are concentrated within this one single figure

Table 3. Data concentration. Location of arithmetic mean values. Annual arithmetic means \bar{x} of the individual daily air dust values northern border of Munich (GSF, Neuherberg) (from Table 2 and Fig. 1) [30f]

	1971	1972	1973	1974	1975	1976
Meteorological parameters:						
Dust concentration ($\mu\text{g}/\text{m}^3$)						
1. Northern border	62	65	64.251	51.571	52.758	53.963
2. City					75.811	84.809
Wind speed (m/s)						
1. Northern border			1.74	1.92	1.52	1.61
2. City					1.31	1.90
Precipitation (mm/d)						
1. Northern border			1.97	2.19	2.29	2.89
2. City					2.92	1.30
Concentrations (per m^3 air):						
Ash of air dust ($\mu\text{g}/\text{m}^3$)						
1. Northern border			29.332	24.332	21.097	19.407
2. City					32.644	39.316
Pb ($\mu\text{g}/\text{m}^3$)						
1. Northern border	(0.7) ^a	0.42	0.490	0.384	0.420	0.235
2. City					1.446	0.743
Cd (ng/m^3)						
1. Northern border			2.99	1.36	2.28	1.20
2. City					3.23	3.02
Zn ($\mu\text{g}/\text{m}^3$)						
1. Northern border			0.148	0.097	0.108	0.105
2. City					0.178	0.180
Ca ($\mu\text{g}/\text{m}^3$)						
1. Northern border	(4.8) ^b		3.357	2.993	3.669	3.414
2. City					5.968	7.315
Contents (per g air dust):						
Ash of air dust (%)						
1. Northern border		53.1	46.338	46.675	39.065	36.368
2. City						
Pb (%)						
1. Northern border	(1.12) ^a		0.776	0.804	0.826	0.449
2. City					2.143	0.980
Cd (ppm)						
1. Northern border			44.561	23.365	42.025	23.967
2. City					45.496	37.667
Zn (%)						
1. Northern border			0.244	0.205	0.204	0.212
2. City					0.240	0.227
Ca (%)						
1. Northern border			5.151	5.440	6.629	6.224
2. City					7.680	8.658

^a August-December 71^b August-November 71

e) outliers (extreme values) have to be eliminated in order to make a "heterogeneous" sample "homogeneous".

Steps a) to c) are independent of the type of distribution. The presumption, however, is that the

data belong to one homogeneous population. The calculation itself gives evidence, whether the sample is heterogeneous and consists of more than one data population, which have to be separated and treated separately. In the reverse case, a data sample can be

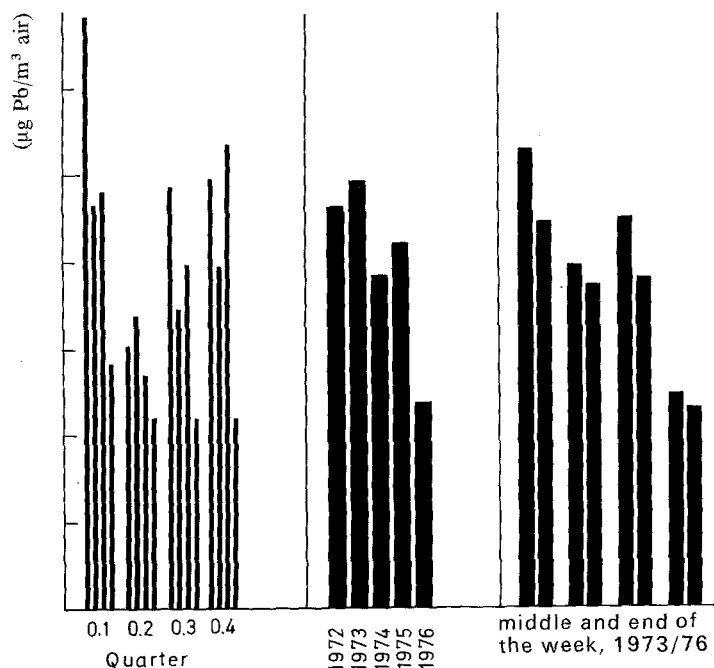


Fig. 2
Data concentration as block diagram of quarterly and annual arithmetic mean values. Air dust data according to Table 3 [30f, 37]

defined as “homogeneous” only due to statistical reasons. This does not imply that the sample can be considered to be “homogeneous”, also based on physical or chemical reasons, as the example given later on in Fig. 11 shows.

The minimum size of the data sample for different data concentration procedures according to a)–e) depends on the procedures which are used. For a) and b), in the case of approximately normal distributed data, “4” is a practical number in this laboratory, when repeating analytical determinations under the same conditions for estimation of the precision. It seems to be a realistic minimum for applying an outlier test, because after elimination of one outlier, the remaining 3 data are the minimum for calculating \bar{x} , s and V_r . A larger number of values would, of course, be much better due to statistical reasons. However, manpower, time and costs increase, whilst measuring more data. Two determinations are, of course, better than only one, however, if 4 are not possible, at least 3 repetitions should be performed. Steps c) and d) need a much larger number of data. For 10 to 20 intervals in the frequency distribution histogram, at least 30 to 50 individual values should be available; more are better.

Another major statistical problem is to infer from the random sample, which has been discussed until now, the characteristics of the whole universe of data, which are to be estimated. These procedures, however, are outside the scope of this paper [3, 10, 18, 20–22, 24].

2.4 Programme ZCH-3/1 (1978) for Data Concentration

For data concentration in routine analysis at the Central Department for Chemical Analysis at the Jülich Centre, the programmes ZCH-2

and ZCH-3/1 (1978) have been developed for the graphical table computer system Tektronix 4051 with 32 k [11, 12, 35–37]. It covers the following steps and calculations (Tables 7, 9–12).

1. *Data Sampling* (Table 4). The n values of the variable x_i form the data sample $x_1, x_2, x_3, \dots, x_n$. After arranging them in increasing order, they are assigned the symbol X_i and the rank number $N: X_1, X_2, X_3, \dots, X_N$, where $X_i \neq x_i$.

2. *Histogram*. The programme ZCH-2 provides the histogram of frequency distribution of the data and also the cumulative distribution.

2. *Type of Distribution*. It can be evaluated visually from the linearity of the probability plots for normal and lognormal distribution as Hazen's straight lines. Moreover, the Kolmogorov-Smirnov goodness-of-fit test gives a quantitative measure for the approximation to the normal and lognormal distribution. In addition the Cramer-von Mises test is used (Programme ZCH-2).

4. *Data Location*. The programme ZCH-3/1 (Table 4) starts again with rank ordering and draws the histogram in absolute as well as in relative frequencies. It continues with location of the centre of gravity of the sample, and calculates different mean values: arithmetic (\bar{x}), geometric (\bar{x}_G) and harmonic (\bar{x}_H) mean. As a distribution-free estimation, the median \tilde{x} is given.

The definition of the median \tilde{x} may be called to mind: It is the middle value of the data sample, when ranked in ascending order; this is equivalent to the 50th percentile or the numerical value of the data with the rank number $N/2$. Its great advantage in describing skew distributions and data samples with outliers is

Table 4. Definition of the data sample and different measures for location of the data [36]

(Programme ZCH-3/1)	
1. Data sample. <i>n</i> values of the variable <i>x</i> are $x_1, x_2, x_3, \dots, x_n$, arranged in increasing order with the rank number <i>N</i> : $X_1, X_2, X_3, \dots, X_N$, where $X_i \neq x_i$. No particular type of distribution is assumed.	
2. Data location	
2.1. Arithmetic mean	$\bar{x} = \frac{\sum x_i}{n}$
2.2. Mean of logarithms	$\bar{x}_{\log} = \frac{\sum \log x_i}{n}$
2.3. Geometric mean	$x_G = 10^{\bar{x}_{\log}}$
2.4. Harmonic mean	$\bar{x}_H = \frac{\sum 1/x_i}{n}$
2.5. Non-parametric estimation: median	
when <i>n</i> is odd	$\bar{x} = X_{(N+1)/2}$
when <i>n</i> is even	$\bar{x} = \frac{X_{N/2} + X_{(N/2)+1}}{2}$

that both ends of the distribution with outliers have no influence on the median.

The central value of a normal distribution is the arithmetic mean \bar{x} , that of a lognormal distribution the geometric mean \bar{x}_G .

5. *Dispersion.* Scattering around the central value according to Table 5 is characterized by the standard deviation *s*, the relative coefficient of variation V_r , the mean deviation of the arithmetic mean, and the geometric standard deviation. The range *R* and 80% interdecile range are distribution free; but use of range is not advisable for *N* > 10. Examples are given in Table 5a for the air dust data.

6. *Confidence Intervals* of the mean values are calculated for statistical probabilities of 95 and 99%. Table 6 considers the confidence intervals of the arithmetic mean, the median, the logarithmic mean and the geometric mean.

7. *Shape of Distribution.* As a measure for skewness, the skewness g_1 and for kurtosis, the kurtosis g_2 according to Fisher are calculated (Table 7).

8. *Outliers.* According to Table 8, the *t*-, Nalimov-, Grubbs- as well as Dixon tests are considered as outlier tests. It should be kept in mind, that they may have different power and eliminate a different number of outliers. After each outlier elimination, the programme ZCH-3/1 starts again with the original data sample to determine the outliers for each test.

