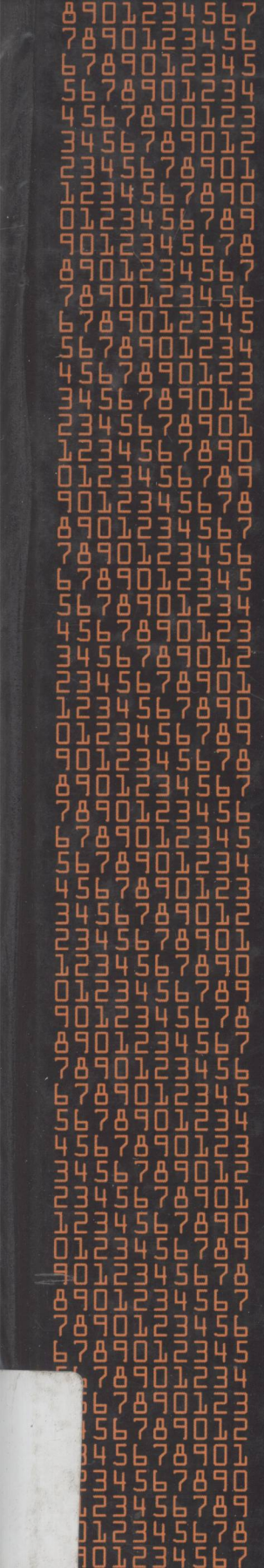

Statistical Treatment of Analytical Data

Zeev B. Alfassi, Zvi Boger and Yigal Ronen

Blackwell Publishing

CRC Press

Statistical Treatment of Analytical Data

Zeev B. Alfassi
Department of Nuclear Engineering, Ben-Gurion University, Israel

Zvi Boger
OPTIMAL – Industrial Neural Systems, Ltd, Beer Sheva, Israel

Yigal Ronen
Department of Nuclear Engineering, Ben-Gurion University, Israel

b
Blackwell
Science

CRC
CRC Press

A Blackwell Publishing Company

Editorial offices:
Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK
Tel: +44 (0)1865 776868
Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia
Tel: +61 (0)3 8359 1011

ISBN 0-632-05367-4

Published in the USA and Canada (only) by
CRC Press LLC, 2000 Corporate Blvd., N.W., Boca Raton, FL 33431, USA
Orders from the USA and Canada (only) to
CRC Press LLC

USA and Canada only:
ISBN 0-8493-2436-X

First published 2005

Library of Congress Cataloging-in-Publication Data:
A catalog record for this title is available from the Library of Congress

British Library Cataloguing-in-Publication Data:
A catalogue record for this title is available from the British Library

Set in 10/12 Times
by Kolam Information Services Pvt. Ltd, Pondicherry, India
Printed and bound in India
by Gopsons Papers Ltd, Noida

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on Blackwell Publishing, visit our website:
www.blackwellpublishing.com

Preface

Chapters 1–11 were written by Zeev B. Alfassi, chapter 12 was written by Yigal Ronen, and chapter 13 was written by Zvi Boger.

Zeev B. Alfassi
Zvi Boger
Yigal Ronen

Contents

1 Introduction

1.1 Statistics and quality assurance, control and assessment

The appraisal of quality has a considerable impact on analytical laboratories. Laboratories have to manage the quality of their services and to convince clients that the advocated level of quality is attained and maintained. Increasingly accreditation is demanded or used as evidence of reliability. At present there are American and European standards (ISO 25 and EN45001) that describe how a laboratory ought to be organized in order to manage the quality of its results. These standards form the basis for accreditation of analytical labs. Terms used frequently are *quality assurance* and *quality control.* Quality assurance is a wider term which includes both quality control and quality assessment.

