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

Blackwell
Science

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

Contents

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

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

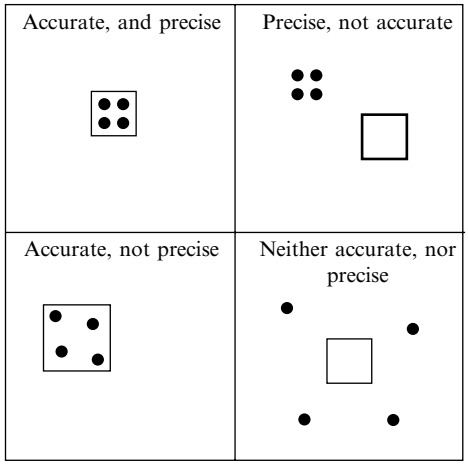

Fig. 1.1 Illustration of the meaning of accuracy and precision.

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 $\overline{x}$ and s, respectively. The values of $\overline{x}$ and s are used to calculate confidence intervals, comparison of precisions and significance of apparent discrepancies. The mean, $\overline{x}$, and the standard deviation, s, of the values $x_1, x_2, \ldots, x_n$ obtained from n measurements is given by the equations:

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

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

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

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

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

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:

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

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

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

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

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

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:

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

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

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

The division by $(n - 1)$ and not by n is done because we do not know the true value of $\overline{x}$, i.e. μ, and instead we used the calculated value of $\overline{x}$. For the calculation of $\overline{x}$, 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.

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

The values of $\overline{x}$ 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 $\overline{x}$. For a small number of repetitions, the equation with $(n-1)$ gives a better estimate of the true σ, which is unknown. The mean $\overline{x}$ 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):

$$\mathrm{SV} = s/\overline{x} \tag{2.4}$$

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

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 $\overline{x}_i = c + (\overline{x_i - c})$ 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 $\overline{x} = 0.074$ 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 $\overline{x} = 0.874$ 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 $\overline{x} = 2.446$ and $s = 1.4584$. Adding 91 to the obtained mean, we get $\overline{x} = 93.446$ 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 $\overline{x} = 93.446$; we should round it off to $\overline{x} = 93.45$ 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:

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

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

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

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:

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

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

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:

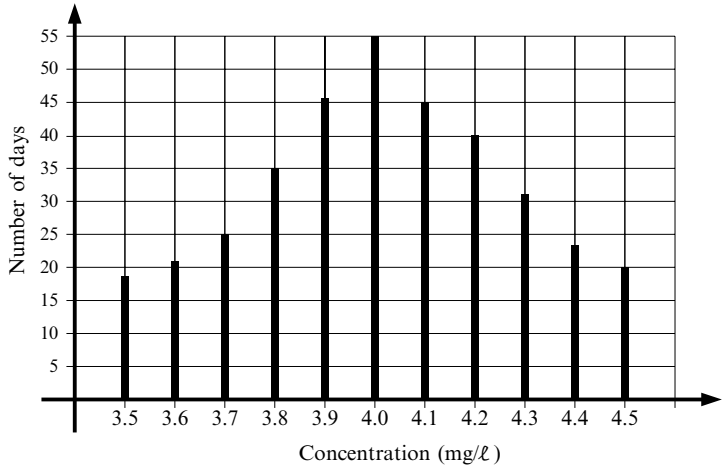

Fig. 2.1 Rod chart.

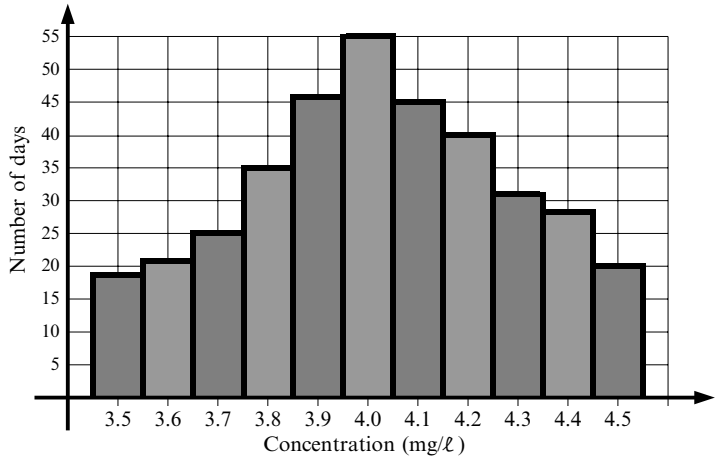

Fig. 2.2 Histogram.