Table 5. Dispersion around the central value [36]

(Programme ZCH-3/1)	
1. Standard deviation	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$
2. Relative coefficient of variation (Rel.-%)	$V_r = s \cdot 100/\bar{x}$
3. Mean deviation from the arithmetic mean	$DM = \frac{\sum (x_i - \bar{x}) }{n}$
4. Standard deviation of logarithms	$s_{\log} = \sqrt{\frac{\sum (\log x_i - \bar{x}_{\log})^2}{(n-1)}}$
5. Geometric standard deviation	= antilog s_{\log}
6. Non-parametric estimates:	
– Range	$R = X_N - X_1$
– 80% Inter decile range	$I_{80\%} = 9\text{th decile} - 1\text{st decile}$

3. Typical Problems of Data Concentration

3.1 Deviation from Normal Distribution

The frequency distribution histogram of the daily air dust concentration at Neuherberg, 1975 [30f, 37], according to Fig. 3, obviously does not have a normal distribution. A few outliers are still not registered and are outside the diagram. The graphical representation as Hazen's straight line in the probability plot fails for normal distribution but gives good agreement for a lognormal one (Fig. 4).

A quantitative estimation can be obtained by applying the Chi-square test or, as within this investigation, by the Kolmogorov-Smirnov as well as the Cramer-von Mises test. A test-statistics is calculated and compared with a tabulated value, which is a function of the number *N* of data. The results in Table 9 confirm the qualitative impression, which is given in Fig. 4 by the probability plots. The data are in excellent agreement with lognormal distribution, but do not fulfil the Gaussian normal distribution.

Skew or logarithmic distributions in environmental, biological, and medical sciences or geochemistry are more often observed than the Gaussian normal distribution. Examples are the distribution of the daily air dust concentration (Fig. 3) and several metal concentrations (Fig. 5) in the air at Neuherberg 1975 [37, 43]. The skew distributions observed frequently in the author's laboratories, within the different fields mentioned before, could often be approximated by the

Table 5a. Different measures for the dispersion of a data sample. Number of data (concentrations) N , standard deviation of single values s , standard deviation of the mean $s_{\bar{x}}$, relative variation coefficient V_r (rel.-%), 80% inter decile range $I_{80\%}$, confidence limits for 95 and 99% statistical probability ($\alpha = 0.05$ and 0.01) [36]. Air dust data from the northern border of Munich 1975 (Neuherberg, GSF), according to Table 2

	N	s ($s_{\bar{x}}$)	DM	V_r (%)	$I_{80\%}$	Confidence limits of \bar{x}	
						95%	99%
Dust for \bar{x} \bar{x}_G \tilde{x}	354	28.6 (1.52)	21.6	5.4	70.4	49.8 – 55.8 43.7 – 48.7 43.8 – 48.7	48.2 – 56.7 43.0 – 49.6 43.2 – 49.0
Ash for \bar{x} \bar{x}_G \tilde{x}	358	14.7 (0.78)	10.9	7.0	30.7	19.6 – 22.6 15.2 – 17.8 16.3 – 19.7	19.1 – 23.1 14.8 – 18.3 15.3 – 20.4
Lead for \bar{x} \bar{x}_G \tilde{x}	362	0.35	0.25	8.4	0.77	0.38 – 0.46 0.34 – 0.39 0.28 – 0.36	0.37 – 0.47 0.33 – 0.40 0.28 – 0.38
Zinc for \bar{x} \bar{x}_G \tilde{x}	362	0.17 (0.009)	0.066	15.5	0.16	0.09 – 0.13 0.07 – 0.09 0.07 – 0.09	0.09 – 0.13 0.07 – 0.09 0.07 – 0.09
Calcium for \bar{x} \bar{x}_G \tilde{x}	359	3.89 (0.21)	2.73	10.6	8.00	3.27 – 4.07 2.09 – 2.58 2.10 – 2.50	3.14 – 4.20 2.03 – 2.66 2.05 – 2.65
Cadmium for \bar{x} \bar{x}_G \tilde{x}	362	4.22 (0.22)	2.35	18.5	5.10	1.84 – 2.72 1.97 – 2.52 0.76 – 1.37	1.71 – 2.86 1.89 – 2.62 0.73 – 1.41

Table 6. Confidence intervals of central values [36]

(Programme ZCH-3/1)	
1. of arithmetic mean	$= \bar{x} \pm \frac{t \cdot s}{\sqrt{n}}$
2. of median	$x_{(h)} \leq \tilde{x} \leq x_{(n-h+1)}, \text{ where}$ $h = \frac{n - t \cdot \sqrt{n-1}}{2}, \text{ reliable when } n < 50$
3. of logarithmic mean	$= \bar{x}_{\log} \pm \frac{t \cdot s_{\log}}{\sqrt{n}}$
4. of geometric mean	$= 10 \left(\bar{x}_{\log} - \frac{t \cdot s_{\log}}{\sqrt{n}} \right) \text{ and}$ $10 \left(\bar{x}_{\log} + \frac{t \cdot s_{\log}}{\sqrt{n}} \right)$

Value t for $(n-1)$ degrees of freedom, corresponding to 95% or 99% probability

lognormal distribution and only in few cases by the Gaussian.

It should be mentioned, however, that the practical examples observed in nature in most cases do not

Table 7. Skewness and kurtosis [36]

(Programme ZCH-3/1)	
Skewness	$g_1 = \frac{\sum_i (x_i - \bar{x})^3}{n \cdot s^3}, \text{ where}$
	$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$
Kurtosis	$g_2 = \frac{\sum_i (x_i - \bar{x})^4}{n \cdot s^4}, \text{ where}$
	$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$

exactly correspond to normal or lognormal distributions, these are only approximations. Therefore, the programme ZCH-3/1 always calculates the statistics for both cases.

Figure 6 represents the most extreme skew distributions observed by one of the authors in the field of environmental analysis. It is the frequency distribution of toxic metal concentrations (in ppm) in 4% acetic acid after leaching ceramic ware for 24 h at room tempera-

Table 8. Tests for outliers [36]

(Programme ZCH-3/1)

1. *t*-test:

$$t = \frac{(x_n - \bar{x}')}{s' \cdot \sqrt{n/(n-1)}} \text{ or } \frac{(\bar{x}' - x_1)}{s' \cdot \sqrt{n/(n-1)}}$$

where x' and s' are calculated from $(n-1)$ observations, omitting the suspected observation. t is compared with the critical value for $(n-2)$ degrees of freedom.

2. Nalimov test:

$$r = \frac{(x_n - \bar{x})}{s \cdot \sqrt{(n-1)/n}} \text{ or } \frac{(\bar{x} - x_1)}{s \cdot \sqrt{(n-1)/n}}$$

for $n < 26$;
 r is compared with the critical value for $(n-2)$ degrees of freedom.

3. Grubbs test:

$$T = \frac{(x_n - \bar{x})}{s} \text{ or } \frac{(\bar{x} - x_1)}{s}$$

for $n \leq 30$;
 T is compared with the critical value corresponding to n

4. Dixon test:

$$r_{i,j} = \frac{(x_n - x_{n-1})}{(x_n - x_{1+j})} \text{ or } \frac{(x_{1+i} - x_1)}{(x_{n-j} - x_1)}$$

where $i = 1, 2; j = 0, 1, 2$. i and j differ for different values of n . $r_{i,j}$ is compared with the critical value corresponding to n .

ture. There are still more outliers far outside the distributions. These distributions are no longer even true lognormal distributions.

3.2 Comparison of Arithmetic Mean \bar{x} , Median \tilde{x} and Geometric Mean \bar{x}_G

Table 10 gives an impression about the differences that may occur between \bar{x} and \tilde{x} in a skew distribution. In these cases, the median is the more correct measure for the central value. The largest difference shows the cadmium content of 50% at the northern border of Munich and of even 70% in the centre. One should be aware that these are differences, not of single measurements, but in the annual means for about 250 measured data!

Table 11 compares \bar{x} , \bar{x}_G and \tilde{x} for some air dust data. A more detailed discussion shows that in the case of skew distributions, which can be approximated by the lognormal distribution, the geometric mean \bar{x}_G is quite similar to the median \tilde{x} , whereas \bar{x} is more different. This is the case for air dust, ash, lead, zinc and cadmium. Only for the cadmium concentration in air, \bar{x} is almost equal to \bar{x}_G and both are different from \tilde{x} .

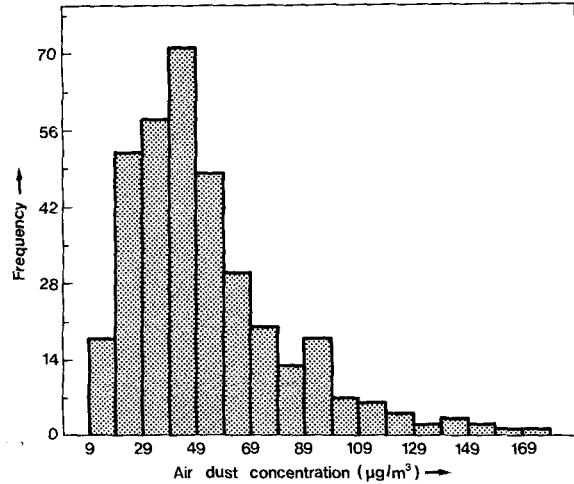


Fig. 3. Frequency distribution of daily air dust concentrations ($\mu\text{g}/\text{m}^3$). Experimental data according to Table 2 and Fig. 1

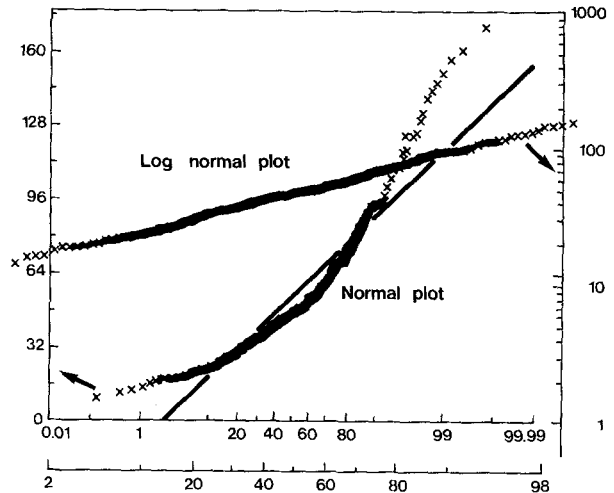


Fig. 4. Cumulative probability plot based on normal and lognormal distribution. Daily air dust concentrations ($\mu\text{g}/\text{m}^3$), data according to Table 2 and Fig. 3. Abscissa: Cumulative probability. Ordinate: Air dust concentration ($\mu\text{g}/\text{m}^3$)

In the case of Gaussian normal distribution, \tilde{x} becomes equal to \bar{x} . Therefore, it might be better to use the median \tilde{x} instead of \bar{x} for an unknown distribution.