Quality control of analytical data (QCAD) was defined by the ISO Committee as: 'The set of procedures undertaken by the laboratory for continuous monitoring of operations and results in order to decide whether the results are reliable enough to be released'. QCAD primarily monitors the batch-wise accuracy of results on quality control materials, and precision on independent replicate analysis of 'test materials'. Quality assessment was defined (Taylor 1987) as 'those procedures and activities utilized to verify that the quality control system is operating within acceptable limits and to evaluate the data'.

The standards of quality assurance (American ISO 25; European EN 45001) were written for laboratories that do analyses of a routine nature and give criteria for the implementation of a quality system which ensures an output with performance characteristics stated by the laboratory. *An important aspect of the quality assurance system is the full documentation of the whole analysis process.* It is essential to have well designed and clear worksheets. On the worksheets both the raw data and the calculated results of the analyses should be written. Proper worksheets reduce the chances of computing error and enable reconstruction of the test if it appears that a problem has occurred. The quality assurance system (or Standard) also treats the problems of personnel, equipment, materials and chemicals. The most important item is the methodology of the analysis. Quality control is not meaningful unless the methodology used has been validated properly. Validation of a methodology means the proof of suitability of this methodology to provide useful analytical data. A method is validated when the performance characteristics of the method are adequate and when it has been established that the measurement is under statistical control and produces accurate results.

'Statistical control' is defined as 'A phenomenon will be said to be "statistically controlled" when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction means that we can state at least

approximately, the probability that the observed phenomenon will fall within the given limits.'

The quality assurance systems required for accreditation of analytical laboratories are very important and are dealt with in several recent books (Kateman & Buydens 1987; Guennzler 1994; Funk *et al.* 1995; Pritchard 1997). However, these systems are well beyond the scope of this book, which will be devoted mainly to *quality assessment* of analytical data.

The quality of chemical analysis is usually evaluated on the basis of its uncertainty compared to the requirements of the users of the analysis. If the analytical results are consistent and have small uncertainty compared to the requirements, e.g. minimum or maximum concentration of special elements in the sample and its tolerances, the analytical data are considered to be of adequate quality. When the results are excessively variable or the uncertainty is larger than the needs, the analytical results are of low or inadequate quality. Thus, the evaluation of the quality of analysis results is a relative determination. What is high quality for one sample could be unacceptable for another. *A quantitative measurement is always an estimate of the real value of the measure and involves some level of uncertainty*. The limits of the uncertainty must be known within a stated probability, otherwise no use can be made of the measurement. *Measurement must be done in such a way that could provide this statistical predictability*.

Statistics is an integral part of quality assessment of analytical results, e.g. to calculate the precision of the measurements and to find if two sets of measurements are equivalent or not (in other words if two different methods give the same result for one sample).

Precise and accurate, which are synonyms in everyday language, have distinctly different meaning in analytical chemistry methodology. There are precise methods, which means that repeated experiments give very close results which are inaccurate since the measured value is not equal to the true value, due to systematic error in the system. For example, the deuterium content of a H_2O/D_2O mixture used to be determined by the addition of $LiAlH_4$, which reduces the water to hydrogen gas. The gas is transferred and measured by a mass spectrometer. However, it was found that although the method is precise, it is inaccurate since there is an isotope effect in the formation of the hydrogen.

Figure 1.1 explains simply the difference between precision and accuracy. *Statistics deals mainly with precision*, while accuracy can be studied by comparison with known standards. In this case, statistics play a role in analyzing whether the results are the same or not.

Old books dealt with only statistical methods. However the trend in the last decade is to include other mathematical methods that are used in analytical chemistry. Many analytical chemists are using computer programs to compute analytically areas of the various peaks in a spectrum or a chromatogram (in a spectrum the intensity of the signal is plotted vs. the wavelength or the mass [in mass spectra], while in the chromatogram it is plotted as a function of the time of the separation process). Another example is the use of the Fourier Transform either in 'Fourier Transform Spectroscopy' (mainly FTIR and FT-NMR, but recently also other spectroscopies) or in smoothing of experimental curves. The combination of statistics