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

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:

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

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

$$= \overline{L}\,\overline{W} + \overline{L} \times \Delta W + \overline{W} \times AL + \Delta W \times \Delta L$$

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

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

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:

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

With this definition the approximation for the standard deviation of y is given by:

$$\sigma_y^2 \simeq \sigma_x^2\left(\frac{\partial y}{\partial x}\right)^2 + \sigma_z^2\left(\frac{\partial y}{\partial z}\right)^2 + 2\sigma_{xz}^2\left(\frac{\partial y}{\partial x}\right)\left(\frac{\partial y}{\partial z}\right) + \ldots \tag{2.17}$$

The first two terms, which will presumably dominate, are averages of squares of deviation. The third term is the average of cross terms. If the functions in x and z are uncorrelated, the cross term will be small and will vanish in the limit of a large number of random observations. Thus for uncorrelated variables:

$$\sigma_y^2 \simeq \sigma_x^2\left(\frac{\partial y}{\partial x}\right)^2 + \sigma_z^2\left(\frac{\partial y}{\partial z}\right)^2 \tag{2.18}$$

Where there are more than two variables, each variable will contribute a similar term.

Addition and subtraction: If y is given by a linear combination of x and z, i.e. $y = ax \pm bz$, then:

$$\frac{\partial y}{\partial x} = a \qquad \frac{\partial y}{\partial z} = \pm\, b$$

Hence:

$$y = ax \pm bz \;\Rightarrow\; \sigma_y^2 = a^2\sigma_x^2 + b^2\sigma_y^2 \pm 2ab\sigma_{xz}^2 \tag{2.19}$$

In most cases the errors in x and z are uncorrelated and the mixed covariance, σ_{xz}, is equal to zero. However if the errors are correlated, σ_y^2 might vanish due to compensations by the covariance σ_{xz}^2. Usually we write:

$$\sigma_y^2 = a^2\sigma_x^2 + b^2\sigma_z^2 \tag{2.20}$$

Multiplication and division: If y is given by $y = axz$ or $y = a(x/z)$ then:

$$y = axz \quad \Rightarrow \quad \frac{\partial y}{\partial x} = az \qquad \frac{\partial y}{\partial z} = ax$$

Thus for multiplication:

$$\sigma_y^2 = a^2z^2\sigma_x^2 + a^2x^2\sigma_z^2 \tag{2.21}$$

Dividing by $y^2 = a^2x^2z^2$ leads to:

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_z^2}{z^2} \tag{2.22}$$

$$y = a\left(\frac{x}{z}\right) \;\Rightarrow\; \frac{\partial y}{\partial x} = \frac{a}{z} \qquad \frac{\partial y}{\partial z} = -\frac{ax}{z^2}$$

$$y = a\left(\frac{x}{z}\right) \;\Rightarrow\; \sigma_y^2 = \left(\frac{a^2}{z^2}\right)\sigma_x^2 + \left(\frac{a^2x^2}{z^4}\right)\sigma_z^2$$

Dividing by $y^2 = a^2x^2/z^2$ will lead to the same equation for the relative standard deviation as in multiplication (Equation (2.22)):

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_z^2}{z^2}$$

Powers: If the function is $y = a\ x^b$, the derivative is:

$$\frac{\partial y}{\partial x} = b\ a\ x^{b-1} = b\ \frac{a x^b}{x} = b\ \frac{y}{x}$$

According to Equation (2.18), the relative σ_y:

$$\sigma_y^2 = b^2\ \frac{y^2}{x^2}\ \sigma_x^2\ \Rightarrow\ \frac{\sigma_y}{\sigma_x} = b\ \frac{y}{x} \quad \text{or} \quad \frac{\sigma_y}{y} = b\ \frac{\sigma_x}{x} \tag{2.23}$$