3.3 Outlier Elimination

Outliers are defined as a limited number of data, which are so far separated in value from the remaining sample, that they give rise to the question of whether they are not from a different population, or that the sampling technique is at fault. Outlier tests ascertain, whether they can be accepted as homogeneous with the rest of the sample [17].

The more a value strays outside the sample, the more serious the error may become, when calculating the arithmetic mean \bar{x} including the outliers. Since it is

Table 9

Results of Kolmogorov-Smirnov goodness-of-fit test for normal and lognormal distribution of air dust concentrations (according to Fig. 3) Approximation within 95%.

	N	Test statistics for				Approximation within 95% ^a	
		normal distribution		lognormal distribution		normal distribution	lognormal
		sample a	table c	sample a	table c		
Dust ($\mu\text{g}/\text{m}^3$)	354	0.1336	0.0723	0.0329	0.0723	—	+
Ash ($\mu\text{g}/\text{m}^3$)	358	0.1053	0.0719	0.0105	0.0719	—	+
Pb ($\mu\text{g}/\text{m}^3$)	362	0.1579	0.0715			—	
Cd (ng/m^3)	362	0.2263	0.0715			—	
Zn ($\mu\text{g}/\text{m}^3$)	362	0.1696	0.0715			—	
Ca ($\mu\text{g}/\text{m}^3$)	359	0.1959	0.0718			—	

^a $a < c$: +; $a > c$: -. + indicates, that data approximate to the given distribution and vice versa

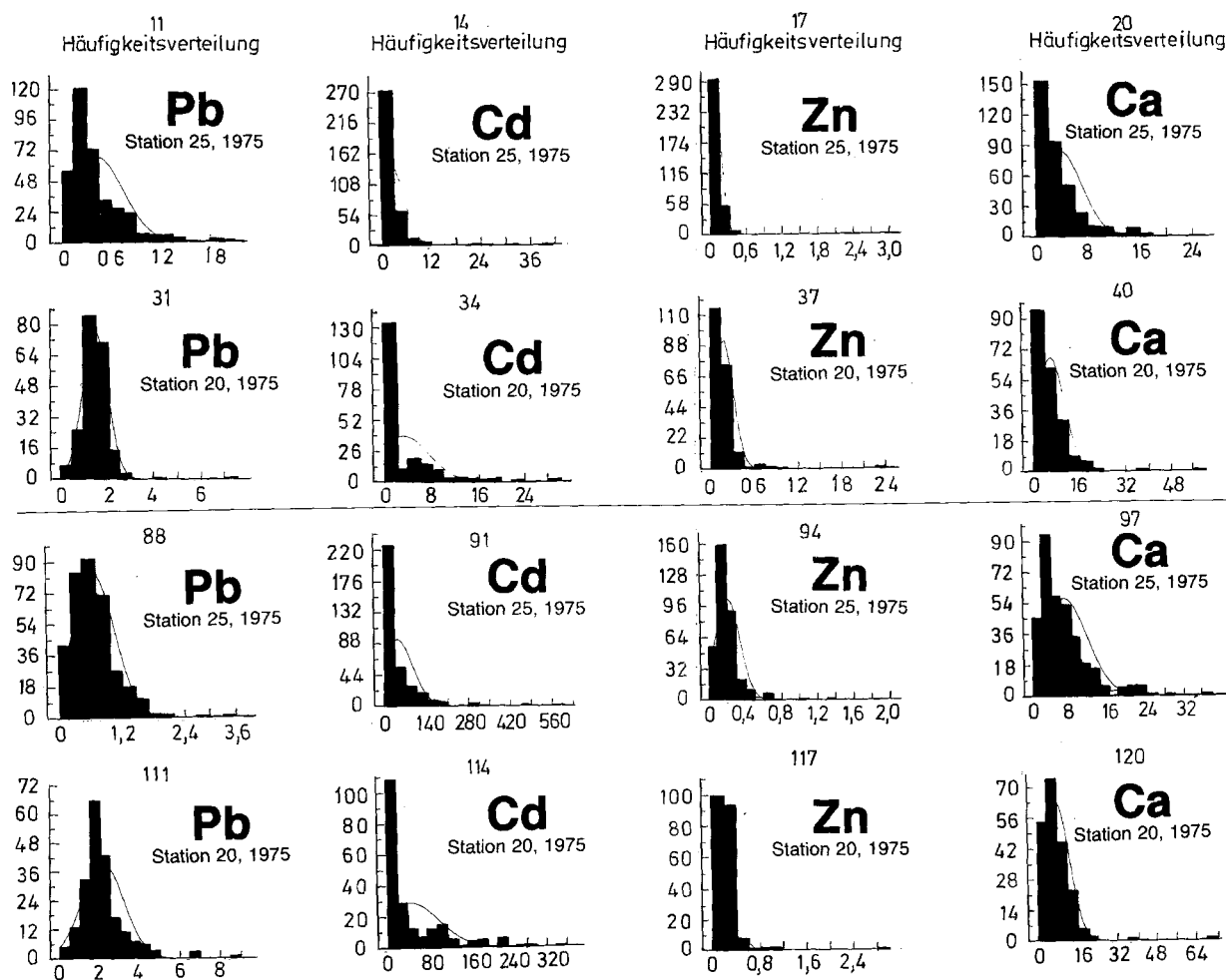


Fig. 5. More examples for skew and approximately normal distributions: Daily values for lead, cadmium, zinc, calcium at two different stations in Munich, 1975 (Station 25 at the northern border, GSF; Station 20: Centre of the city, Leuchtenberg-Unterführung). Data according to Table 2. Upper diagram: concentrations in air; lower diagram: contents in air dust

assumed per definition, that outliers do not belong to the same data sample, they have to be eliminated in order to make the remaining sample homogeneous.

This, however, does not imply that an outlier is always something, which is not important and has to be

neglected! The reverse can be true. In the case of toxic metal concentrations in biological and environmental systems, high outlier values are often more important than the remaining homogeneous data sample. These outliers signalize danger for man or the biosystem.

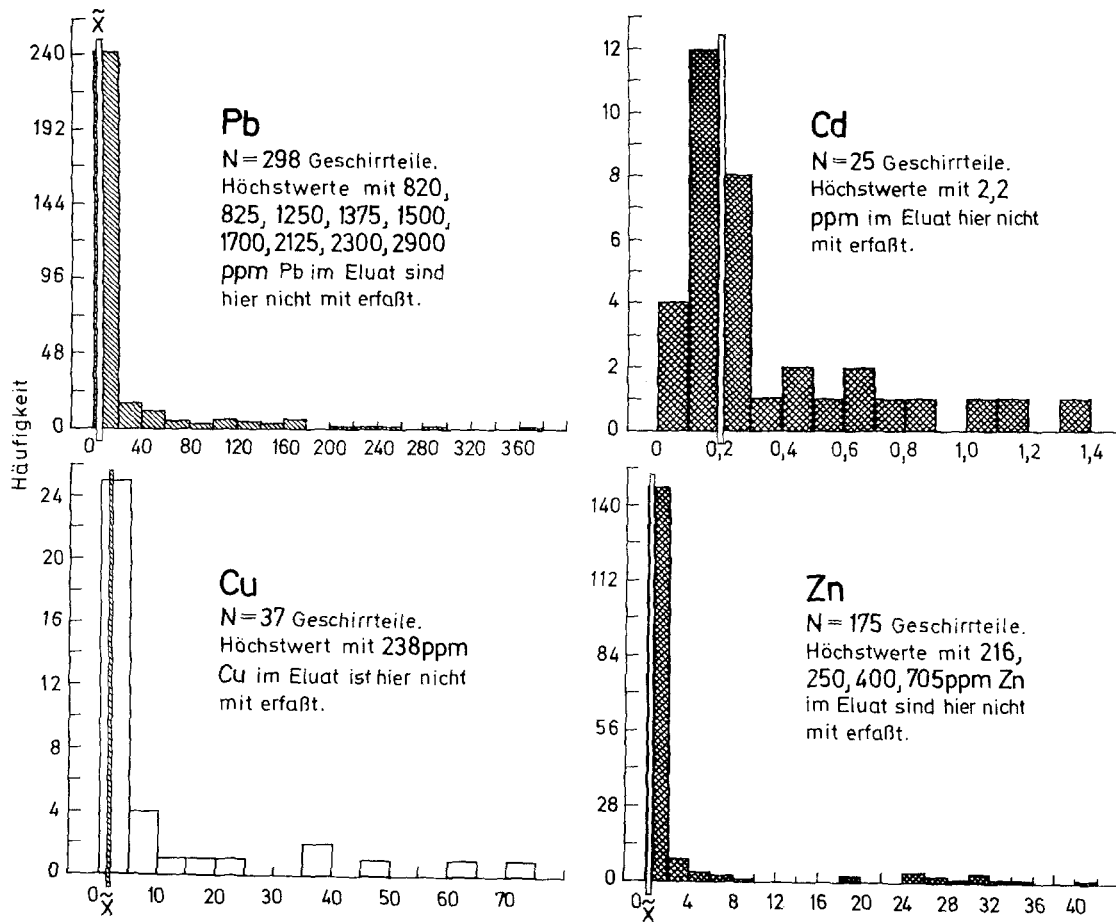


Fig.6. Examples for extremely skew distributions: Metal concentrations in the elutriate of ceramic ware. Leaching with 4% acetic acid for 24 h at room temperature under exclusion of light; metal concentrations determined by atomic absorption spectrometry [39], Munich 1972/75