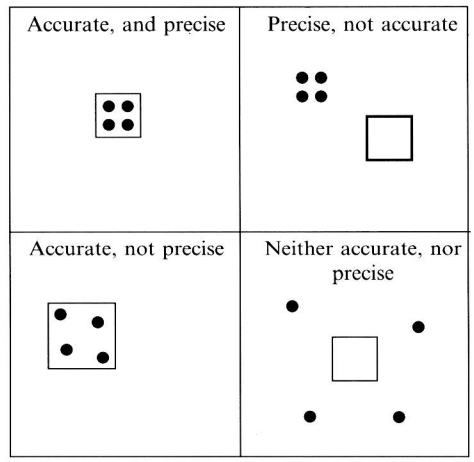

Fig. 1.1 Illustration of the meaning of accuracy and precision.

and other mathematical methods in chemistry is often referred to as chemometrics. However due to the large part played by statistics, and since many are 'afraid' of the general term of chemometrics, we prefer the title of Statistical and Mathematical Methods. These methods can be used as a black box, but it is important for educated analysts to understand the basic theory in order to take advantages of the full possibilities of these techniques and to choose intelligently the parameters as well as recognizing the limitation of these methods. It is clear that the choice of the mathematical tools is subjective, hence some methods are not included in this book because the authors feel that they are less important. Including the other methods would make this book too large.

1.2 References

Funk, W., Damman, V. & Donnevert, G. 1995, *Quality Assurance in Analytical Chemistry*, VCH, Weinheim.
Guennzler, H. 1994, *Accreditation and Quality Assurance in Analytical Chemistry*, Springer, Berlin.
Kateman, G. & Buydens, L. 1987, *Quality Control in Analytical Chemistry*, John Wiley, New York.
Pritchard, E. 1997, *Quality in the Analytical Chemistry Lab*, John Wiley, Chichester, UK.
Taylor, G. K. 1987, *Quality Assurance of Chemical Measurements*, John Wiley, Chichester, UK.

2 Statistical measures of experimental data

2.1 Mean and standard deviation

One of the best ways to assess the reliability of the precision of a measurement is to repeat the measurement several times and examine the different values obtained. Ideally, all the repeating measurements should give the same value, but in reality the results deviate from each other. Ideally, for a more precise result many replicate measurements should be done, however cost and time usually limit the number of replicate measurements possible. Statistics treats each result of a measurement as an item or individual and all the measurements as the *sample*. All possible measurements, including those which were not done, are called the *population*.

The basic parameters that characterize a population are the *mean*, μ, and the *standard deviation*, σ. In order to determine the *true* μ and σ, the entire population should be measured, which is usually impossible to do. In practice, measurement of several items is done, which constitutes a sample. Estimates of the mean and the standard deviation are calculated and denoted by $\bar{x}$ and s, respectively. The values of $\bar{x}$ and s are used to calculate confidence intervals, comparison of precisions and significance of apparent discrepancies. The mean, $\bar{x}$, and the standard deviation, s, of the values $x_1, x_2, \ldots, x_n$ obtained from n measurements is given by the equations:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{2.1a}$$

$$s = \sqrt{\left(\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}\right)} \tag{2.2a}$$

These equation can be written in a shorter way using the Σ notation:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1b}$$

$$s = \sqrt{\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}\right)} = \sqrt{\left(\frac{\sum_{i=1}^{n} x_i^2}{n - 1}\right) - \frac{n(\bar{x})^2}{n - 1}} = \sqrt{\left(\frac{\sum_{i=1}^{n} x_i^2}{n - 1} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n(n - 1)}\right)} \tag{2.2b}$$

In some older books the use of the term 'average' instead of 'mean' (Youden 1994), can be found, but the common term nowadays is 'mean'. There are different kinds of 'means' (Woan 2000) (e.g. arithmetic mean, harmonic mean), but if not

explicitly written the 'mean' is meant to be the arithmetic mean as defined by Equation (2.1).