Exponential: If the function is $y = a\ \mathrm{e}^{bx}$, the derivative is:

$$\frac{\partial y}{\partial x} = a\ b\ \mathrm{e}^{bx}\ \Rightarrow\ \frac{\partial y}{\partial x} = by$$

Hence the variance of y and the relative σ_y are:

$$\sigma_y^2 = b^2\ y^2\ \sigma_x^2\ \Rightarrow\ \frac{\sigma_y}{\sigma_x} = by\ \Rightarrow\ \frac{\sigma_y}{y} = b\,\sigma_x \tag{2.24}$$

If the base is not e but a constant c, we can write:

$$c^{bx} = \left(e^{\ln c}\right)^{b\,x} = e^{b \ln c\, x}$$

and using Equation (2.24) gives:

$$y = a\ c^{b\,x}\ \Rightarrow\ \frac{\sigma_y}{y} = b\ \ln\ c\ \sigma_x \tag{2.25}$$

Logarithm: If the function is $y = a \ln b\ x$, the derivative is:

$$\frac{\partial y}{\partial x} = \frac{a}{x}\ \Rightarrow\ \sigma_y^2 = \frac{a^2}{x^2}\ \sigma_x^2\ \Rightarrow\ \sigma_y = a\ \frac{\sigma_x}{x} \tag{2.26}$$

Looking at the functional form $y = a \ln b + \ln x^a$, we can see that the first term, which is a constant, has no influence and that σ_y for $y = x^a$ and $y = \ln x^a$ is the same.

In both cases,

$$\sigma_y = a\ \frac{\sigma_x}{x}$$

2.4 References

Cantrell, C. D. 2000, *Modern Mathematical Methods for Physicists and Engineers*, Cambridge University Press, p. 41.

Woan, G. 2000, *The Cambridge Handbook of Physics Formulas*, Cambridge University Press, p. 27.

Youden, W. J. 1994, *Experimentation and Measurement*, NIST Special Publication 672, US Department of Commerce.

3 Distribution functions

3.1 Confidence limit of the mean

The mean of a sample of measurements, $\overline{x}$, provides an estimate of the true value, μ, the quantity we are trying to measure. However, it is quite unlikely that $\overline{x}$ is exactly equal to μ, and an important question is to find a range of values in which we are certain that the true value lies. This range depends on the measured mean but also on the distribution of the various x_i, on the number of measurements done *and on the question of how certain we want to be*. The more certain we want to be, the larger the range we have to take. The larger the number of experiments done, the closer $\overline{x}$ is to μ, and a smaller range has to be taken for the same percentage of certainty. Usually statistics tables refer not to the number of repeated experiments but to the *degrees of freedom* (usually given as ν or df). In the previous chapter it was seen that in the calculation of the standard deviation, the number of degrees of freedom is $(n-1)$, where n is the number of repeated measurements. The number of degrees of freedom refers to the number of independent variables (units of information). When calculating the standard deviation n terms of $(x_i - \overline{x})^2$ are used, but only $(n-1)$ terms are independent, since the nth term can be calculated from the equation:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

i.e. from the definition of the mean. The standard deviation tells us if the spread of the data is large or small or, in other words, if the distribution function is wide or narrow. But from the few data we usually have, this cannot be deduced from the distribution function itself. Various distribution functions are *assumed* in the use of statistics. The width of the distribution function that is assumed comes from the confidence limit of the calculated mean. The narrower the distribution function, the better the results and the smaller the confidence limit (range).

3.2 Measurements and distribution functions

In some measurements only discrete results are available, e.g. the reaction occurs or does not occur. Also in the case of a colored product, although we can characterize the product by a continuous variable, the wavelength of maximum absorption (peak of absorption), we will usually refer to the color by discrete values i.e. the names: black, white, green, etc. In most analytical chemistry measurements the value of the result is of a continuous nature. Yet in many cases we will gather the various results into several groups, making the results discrete as each group has a unique value and the number of groups is limited (albeit by us).