Table 10. Difference between median \tilde{x} and arithmetic mean \bar{x} . Daily air dust data from the northern border of Munich (GSF, Neuherberg), 1975, according to Fig. 1 and Table 2 [28, 29, 32]

Component	N	Annual central value		
		Arithmetic mean \bar{x}	Median \tilde{x}	Difference* Rel.-%
Northern border of Munich (Neuherberg, GSF) 1974:				
Air dust	358	51.6	47.8	- 7.4
Ash of air dust	352	24.3	20.3	-16.5
Pb	355	0.38	0.33	-13.2
Cd	335	0.0014	0.0007	-50.0
Zn	354	0.097	0.090	- 7.2
Ca	349	2.99	1.80	-39.8
City of Munich (Leuchtenberg-Unterführung), 1975:				
Air dust	211	75.8	75.1	- 0.9
Ash of air dust	207	32.6	29.1	-10.7
Pb	208	1.44	1.43	- 0.7
Cd	208	0.0032	0.001	-70.3
Zn	208	0.18	0.14	-22.2
Ca	205	5.97	4.43	-25.8

* $(\tilde{x} - \bar{x})/\bar{x}$

Table 11. Comparison of arithmetic mean \bar{x} , geometric mean \bar{x}_G and median \tilde{x} for skew distributions (30e). Air dust data according to Fig. 1 and Table 2

	N	\bar{x}	\bar{x}_G	\tilde{x}
Dust	354	52.8	46.1	46.1
Ash	352	21.1	16.5	18.0
Pb	362	0.42	0.36	0.31
Cd	362	2.28	2.23	0.95
Zn	362	0.11	0.08	0.08
Ca	359	3.67	2.32	2.18

Outlier-patients need further consideration and help, not the normal population [30d].

Figure 7 gives the annual arithmetic mean \bar{x} of the dust concentration in air at the northern border of Munich, 1975, as a function of increasing number of outlier eliminations. For the total of 354 daily values, the arithmetic mean \bar{x} is about $52.8 \mu\text{g dust}/\text{m}^3$. The outlier-t-test eliminates one outlier after the other on the high value side and stops at an $\alpha = 0.05$ level after elimination of 28 outliers. The corrected arithmetic

mean \bar{x}' for the remaining $N' = 332$ values becomes 47.1 instead of $52.8 \mu\text{g}/\text{m}^3$. It might be remarked that the corresponding median \tilde{x} gives a value of 46.7 without laborious outlier elimination, which is very close to $\bar{x}' = 47.1$ and quite different from $\bar{x} = 52.8$. The similarity of the outlier-corrected \bar{x}' with \tilde{x} and the difference between \bar{x} and \tilde{x} is due to the skewness of the distribution [30b, 30e] introduced by the outliers.

The second example shows, that an outlier elimination, which theoretically is based on a normal distribution, may become absurd, when applied to skew distributed data. In 1974, one of the authors (S.) has applied Nalimov's r -test [22] in the version of [7, 16] to the data sample of 110 lead contents in dental calculus from men and women, collected from normal patients in the Dental Clinical Hospital of the University of Munich (Director: Prof. Dr. E. Sonnabend) [30b, 30d, 30e]. The data distribution is given in Fig. 8, omitting the maximum content of $465 \mu\text{g Pb/g}$. On an $\alpha = 0.01$ level, the r -test eliminates 36 outliers in groups or as single values. The remaining sample of 76 data as well as the eliminated groups of outliers are quite normal distributed and "homogeneous". By this procedure, however, the number of data dropped from 110 to only 76, the arithmetic mean from $\bar{x} = 8.81$ to only $\bar{x}' = 1.46$ and the corresponding standard deviation from $s = \pm 44.98$ to only $s' = \pm 0.86 \mu\text{g Pb/g}$! Within a philosophy of the "Quasi-normal distribution", to be mentioned later, the correct way would be to present the corrected $\bar{x}' = 1.46$ together with $s' = \pm 0.86 \mu\text{g Pb/g}$ and to list the 34 outliers separately. It becomes quite obvious, that this way of thinking in terms of normality is nonsense. — Recently, the version [7, 16] of the r -test has been criticized [46]. Programme ZCH-3/1, however, according to Table 8 uses the original version of Nalimov's test [22], which is restricted to about $N < 26$ values. Due to this reason, Table 12 gives a similar treatment for the 104 data, using the t -test for eliminating the outliers, one after the other, and omitting the 6 values below the detection limit. After outlier elimination by t -test, the number of values has dropped from $N = 110$ to $N' = 84$, the arithmetic mean from $\bar{x} = 9.36$ to $\bar{x}' = 2.03$ and the corresponding standard deviation from $s = 46.21$ to $s' = 1.16 \mu\text{g Pb/g}$. Comparison of these results between Nalimov- and t -test demonstrates, however, that the use of different outlier tests gives different results; but the things are not so fundamentally changed when using one or the other test. This is also shown in the next example.

A third example informs us that different outlier tests, because of their different mathematical structure, may eliminate a somewhat different number of outliers and, therefore, may lead also to different corrected mean values \bar{x}' . Within the second example of the lead content in dental calculus, according to Fig. 10, the

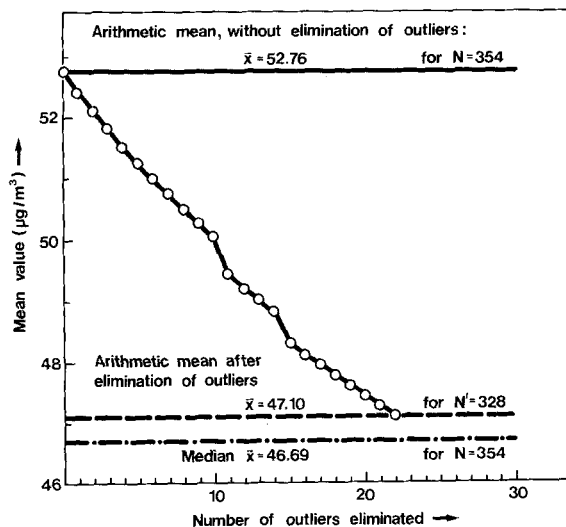


Fig. 7. Influence of increasing number of outlier eliminations on the arithmetic mean of a non-normal and skew data distribution. Daily air dust concentrations, 1975, northern border of Munich (GSF, Neuherberg). Outlier elimination according to t -test on a $\alpha = 0.05$ level. Data from Table 2 and Fig. 3

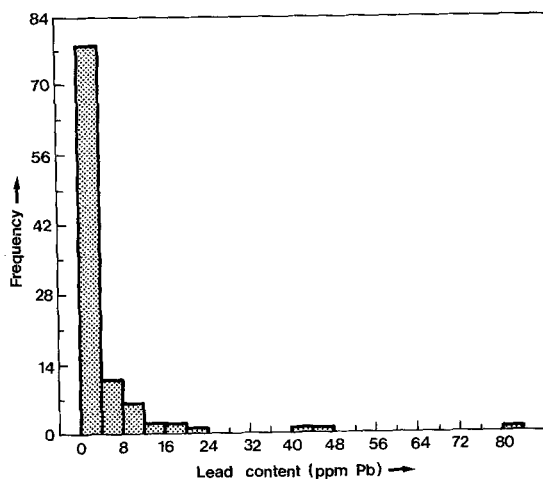


Fig. 8. Frequency distribution of lead content in dental calculus, normal patients (men and women), Munich 1973/74

Nalimov- r -test was more powerful than the t -test. The r -test eliminates 36 instead of only 22 outliers. This corresponds to an $\bar{x}' = 1.5$ instead of 2.0, as compared with a median of $\tilde{x} = 2.0 \mu\text{g Pb/g}$. Due to these reasons, the programme ZCH-3/1 calculates simultaneously four different outliers tests: the t -, Nalimov-, Grubbs- and Dixon-test. Some of them are restricted to a definite range of N . To which test preference should be given, must be decided for each individual case, when discussing the shape and other descriptive statistics of the frequency distribution.