There are several reasons why the arithmetic mean and not the other ones is chosen. The main reason is because it is the simplest one:

$$\text{Arithmetic mean:} \quad \bar{x}_a = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

$$\text{Geometric mean:} \quad \bar{x}_g = (x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{1/n}$$

$$\text{Harmonic mean:} \quad \bar{x}_h = n\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)^{-1}$$

Another reason to choose the arithmetic mean is that it fulfils the least squares criterion (Cantrell 2000), i.e. $\bar{x}_a$ fulfils the requirement:

$$\sum_{j=1}^{n} (x_j - \bar{x}_a)^2 = \text{minimum}$$

The names of these means come from the corresponding sequences. If we have an odd number of consecutive terms of a geometric sequence, then the middle term is given by the geometric mean of all these terms. The same is true for the arithmetic mean (in the case of an arithmetic sequence) and for the harmonic mean (in the case of an harmonic sequence). From now on we will use only the arithmetic mean and will refer to it in the general form:

$$\bar{x} = \bar{x}_a = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{2.1c}$$

The mean of the sum of squares of the deviation of the observed data from the mean is called the *variance*:

$$V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1} \tag{2.3}$$

The division by $(n - 1)$ and not by n is done because we do not know the true value of $\bar{x}$, i.e. μ, and instead we used the calculated value of $\bar{x}$. For the calculation of $\bar{x}$, we use one degree of freedom (one unit of information), and this is the reason that we divide by $(n - 1)$ (the number of degrees of freedom, i.e. the number of free units of information which were left).

The dimension of *the variance*, V, is the square of the dimension of our observation and in order to get the same dimension we take the square root of V, which is called the *standard deviation*, s. In many cases the variance is not denoted by V, but is written as s^2.

$$s = \sqrt{\left(\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}\right)} = \sqrt{\left(\frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n(n-1)}\right)} \tag{2.2c}$$

The values of $\bar{x}$ and s can be calculated using a computer program or a calculator. It is important to note that all scientific calculators have two keys, one depicted

as σ_n and the other one as σ_{n-1}. Equation (2.2) fits the key σ_{n-1}. The other key uses n instead of $(n-1)$ in Equation (2.2). The key σ_{n-1} gives the standard deviation of our sample, but not of the whole population, which can be obtained by doing an infinite number of repeated measurements. In other words, σ_n is the standard deviation if the true mean μ is known. Otherwise, one degree of freedom is lost on the calculation of $\bar{x}$. For a small number of repetitions, the equation with $(n-1)$ gives a better estimate of the true σ, which is unknown. The mean $\bar{x}$ is a better estimate for the true value than one measurement alone. The standard deviation σ (or its estimate s) represents the dispersion of the measured values around the mean. The standard deviation has the same dimension as that of the measured values, x_i. Often, analysts prefer to use a dimensionless quantity to describe the dispersion of the results. In this case they use the *relative standard deviation* as a ratio (SV) (also called the *coefficient of variation*, CV) or as a percentage (RSD):

$$\text{SV} = s/\bar{x} \tag{2.4}$$

$$\text{RSD} = \text{CV} \times 100 \tag{2.5}$$

When calculating small absolute standard deviations using a calculator, sometimes considerable errors are caused by rounding, due the limited number of digits used. In order to overcome this problem, and in order to simplify the punching on the calculator, it is worth subtracting a constant number from all the data points, so that x_i will be not large numbers but rather of the same magnitude as their differences. The standard deviation will be unchanged but *the subtracted constant should be added to the mean.* In other words, if we have n data points, $x_1, \ldots, x_n$, which are large numbers, it is better to key into the calculator $(x_1 - c)$, $(x_2 - c), \ldots, (x_n - c)$ such that $(x_i - c)$ are no longer large numbers. The real mean of x_i is $\bar{x}_i = c + (\overline{x_i - c})$ and the standard deviation remains the same, $s(x_i) = s(x_i - c)$. Thus for calculating the mean and standard deviation of 50.81, 50.85, 50.92, 50.96, 50.83, we can subtract the constant 50.8, key 0.01, 0.05, 0.12, 0.16, 0.03 and obtain $\bar{x} = 0.074$ and $s = 0.06348$. The real mean is $50.8 + 0.074 = 50.874$ and s remains the same i.e. 0.06348. We could subtract only 50, key 0.81, 0.85, 0.92, 0.96, 0.83 and will obtain $\bar{x} = 0.874$ and $s = 0.06348$. The real mean is $50 + 0.874 = 50.874$ and s is 0.06348 as before.