Let us assume that we have k groups of balls of varying sizes. Each group contains balls of a range of sizes and there is no overlapping between the groups, i.e according to the size of the ball we know unequivocally to which group it belongs. If the ith group has a radius in the range a_i to b_i and there are n_i balls in the group, then the total number of balls is

$$\sum_{i=1}^{k} n_i$$

The probability that in choosing a ball at random its size will be between a_i and b_i is

$$\left(n_i \Big/ \sum_{i-1}^{k} n_i \right)$$

The plot of n_i vs. i gives us the distribution function of the size of the balls as a histogram rather than as a rod chart. The sum of all the rectangles in the histogram is $\sum n_i$. If we plot the ratio

$$\left(n_i \Big/ \sum_{i-1}^{k} n_i \right) \text{ vs. } i$$

we will get the *probability function* for a ball chosen at random to have a radius in the range a_i to b_i. The difference between these two plots (either n_i or the ratio $n_i / \sum n_i$) is the multiplication of the first one by the factor $(1/\sum n_i)$ in order to change the area under the function curve (histogram) from $\sum n_i$ to 1, since the *total probability must be unity*. This multiplication is usually called *normalization*. A normalized distribution function is one which has an area under its curve (the integral from $-\infty$ to ∞ for a continuous distribution function) of unity, and hence can be used as a *probability function*.

If the distribution function of an experiment is known, it can be used to calculate the confidence limits of the mean. This confidence limit depends on the *certainty* with which we want to know it. If, for example, we want to know the confidence limit for 95% certainty (5% uncertainty), we will take 0.025 from both sides of the maximum of the distribution function area. This area of $0.05 = (2 \times 0.025)$ represents 5% of the total area under the distribution (probability) function curve. For common distribution functions the values which fit different areas (uncertainties) are tabulated, since the integrals cannot be calculated analytically.

3.3 Mathematical presentation of distribution and probability functions

A probability function is a function that assigns a probability of occurrence to each outcome of an experiment or a measurement. In the case of discrete outcomes, a *probability function (PF)* $f(x)$ is defined as that which gives the probability of getting an outcome equal to x. In the case of continuous outcomes, a *probability density function* (*PDF*) $f(x)\,dx$ is defined as that which gives the probability that an outcome will lie in the interval between x and $(x + dx)$.

The probability must have a real non-negative value not greater than 1, so in both cases of discrete and continuous outcomes:

$$0 \leq \mathrm{f}(x) \leq 1$$

Since in any experiment we have some outcomes (results), the probabilities must sum to unity. Thus:

$$\text{for PF} \quad \sum_i \mathrm{f}(x_i) = 1 \qquad \text{and for PDF} \quad \int_{-\infty}^{\infty} \mathrm{f}(x)\, \mathrm{d}x = 1 \tag{3.1}$$

The summation or integration is done for all allowed values of x. For a PDF of common distribution functions, $\mathrm{f}(x)$ can be expressed analytically, however the integration must be done numerically. The numerical integration is tabulated for the *cumulative probability function* (*CPF*) $\mathrm{F}(x)$ defined as:

$$\mathrm{F}(x) = \int_{-\infty}^{x} \mathrm{f}(u)\, \mathrm{d}u$$

i.e. $\mathrm{F}(x)$ is the area under the probability density function $\mathrm{f}(u)$ from $-\infty$ to x.

The probability that the outcome of a measurement will lie between the limits a_1 and a_2 is given by $\mathrm{F}(a_2) - \mathrm{F}(a_1)$, i.e.

$$\mathrm{p}\,(a_1 \leq x \leq a_2) = \int_{a_1}^{a_2} \mathrm{f}(u)\, \mathrm{d}u = \int_{-\infty}^{a_2} \mathrm{f}(u)\, \mathrm{d}u - \int_{-\infty}^{a_1} \mathrm{f}(u)\, \mathrm{d}u = \mathrm{F}(a_2) - \mathrm{F}(a_1)$$

The mean and the variance are connected to the distribution function by definition of the *moments of the distribution function*.