Table 12. Absurdity of outlier elimination from skewly distributed data samples by an outlier test for normal distributed data (*t*-test). Elimination of outliers by *t*-test (on a $\alpha = 0.01$ level) for lead contents in 110 dental calculus samples from a patient collective (men and women) of the Dental Hospital of the Univ. of Munich 1973/74. Sampling by E. Sonnabend, J. Herzog; lead determination by flameless atomic absorption by B. Sansoni, W. Kracke, F. Dietl. From the beginning, 6 values below detection limit have been neglected, the outlier-*t*-test starts with $N = 104$ values and ends with $N' = 84$ values. This remaining sample should be quasi-normal distributed and homogeneous

Number of outlier-eliminations	Number of patients, N		Outlier eliminated ($\mu\text{g Pb/g}$) x_{max}	Lead content in dental calculus ($\mu\text{g Pb/g}$)			
				Arithmetic mean \bar{x}		Standard deviation s	
	before	after		before	after	before	after
	outlier-elimination by <i>t</i> -test			outlier-elimination by <i>t</i> -test		outlier-elimination by <i>t</i> -test	
	N	N'	\bar{x}	\bar{x}'	s	s'	
1.	104	103	465.0	9.36	4.93	46.21	10.07
2.	103	102	81.0	4.93	4.19	10.07	6.67
3.	102	101	45.0	4.19	3.78	6.67	5.30
4.	101	100	40.0	3.78	3.42	5.30	3.88
5.	100	99	20.0	3.42	3.25	3.88	3.50
6.	99	98	18.0	3.25	3.09	3.50	3.16
7.	98	97	17.4	3.09	2.94	3.16	2.82
8.	97	96	15.2	2.94	2.82	2.82	2.53
9.	96	95	14.4	2.82	2.70	2.53	2.25
10.	95	94	11.4	2.70	2.60	2.25	2.07
11.	94	93	9.40	2.60	2.53	2.07	1.95
12.	93	92	9.30	2.53	2.46	1.95	1.83
13.	92	91	8.90	2.46	2.39	1.83	1.71
14.	91	89	8.00	2.39	2.26	1.71	1.50
15.	89	88	7.20	2.26	2.21	1.50	1.41
16.	88	87	6.60	2.21	2.16	1.41	1.34
17.	87	86	6.50	2.16	2.10	1.34	1.26
18.	86	85	5.50	2.10	2.06	1.26	1.21
19.	85	84	5.40	2.06	2.03	1.21	1.16
20.	84	—	—	2.03	—	1.16	—

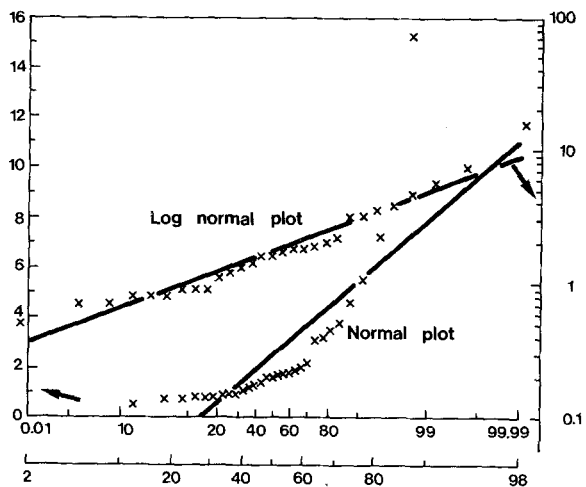


Fig. 9. Normal and lognormal plot of lead content in dental calculus from 110 patients of the dental clinical hospital, Univ. of Munich, 1973/74. Data according to Fig. 8. Abscissa: Cumulative probability. Ordinate: Lead content (ppm)

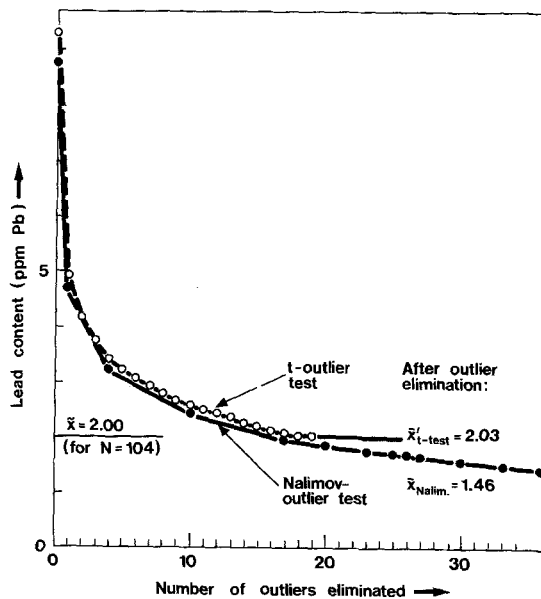


Fig. 10. Comparison of outlier elimination by *t*- and Nalimov's *r*-test. Lead content according to Fig. 8. In this case, values below detection limit have been eliminated, therefore actual $N = 104$. Nalimov's *r*-test in the version of [7, 16]

Table 13. Influence of excluding or including concentrations below the detection limit on the arithmetic mean \bar{x} . Excl.: excluding the values below the detection limit DL; incl. O: including as zero value; incl. DL: including as detection limit value. N : Number of data, \bar{x} : arithmetic mean, DL: Detection limit. Original data samples from the examples mentioned earlier

System	Data sample, N			Arithmetic mean \bar{x} (ppm)		
	$N_{\text{excl.}}$	$N_{\text{incl. O}}$	$N_{\text{incl. DL}}$	$\bar{x}_{\text{excl.}}$	$\bar{x}_{\text{incl. O}}$	$\bar{x}_{\text{incl. DL}}$
Lead content in dental calculus, Munich 1973/74, men + women	104	110	110	9.36	8.84	8.86
Lead leaching from ceramic ware into 4% acetic acid	385	831	831	65.1	30.14	30.3
Cadmium leaching from ceramic ware into 4% acetic acid	43	831	831	1.57	0.081	0.13

3.4 Concentrations below Detection Limit

Another sometimes serious and not always satisfactorily solved problem is the handling of concentrations or contents *below* the detection limit. Several different possibilities exist:

- to exclude from the data sample and list separately,
- to include as zero values
- to include as detection limit values
- to include as determination limit values
- to include as 50% of the values in c) or d).

Here only the influence of a) and b) on the arithmetic mean \bar{x} shall be discussed.

It is obvious, that the difference in both cases a) and b) – excluding or including the detection limit – becomes larger with increasing number of values below the detection limit. Secondly, the frequency distribution of element concentrations becomes skew, if the concentration range determined is still measurable, but near the detection limits.

Table 13 adds to the two data samples already mentioned before as a third example the lead and cadmium concentrations from leaching of ceramic ware with 4% acetic acid for 24 h at room temperature [39]. This is an official screening test for leaching of toxic elements from ceramic ware.

In the extreme case of lead leaching from ceramic ware, among 831 ceramics pieces investigated in the period from 1970–1978 at GSF Neuherberg, 385 were found to be below the detection limit of 0.04 ppm lead in the elutriate. If they are excluded, the arithmetic mean becomes 65.1 ppm instead of only 30.3 ppm when including them as zero (or detection limit). This corresponds to a difference of almost 100%! Again one should be aware that this is not an error for a single value, but for the mean of a large data sample containing as many as 831 measurements! The analytical error of a single lead determination by flame atomic

absorption spectrometry might be around ± 2 to $\pm 3\%$ only [39].

In environmental protection, the health of man in his environment has to be controlled. Therefore it is quite obvious, that values below the detection limit have to be included as determination limit values, in order to be on the safe side, when comparing the monitored contamination by a toxic substance with its tolerance level. On the other hand, it seems to be sufficient to use the somewhat smaller detection limit instead of the determination limit.

3.5 Homogeneous Sample [30e, f]

One of the fundamental presumptions in applied statistics is that the data sample to be treated is “homogeneous”. That means, it is normal distributed, free of outliers, continuous and has only one mode in its frequency distribution.

There seems to be no doubt that the frequency distribution shown in Fig. 11, after elimination of the three outliers between 34 and 40, is a good approximation to a Gaussian normal distribution. This impression is confirmed by the probability plots in Fig. 12.

These values represent 104 gamma-dose rate measurements of the natural gamma-radioactivity in Wunsiedel. This is a town with about ten thousand inhabitants in the Fichtelgebirge, a mountain area at the northern border of Bavaria, at an altitude of 500 m up to 1050 m above sea level [30e]. This region belongs to two or three of the highest natural radioactivity background areas in the Federal Republic of Germany. Sampling of the data was done just by walking around in the centre of the town and measuring the dose rate of the natural background gamma-radioactivity for points of personal interest, e.g. walls of houses, fountains, road paving, bitumen, marble rocks, sand, soil, etc. The dose rate meter MAB 604 from Münchner Apparatebau with the sensitive scale ranges of 0–30 and 0–100 $\mu\text{R/h}$ has been used. From the measured data, the frequency distribution in Fig. 11 has been drawn. From

this it is obvious that sampling was not systematical, it was done more or less at random. Therefore, the distribution in Fig. 11 does not represent the real natural radioactivity level of the whole town in the right proportions. However, there is reason to assume, that all levels of dose rates, which might occur outside the houses, are included.

Probability plots in Fig. 12, as well as more quantitatively the Kolmogorov-Smirnov test, establish that within a significance level of $\alpha = 0.05$, this distribution is a normal Gaussian, considering the three values above $30 \mu\text{R/h}$ as outliers. The sample is statistically homogeneous. It is important to draw attention to the fact that this does not imply that the measurements are homogeneous, based also on physical and chemical considerations. In this special case, each measurement can be related to a well-defined material. A more careful investigation shows that all dose rates above about 16 to $18 \mu\text{R/h}$ – this is just the region of \bar{x} - and \tilde{x} -correspond to granitic material. The granites of Fichtelgebirge are all slightly radioactive due to radionuclides of the natural uranium-238 and the thorium-232 decay series, which are produced by natural contents of uranium and thorium. Different types of granites from Fichtelgebirge are showing slightly different radioactivity levels. There are some regions with quarries, which contain uranium minerals. Granite stones from such quarries have been used for building houses and road pavings. The radioactivity of granites from Fichtelgebirge is so characteristic, that dose rate measurements can be used for monitoring granite in this area. Even under a layer of asphalt, an old granite paving has been detected with the gamma dose rate meter.

Due to these reasons, it may be doubtful, to define the data sample in Fig. 11 as “homogeneous”. The question arises of separating the non-granitic from the granitic material and of treating them separately despite the fact that statistics declares the material as homogeneous. This is a limitation, which should be kept in mind.

Even when sampling was not systematic, the results from Wunsiedel can be compared with measurements in an other similar town, provided that the measurements have been made in the same way, paying attention to as constant a sampling technique as possible.