Usually we choose the constant c as the smallest integer number of our data, so that the smallest number of $(x_i - c)$ is less than one. For example, if the data points are 92.45, 93.16, 91.82, 95.43, 94.37, we subtract 91 from all the data points, and calculate the mean and standard deviation of 1.45, 2.16, 0.82, 4.43, 3.37. The calculator will give $\bar{x} = 2.446$ and $s = 1.4584$. Adding 91 to the obtained mean, we get $\bar{x} = 93.446$ and $s = 1.4584$. Some will find it more easy to subtract just 90.

2.1.1 *Significant figures*

At this stage it is important to emphasize the importance of significant figures, especially nowadays when all calculations are made with calculators or computers, which yield results with many digits. Since our original data were given with two

digits after the decimal point, any additional digits are meaningless. Consequently in the previous example there is no point giving $\bar{x} = 93.446$; we should round it off to $\bar{x} = 93.45$ and similarly $s = 1.46$. Usually the number of significant figures does not refer to the number of decimal digits but to the total number of figures. Thus, for example, the number 92.45 has four significant figures. This means that our precision of the measurement is 10^{-4}. In this case a result should not be given as 25.3 but rather as 25.30, in order to emphasize the precision of the measurement. *The mean of values should have the same number of significant figures as the values themselves.* However, the standard deviation, which is usually smaller, should have the same number of decimal digits as the measurements themselves, rather than the same number of significant figures. Thus, in our example we use for s only three significant figures i.e. $s = 1.46$, since the important factor is the decimal digits.

2.1.2 *Frequency tables*

When large numbers of measurements are made (on the same aliquot if it is not consumed by the measurement, or on different aliquots of the same sample or on different samples), some values are obtained more than once. Sometimes, instead of discrete values, a range of values is chosen as one value. In both cases it is simpler to concentrate the data in a *frequency table* – a table that gives the number of times (named frequency) each value was obtained. For example, the concentration of salt in drinking water was measured each day for a whole year. The results are given in Table 2.1 (given to two significant figures).

In this case the mean and the standard deviation are calculated by the equations:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} \quad \Rightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \tag{2.6}$$

Table 2.1 Concentration of salt in drinking water measured each day for one year.

Concentration (mg/ℓ) x_i	Numbers of days f_i
3.5	18
3.6	22
3.7	25
3.8	35
3.9	46
4.0	55
4.1	45
4.2	40
4.3	32
4.4	27
4.5	20

$$s = \sqrt{\left(\frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \ldots + f_n(x_n - \bar{x})^2}{f_1 + f_2 + \ldots + f_n - 1}\right)} \Rightarrow s = \sqrt{\left(\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} f_i\right) - 1}\right)} \tag{2.7}$$

The summation is carried out over all the various values of x_i (n different values) and the total number of measurements is:

$$f_1 + f_2 + \ldots + f_n = \sum_{i=1}^{n} f_i \tag{2.8}$$

Most scientific calculators can calculate the mean value and the standard deviation from frequency tables. In our example the following results will be obtained:

$\bar{x} = 4.0,\ s = 0.3$

(remember to use the n−1 key). The units of both $\bar{x}$ and s are the same as each sample, i.e. mg/ℓ. In short the concentration of the salt is written as 4.0 ± 0.3 mg/ℓ.

2.2 Graphical distributions of the data – bar charts or histograms

The standard deviation gives a measure of the spread of the results around the mean value. However, it does not indicate the shape of the spread.

Frequency tables and, even more so, drawing them as a rod diagram or as a histogram give a clearer picture of the spread of the measurement. A histogram describes the real situation better than bar charts since the real values are not discrete values of only two significant digits, and 3.7 mg/ℓ stands, for example, for the range 3.650 01 to 3.750 00. If the table were to three rather than two significant digits, there would be many more columns in the histogram. Increasing the number of measurements and the number of significant figures will lead to a continuous distribution.

Most spreadsheet data programs, such as Lotus 1-2-3, Quattro Pro, Excel or Origin can draw the frequency table in the form of column charts or histograms. Figures 2.1 and 2.2 are, for example, the result of the use of the chart wizard of Excel on the data of Table 2.1.

2.3 Propagation of errors (uncertainties)

In some cases we are interested in a value of a variable, which cannot be determined directly but can be calculated from several measurements of different properties. Thus for the measurement of the area of a rectangle we need to measure both its length L and the width W. The area A, is given by:

$A = L \times W$

For the volume of a box, V, we need in addition to measure its height, H:

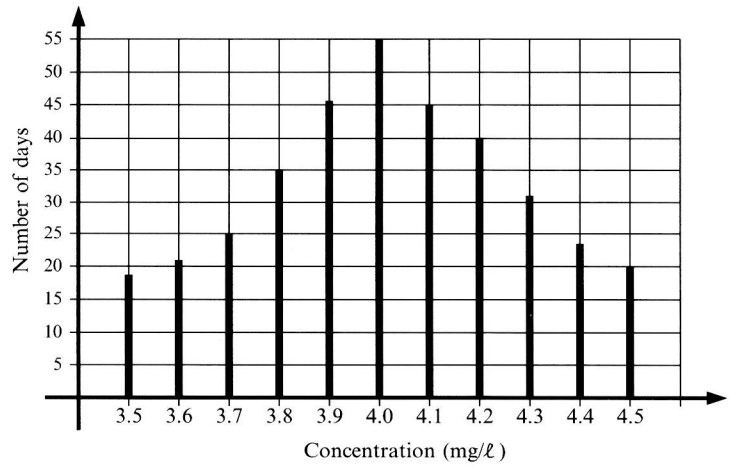

Fig. 2.1 Rod chart.

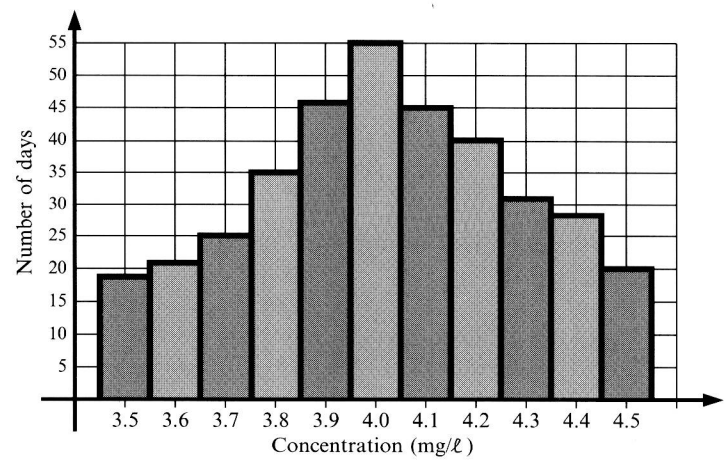

Fig. 2.2 Histogram.