3.6 Heterogeneous and Multimodal Data Samples

The town of Marktredwitz is situated 8 km southeast of Wunsiedel. It has about sixteen thousand inhabitants and a similar structure, but a somewhat smaller fraction of houses and other buildings are constructed out of granitic material.

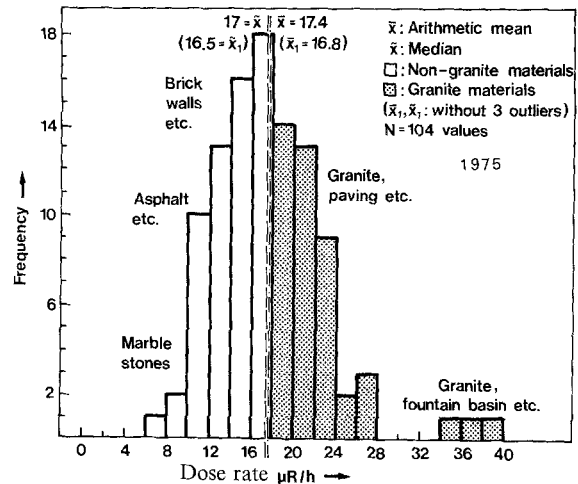


Fig. 11. Frequency distribution of gamma radiation dose rate measurements in the town of Wunsiedel/Fichtelgebirge, FRG [30e]

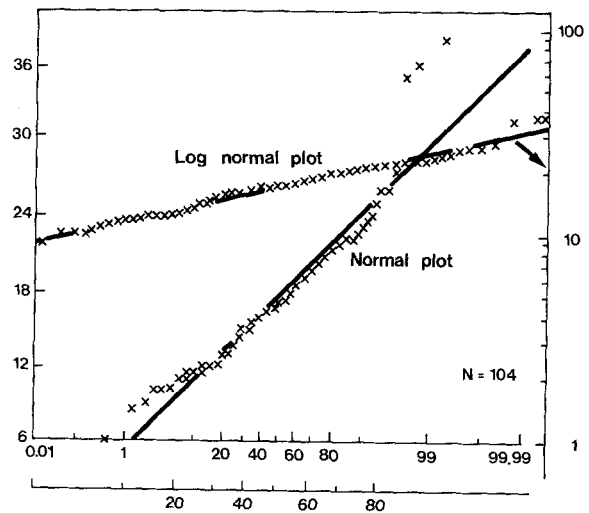


Fig. 12. Normal and lognormal plot of dose rate measurements at Wunsiedel/Fichtelgebirge. Data according to Fig. 11. Abscissa: Cumulative probability. Ordinate: Dose rate ($\mu\text{R/h}$)

The normal and lognormal probability plots for similar dose rate measurements in Figs. 13 and 14 are not straight lines. They are S-shaped and quite different from those of Wunsiedel.

The explanation is given by the frequency distribution shown in Fig. 15. The histogram is trimodal and the data sample is obviously a composite sample derived from at least two different populations.

If the whole distribution is treated as one data sample, the arithmetic mean \bar{x} becomes $30.1 \mu\text{R/h}$. This is, curiously, a value which no one measurement or material has shown in Marktredwitz! This value is also quite unreasonable because the usual granitic material, as to be found in Wunsiedel, does not reach such a high value.

From this it becomes quite clear, that the three different data groups have to be treated separately (Fig. 16). Let us concern ourselves only with the two

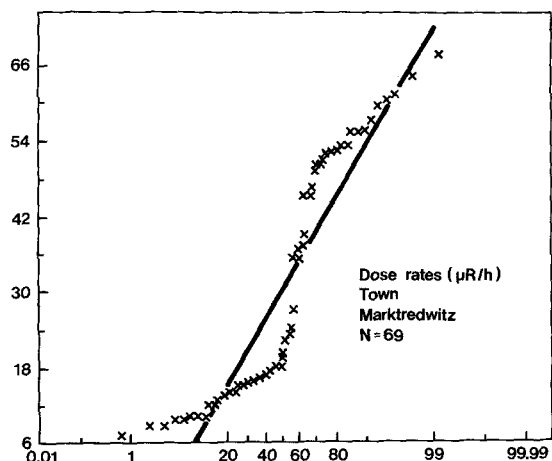


Fig. 13. Normal plot of gamma radiation dose rate measurements in the town of Marktredwitz/Fichtelgebirge [30e]. Abscissa: Cumulative probability. Ordinate: Dose rate ($\mu\text{R/h}$)

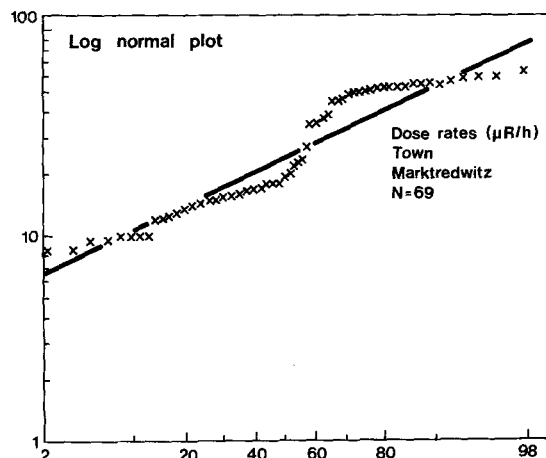


Fig. 14. Lognormal plot of gamma radiation dose rate measurements in the town of Marktredwitz/Fichtelgebirge. Abscissa: Cumulative probability. Ordinate: Dose rate ($\mu\text{R/h}$)

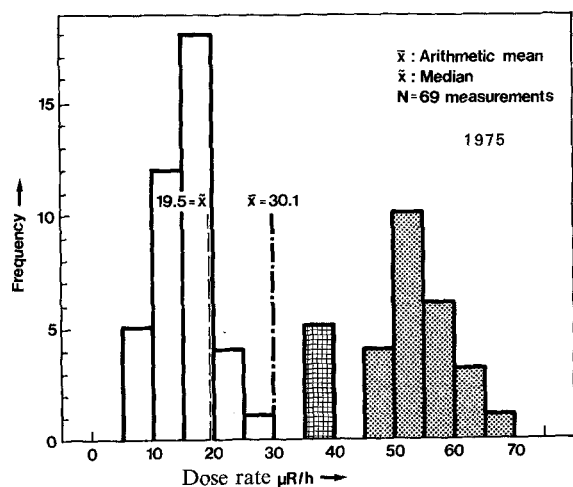


Fig. 15. Frequency distribution of a heterogeneous data sample with three different data populations [30e]. 69 gamma dose rate measurements in the town of Marktredwitz/Fichtelgebirge. Experimental conditions as in Fig. 11

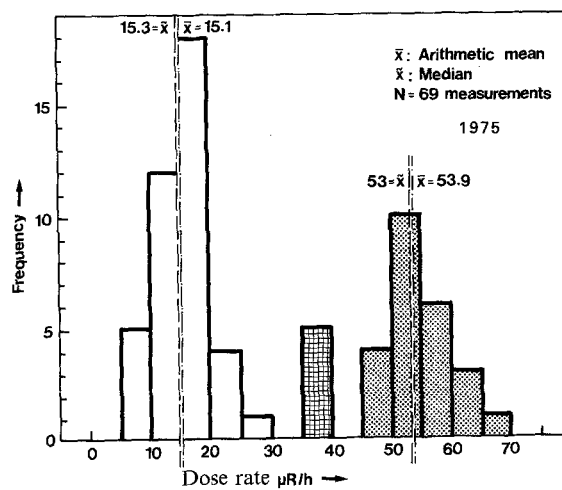


Fig. 16. Heterogeneous data sample treated as two separated homogeneous and approximately normal distributed samples [30e]. Gamma dose rate measurements in the town of Marktredwitz. Samples from 5 to 30 and 45 to 70 $\mu\text{R/h}$ are treated separately as almost normal distributed samples. Experimental conditions and original data as in Fig. 15

larger ones. When calculating separately, according to Figs. 17 and 18, both are quite normally distributed. The normal character is confirmed by the good agreement of arithmetic mean/median in both cases, which are 15.3/15.1 and 53.0/53.9 $\mu\text{R/h}$, respectively.

The sample with the lower values can be compared very well with the distribution at Wunsiedel (Fig. 11). This is due to the almost similar materials in both towns. The measured material in Marktredwitz has perhaps a little less granite ($\bar{x} = 15.1 \mu\text{R/h}$) as compared with Wunsiedel ($\bar{x} = 17.4 \mu\text{R/h}$). This confirms the qualitative impression mentioned above.

Completely different, however, is the third sample between 45 and 70 $\mu\text{R/h}$. It has no equivalent in Wunsiedel. Not a single value of this magnitude has

been measured within this investigation in Wunsiedel and the whole Fichtelgebirge area. It could be identified that the high dose rates at Marktredwitz, between 45 and 70 $\mu\text{R/h}$, correspond to a road paving made from slag stones in the main street, the market place and partly on a few other roads. These slag stones, which are very hard and durable, have been imported from Thuringia at the turn of the century and are almost unchanged in the road paving since that time. Neutron activation analysis has found about 90 to 100 ppm of uranium, compared with only 2 to 4 in normal soil material (by courtesy of Mr. B. Kayßer, ZCH, Jülich).

It was an interesting discovery that the highest natural radioactivity found in the Fichtelgebirge within

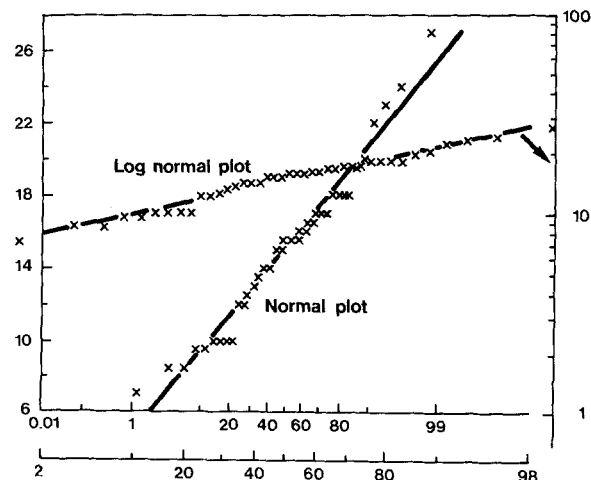


Fig. 17. Normal and lognormal plot of the dose rate, sample 5 to 30 $\mu\text{R/h}$ at Marktredwitz. Abscissa: Cumulative probability. Ordinate: Dose rate ($\mu\text{R/h}$)

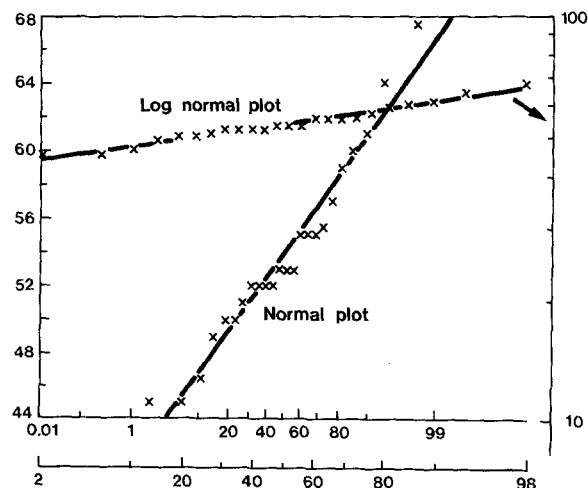


Fig. 18. Normal and lognormal plot of the dose rate; data sample 45 to 70 $\mu\text{R/h}$ at Marktredwitz. Abscissa: Cumulative probability. Ordinate: Dose rate ($\mu\text{R/h}$)

this study, was not in a granite quarry with uranium minerals, but in the main street of the largest town in the centre of the Fichtelgebirge! It should be mentioned, however, that even these relatively high values should not be harmful to the inhabitants. Moreover, they are now removed to a large extent during the new construction of the centre of the town.

This example has shown that the different data populations observed by statistical methods also correspond to quite different chemical and physical materials. Here is no doubt about heterogeneity.

4. Conclusions

A part of the scope of this paper was to demonstrate in an elementary way, that the trace analyst should take care regarding correct processing and interpretation of his analytical results not only with respect to precision and accuracy.

It has been assumed that the analytical error of the individual analytical data is negligibly small compared with its natural variations as a first approximation.

Attention has been drawn to the facts that:

- use of the arithmetic mean \bar{x} , standard deviation s and relative coefficient of variation V_r is restricted to normal distributions, from which outliers have been eliminated,

- investigation of the type of distribution of the data material is the primary and most important step before starting with other procedures for appropriate data concentration,

- different series of measurements can only be compared, if all conditions and presumptions of statistical data concentration including the same type of distribution have been observed.

- Furthermore, a computer programme ZCH-3/1 for the concentration of data sample, independent of the type of their distribution, has been presented. It considers all the requirements treated within this lecture and is in routine use in this laboratory since two years.

It could be demonstrated that incorrect or insufficient data concentration, which overlooks the skewness of the distribution and uses only arithmetic means for data location, may easily produce errors of the order of 10, 20, 50 or even 100%, compared with an analytical error of a single value of only 2–5%! The trace element analyst, unlike his colleague from biology, medicine, agriculture, anthropology, geochemistry, behaviour or social science often enters statistical data treatment via the Normal Law of Errors and the corresponding Gaussian Normal Distribution, when estimating precision and accuracy of his analytical results and the analytical method. This has led to an overestimation of the normal distribution, underestimating other types of distributions. An exception is only the Poisson Distribution in the case of radioactivity measurements. Due to this reason, trace element analysis today still sometimes finds itself in a position, which was held by professional statisticians about four decades ago, before the ideas of distribution-free and non-parametric statistics had developed. For this situation, the wrong treatment in Tables 3 and 12 as well as in Fig. 2 is characteristic.

Bradley [2] has described this situation as follows: The distribution, which is today called "normal", has been discovered by De Moivre in 1733, about half a century later by Laplace and again by Gauss. De Moivre has developed it as the limiting form of the binomial distribution. The last two mathematicians have tended to describe the distribution of errors in astronomical observations. Basic fundamentals are the Central Limit Theorem on one side and the Normal Law of Error on the other. These are quite different aspects compared with e. g. the distribution of trace elements in biological, environmental and geochemical systems. According to the Normal Law of Error, the error of astronomical (or instrumental-analytical) observations should be as likely positive as negative; should become less and less frequent as they increase in absolute magnitude; the empirical error distribution should be unimodal and symmetric about the true value; on both sides it seems to decrease monotonically; the slope of the error curve is zero and the curve horizontal at the centre, the error curve is zero when the relative frequency of the error is zero at either end; the error function is continuous and infinitely distributed. Therefore, the well-known bellshaped, symmetrical normal distribution curve is the result.

The normal error distribution began to fit and to explain almost everything. That which was Gaussian was considered normal and that which was non-Gaussian was regarded as abnormal. One could not become reconciled with the thought of the possible presence of "skew" frequency distributions, although numerous data offered completed defiance of the Gaussian dogma and exhibited markedly skew frequency distributions [6], as for example the data in Table 12. In the period of the "Mystery of Normal Distribution", the arithmetic mean appears to have been regarded as the truest value of all. Interest centered upon identification of the arithmetic mean, variance and the standard deviation, normality being taken for granted [2].

However, as larger quantities of data became available, the non-normality of more and more empirical distributions became apparent. But the central 80–90% of the distribution fitted normality often quite well. Therefore, the idea of the "quasi-universal normality" became replaced by a belief in "quasi-universal quasi-normality and quasi-homogeneity" [2].

After some time, it became evident that the values outside this 80% of the empirical quasi-normal distribution fit only increasingly poorly at increasingly remote tail regions. One became aware that in some areas, the non-normal and skew distributions are the "normal" case. Statistical evaluation of non-normal distributed data samples was much promoted by distribution-free statistics and order statistics since around 1936 [2, 3, 10]. With this in mind, Ahrens, for example, has considered the lognormal frequency distribution of chemical elements in minerals and rocks as the "Fundamental Law of Geochemistry"¹. In a similar way, the trace element distribution in biological, medical and environmental materials studied in the last twelve years in one of the author's laboratories, as mentioned before, fails in most cases to fulfil the Mystery of Normality and follows a quasi-lognormal and skew distribution. In these fields, values outside the quasi-normal distribution, which are responsible for the skewness, have been frequently found to be the most

important ones. Outliers are tested for not discarding; they merit the highest interest, whereas the mass of the remaining quasi-normal distributed values often have less importance.

As a last conclusion, it is mentioned that the trace element analyst shall, as before, be concerned with Gaussian normal distributions in the case of error estimation of analytical measurements. But when evaluating the meaning of his trace element concentrations for a given problem – taking the analytical error of every single data as negligibly small – he must be aware of the problem and treatment of skew and especially logarithmic distributions. The median and the geometric mean instead of the arithmetic mean, and confidence limits or interdecile ranges instead of variance and standard deviation are useful tools. Distribution-free statistics and order statistics give valuable help. On the other hand, he should not fall from the Mystery of the Normal Gaussian into the other extreme of a Mystery of the Lognormal Distribution.

Acknowledgement. One of the authors (S.) likes to thank Prof. Dr. G. Gottschalk, Gmund am Tegernsee, for introduction into statistical data evaluation from 1963–1975; participants of the GDCh-Training Courses on Information Processing in Analytical Chemistry, organized from 1970–1976 at the authors laboratory at Neuherberg, for valuable discussions; his former coworkers at GSF, Prof. Dr. J. Fischer†, Ing. (grad.) H. Ringer, cand. chem. M. Haimerl, W. Schulz, Ing. W. Kracke, Ing. (grad.) F. Dietl; his coworker at KFA Jülich Ing. (grad.) W. Matthes. Mrs. R. Kurth had made the calculations with the Tektronix 4051, Mr. H. Ringer had performed programming and calculating the raw data from air dust monitoring during 1970–1975. Dr. W. Schmid, Unternehmensberatung, München, has critically reviewed the program ZCH-3/1.

The other author (I.) thanks his authorities of Bhabha Atomic Research Centre at Bombay and Nuclear Research Establishment of Jülich for a Fellowship within the Indo-German Scientific Agreement in 1977/78 at Jülich.

References

1. Bevington, R. P.: Data reduction and error analysis for physical science. New York, Toronto, London: McGraw-Hill Book Comp. 1969
2. Bradley, J. V.: Distribution-free statistical tests. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. 1968
3. Dixon, W. J., Massey jur., F. J.: Introduction to statistical analysis, 3rd edit. Tokyo: McGraw-Hill Kogakusha Ltd. 1969
4. Doerffel, K.: Beurteilung von Analysenverfahren und -ergebnissen, 2. Aufl. Berlin, Heidelberg, New York: Springer 1965
5. Eckschlager, K.: Errors, measurements and results in chemical analysis. London, New York, Toronto, Melbourne: Van Nostrand Reinhold Comp. 1969
6. Fisher, A.: The mathematical theory of probabilities, Vol. 1, 2nd edit. (English translation), The Macmillan Comp., 1923; cited according to [2]
7. Gottschalk, G., Kaiser, R. E.: Einführung in die Varianzanalyse und Ringversuche, BI-Hochschultaschenbücher Nr. 775, Mannheim 1972

¹ Dr. K. Gundlach, Bundesanstalt für Bodenforschung, Hannover, has kindly drawn my attention to this fact (S.)