$$V = L \times W \times H$$

How do the uncertainties (possible errors) in the estimation of L and W affect the resulting uncertainty in the value of the area, A? One way to calculate the possible error in A is to take the highest values of L and W, calculate from them the obtained A and compare it with the average value and the minimal values. Thus:

$$\overline{L},\ \overline{W} \Rightarrow \overline{A} = \overline{L} \times \overline{W}$$

$$\overline{L} + \Delta L,\ \overline{W} + \Delta W \Rightarrow \overline{A} + \Delta A = (\overline{L} + \Delta L) \times (\overline{W} + \Delta W)$$
$$= \overline{L}\,\overline{W} + \overline{L} \times \Delta W + \overline{W} \times AL + \Delta W \times \Delta L$$

$$\overline{L} - \Delta L,\ \overline{W} - \Delta W \Rightarrow \overline{A} - \Delta A = (\overline{L} - \Delta L) \times (\overline{W} - \Delta W)$$
$$= \overline{L}\,\overline{W} - (\overline{L} \times \Delta W + \overline{W} \times AL) + \Delta W \times \Delta L$$

If we assume that the last term ($\Delta W \times \Delta L$) can be neglected, due to the fact that the product of two small terms will lead to a smaller term, we can see that both directions will lead to the same value of ΔA:

$$\Delta A = \overline{L} \times \Delta W + \overline{W} \times \Delta L \tag{2.9}$$

The same equation will be obtained by calculus:

$$\mathrm{d}A = \frac{\partial A}{\partial L}\,\mathrm{d}\ell + \frac{\partial A}{\partial W}\,\mathrm{d}W \;\Rightarrow\; \mathrm{d}A = W\,\mathrm{d}\ell + \ell\,\mathrm{d}W \tag{2.10}$$

In the general case, where y was measured from the separate quantities x, z, etc., we can write:

$$y = f(x,\ z,\ \ldots) \tag{2.11}$$

The mean of y is calculated from the mean values of the different quantities:

$$\overline{y} = f(\overline{x},\ \overline{z},\ \ldots) \tag{2.12}$$

The different values of y can be written as:

$$y_i - \overline{y} \cong (x_i - \overline{x})\,\frac{\partial y}{\partial x} + (z_i - \overline{z})\,\frac{\partial y}{\partial z} + \ldots \tag{2.13}$$

and the variance σ_y^2 is:

$$\sigma_y^2 = \lim_{N\to\infty} \frac{1}{N} \sum (y_i - \overline{y})^2 \tag{2.14}$$

The variance, σ_y^2, can be expressed in terms of the variance of the separate measured quantities σ_x^2, σ_z^2, etc:

$$\sigma_y^2 \cong \lim_{N\to\infty} \frac{1}{N} \sum \left[(x_i - \overline{x})\,\frac{\partial y}{\partial x} + (z_i - \overline{z})\,\frac{\partial y}{\partial z} + \ \ldots \right]^2$$

$$\sigma_y^2 = \lim_{N\to\infty} \frac{1}{N} \sum \left[(x_i - \overline{x})^2 \left(\frac{\partial y}{\partial x}\right)^2 + (z_i - \overline{z})^2 \left(\frac{\partial y}{\partial z}\right)^2 + 2(x_i - \overline{x})\,(z_i - \overline{z}) \left(\frac{\partial y}{\partial x}\right) \left(\frac{\partial y}{\partial z}\right) + \ \ldots \right]$$

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x}\right)^2 \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \overline{x})^2 + \left(\frac{\partial y}{\partial z}\right)^2 \lim_{N\to\infty} \frac{1}{N} \sum (z_i - \overline{z})^2 + 2\,\frac{\partial y}{\partial x}\,\frac{\partial y}{\partial z} \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \overline{x})\,(z_i - \overline{z}) + \ldots$$

The first two sums are σ_x^2 and σ_z^2 respectively:

$$\sigma_x^2 = \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \overline{x})^2 \qquad \sigma_z^2 = \lim_{N\to\infty} \frac{1}{N} \sum (z_i - \overline{z})^2 \tag{2.15}$$

Similarly the third sum can be defined as σ_{xz}^2:

$$\sigma_{xz}^2 = \lim_{N\to\infty} \frac{1}{N} \sum (x_i - \overline{x})(z_i - \overline{z}) \tag{2.16}$$