8. Gottschalk, G., Sansoni, B.: Information on processing in chemical analysis I and II, GDCh training courses, GSF, Neuherberg near Munich, 1973 to 1976 (printed manuscripts)
9. Gottschalk, G. W.: Auswertung quantitativer Analyseergebnisse, in: Kienitz, H., Bock, R., Fresenius, W., Huber, W., Tölg, G.: Analytiker Taschenbuch, Bd. 1, p. 63. Berlin, Heidelberg, New York: Springer 1980
10. Hays, W. L.: Statistics for the social sciences, 2nd edit. London, New York, Sydney, Toronto: Holt, Rinehart and Winston 1977
11. Iyer, R. K., Sansoni, B.: Some aspects of the statistical evaluation of trace element data, Proceedings of the 6th Indo-German Seminar on Trace Element Analysis and Characterization, Methods and Selected Applications, Jül-Spez. Report (in press) (1981)
12. Iyer, R. K., Sansoni, B.: Use of statistical techniques in the evaluation of environmental data, Proceedings of the International Seminar on Management of Environment, BARC, Bombay (in press) (1981)
13. Iyengar, G. V., Sansoni, B.: Sample preparation of biological materials for trace element analysis, in: N.N., Elemental analysis of biological materials: Current problems and techniques with special reference to trace elements, Technical Report Series No. 197, International Atomic Energy Agency, Vienna 1980, p. 73
14. Kaiser, H., Specker, H.: Fresenius Z. Anal. Chem. **149**, 46 (1956)
15. Kaiser, H.: Grundlagen zur Beurteilung von Analyseverfahren, Methodicum Chemicum, Bd. 1 (Analytik), Teil 1, S. 1. Stuttgart: G. Thieme and New York, London: Acad. Press 1973
16. Kaiser, R., Gottschalk, G.: Elementare Tests zur Beurteilung von Meßdaten, BI-Hochschultaschenbücher Nr. 774, Mannheim 1972
17. Kendall, M. G., Buckland, W. R.: A dictionary of statistical terms, 3rd edit., reprint. London: Longman Group Ltd. 1976
18. Kreyszig, E.: Statische Methoden und ihre Anwendungen, 3. Aufl. Göttingen: Vandenhoeck and Ruprecht 1968
19. Laitinen, H. A., Harris, W. E.: Chemical analysis, an advanced text and reference, 2nd edit., p. 531. New York, Toronto, London: McGraw-Hill Book Comp. 1975
20. Mandel, J.: The statistical analysis of experimental data, Intersci. Publ., a Division of John Wiley and Sons, New York, London, Sydney 1964
21. Massart, D. L., Dijkstra, A., Kaufmann, L.: Evaluation and optimization of laboratory methods and analytical procedures, a survey of statistical and mathematical techniques. Amsterdam: Elsevier Sci. Publ. Comp. 1978
22. Nalimov, V. V.: The application of mathematical statistics to chemical analysis. Oxford, London, Paris, Frankfurt: Pergamon Press 1963
23. Panday, V. K., Sansoni, B.: Ashing; in: Sansoni, B., Iyengar, V., Sampling and sample preparation methods for the analysis of trace elements in biological material, Jül-Spez. Report, May 1978, ISSN 0343-7639, p. 65
24. Sachs, L.: Angewandte Statistik, Planung und Auswertung, Methoden und Modelle. Berlin, Heidelberg, New York: Springer 1974
25. Sansoni, B.: Schnellmethoden zur Überwachung der Radioaktivität in der Bundesrepublik Deutschland, in: Rapid methods for measuring radioactivity in the environment, Proceedings of an International Symposium, Neuherberg, 5–9 July 1971, Proceedings Series, International Atomic Energy Agency, Vienna 1971, p. 655
26. Sansoni, B.: Analytische Methoden der Umweltüberwachung, GSF-Bericht S 188 (1972), Neuherberg, near Munich; also in: Umweltschutz bei nuklearer und konventioneller Energiegewinnung (Hrsg.: Braun, H., Henck, F., Ladner, H.-A., Messerschmidt, O., Niklas, K., Rausch, L.), p. 238. Stuttgart: G. Thieme 1973
27. Sansoni, B.: Einführung in die Statistische Verarbeitung und Beurteilung von Daten in der Analytik, GSF-Bericht S-343, GSF, Neuherberg near Munich, 1975
28. Sansoni, B.: Fundamentals of trace element analysis and its future trends, 6th Indo-German Seminar on Trace Element Analysis and Characterization, Methods and Selected Applications, Maria Laach, Oct. 15. – 20., 1978, Proceedings (in preparation)
29. Sansoni, B.: Reliability of analytical environmental data – an important but often overlooked problem in the management of environment, International Seminar on Management of Environment, Bombay, February 11.–15., Proceedings (in preparation) (1981)
30. Sansoni, B.: a) Analytical data and their evaluation in environment protection, GDCh training course on Information Processing in Chemical Analysis, GSF, Neuherberg, near Munich, May 1976. – b) Reliability of data in trace analysis of elements and radionuclides in the environment, Indo-German Workshop on Chemistry of Metals in Trace Levels, BARC, Bombay, 1977. – c) Principles of trace element analysis and its future trends (plenary lecture) IAEA-Postcongress Symposium on Future Trends in Chemistry, Kyoto 1977. – d) Beurteilung spurenanalytischer Verfahren und Analysendaten, Methoden-kolloquium „Geochemie umweltrelevanter Spurenstoffe“, Deutsche Forschungsgemeinschaft, Schwäbisch Gmünd, Juni 1977. – e) Die Beurteilung von Umweltdaten am Beispiel toxischer Metalle, Vortragsveranstaltung der Arbeitsgemeinschaft der Großforschungszentren über Umweltchemikalien, Bonn-Bad Godesberg, Oktober 1979. – f) Unpublished
31. Sansoni, B., Alian, A., Iyer, A. K., Petri, H.: Trans. Amer. Nucl. Soc. **32**, 134 (1979)
32. Sansoni, B., Fischer, J., Ringer, H., Schultz, W., Dietl, F., Haimerl, H., Kurth, R.: Investigations on air dust monitoring for toxic metals at the Northern Border and in the city of Munich, 1970 to 1976; to be published, preliminary presentation
33. Sansoni, B., Iyengar, G. V.: Sampling and sample preparation methods for the analysis of trace elements in biological materials, Jül-Spez-13, Report KFA Jülich, ISSN 0343-7639, May 1978
34. Sansoni, B., Iyengar, G. V.: Sampling and storage of biological materials for trace element analysis, in: N.N., Elemental analysis of biological materials: Current problems and techniques with special reference to trace elements, Technical Report Series No. 197, International Atomic Energy Agency, Vienna 1980, p. 57
35. Sansoni, B., Iyer, R. K.: Data concentration of environmental analytical data with toxic metal contents as an example, Annual Symposium on the Analytical Chemistry of Pollutants, Dortmund, May 28 – 30, 1980
36. Sansoni, B., Iyer, R. K.: to be published
37. Sansoni, B., Iyer, R. K., Kurth, R., Fischer, J., Ringer, H., Schultz, W., Dietl, F., Haimerl, H.: Results of continuous monitoring of lead, cadmium, zinc and calcium in the air at the northern border of Munich 1973–1976, 10th Annual Symposium on the Analytical Chemistry of Pollutants, Dortmund, May 28 – 30, 1980
38. Sansoni, B., Kracke, W., Dietl, F., Fischer, J.: Mikrospurenbestimmung von Blei in verschiedenartigen Umweltproben durch flammenlose Atomabsorption nach externer Nassveraschung mit H_2O_2/Fe^{2+} , in: N.N., Proceedings of the International Symposium on Environmental Health Aspects of Lead, European Community, Amsterdam (Netherlands), 1972, EUR 5004 d-e-f; p. 1107
39. Sansoni, B., Kracke, W., Dietl, F.: Lead and other toxic metal leaching from ceramic ware, Munich 1971 to 1978 (to be published)
40. Sansoni, B., Kracke, W., Ringer, H., Schmidt, W., Dietl, F., Fischer, J., Kreuzer, W.: Der Cadmiumgehalt ausgewählter Umweltproben 1971/73, in: N.N., Proceedings International

- Symposium Problems of the Contamination of Man and his Environment by Mercury and Cadmium, Commission of the European Communities, Luxemburg, 1974, EUR 5075, p. 149
41. Sansoni, B., Kracke, W., Winkler, R.: Rapid assay of environmental radioactive contamination with special references to a new method of wet ashing, in: Environmental contamination by radioactive materials, Proceedings Series, International Atomic Energy, Vienna, 1969, p. 487
 42. Savage, I. R.: J. Amer. Stat. Assoc. **48**, 844 (1953)
 43. Schmidt, W., Dietl, F., Sansoni, B.: Untersuchungen der Luft auf Schwebstoffe, Blei, Cadmium, Zink und andere Metallspuren I: Bestimmung von Blei, Cadmium, Calcium, Eisen, Kalium, Kupfer, Magnesium, Mangan, Natrium und Zink durch Flammen-Atomabsorption, GSF-Report S 371 (Dezember 1975), GSF-Neuherberg near Munich, 1975
 44. Sonnabend, E., Sansoni, B., Herzog, J., Kracke, W.: Dtsch. Zahnärztl. Z. **31**, 189 (1976); further results to be published
 45. Shaw, D. M.: Evaluation of data, in: K. H. Wedepohl (Ed.), Handbook of Geochemistry, p. 324. Berlin, Heidelberg, New York: Springer 1969
 46. Streuli, H.: Fresenius Z. Anal. Chem. **303**, 406 (1980)

Received November 28, 1980