The *kth moment* of a distribution about zero (the origin) which is assigned by $\mathrm{E}\left(x^{\mathrm{k}}\right)$ is defined by the equations:

$$\mathrm{E}\left(x^{\mathrm{k}}\right) = \sum_i u_i^k\, \mathrm{f}(u_i) \quad \text{for } \mathbf{PF} \text{ (discrete outcomes)}$$

$$\mathrm{E}\left(x^{\mathrm{k}}\right) = \int_{-\infty}^{\infty} u^k\, \mathrm{f}(u)\, \mathrm{d}u \quad \text{for } \mathbf{PDF} \text{ (continuous outcomes)}$$

The *mean* of a distribution (known also as its *expectation value*) is defined as the first moment about zero:

$$\mu = \mathrm{E}(x) = \sum_i u_i\, \mathrm{f}(u_i) \quad \text{for } \mathbf{PF} \quad \text{and} \quad \mu = \mathrm{E}(x) = \int_{-\infty}^{\infty} u\, \mathrm{f}(u)\, \mathrm{d}u \quad \text{for } \mathbf{PDF} \tag{3.2}$$

where u_i are all the possible values of the random variable x.

When we talk about the mean of a sample (a portion of the whole population) we call it $\overline{x}$ and similarly the standard deviation is designated by s. When we refer

to the whole population and similarly to the distribution function we use μ and σ for the mean and the standard deviation, respectively.

The moment about zero is a special case of the moment about the mean (center), where the center is chosen as zero. The definition of the *kth moment* of distribution about the center (the mean) is defined as:

$$\begin{aligned} \mathrm{E}\left[(x-\mu)^k\right] &= \sum (u_i-\mu)^k \, \mathrm{f}(u_i) \quad \text{for } \mathbf{PF} \text{ (discrete probability)} \\ \mathrm{E}\left[(x-\mu)^k\right] &= \int (u-\mu)^k \, \mathrm{f}(u) \, \mathrm{d}u \quad \text{for } \mathbf{PDF} \text{ (continuous probability)} \end{aligned} \tag{3.3}$$

where $\mu = \mathrm{E}(x)$

The *variance* (the square of the standard deviation) is defined as the second central moment (the second moment about the mean):

$$\sigma^2 = \mathrm{V}(x) = \mathrm{E}\left[(x-\mu)^2\right] = \sum (u_i-\mu)^2 \, \mathrm{f}(u_i) \quad \text{or} \quad \int (u_i-\mu)^2 \, \mathrm{f}(u) \, \mathrm{d}(u) \tag{3.4}$$

The variance, σ^2, can be transformed to a moment about zero using the following treatment:

$$\begin{aligned} \sigma^2 = \sum (x_i-\mu)^2 \, \mathrm{f}(x_i) &= \sum \left(x_i^2 - 2\mu x_i + \mu^2\right) \mathrm{f}(x_i) \\ &= \sum x_i^2 \, \mathrm{f}(x_i) - 2\mu \sum x_i \, \mathrm{f}(x_i) + \mu^2 \sum \mathrm{f}(x_i) \end{aligned}$$

As $\sum x_i \, \mathrm{f}(x_i)$ is defined as μ, and since the distribution function is normalized, $\sum \mathrm{f}(x_i) = 1$:

$$\sigma^2 = \sum x_i^2 \, \mathrm{f}(x_i) - 2\mu^2 + \mu^2 \quad \Rightarrow \quad \sigma^2 = \sum x_i^2 \, \mathrm{f}(x_i) - \mu^2 \tag{3.5a}$$

In the moments notation (exception value) we can write:

$$\sigma^2 = \mathrm{E}\left(x^2\right) - [\mathrm{E}(x)]^2 \quad \Rightarrow \quad \sigma^2 = \mathrm{E}\left(x^2\right) - \mu^2 \tag{3.5b}$$

In order to see an example for these definitions let us look at the throwing of a die. This is a discrete probability function, which for the six possible outcomes 1, 2, 3, 4, 5, 6 has a value of 1/6 , and for all other numbers the probability is zero.

$$\begin{aligned} \mu &= \sum (u_i) \, \mathrm{f}(u_i) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6} = 3.5 \\ \sigma^2 &= \sum (u_i-\mu)^2 \, \mathrm{f}(u_i) = \sum (u_i - 3.5)^2 \, \mathrm{f}(u_i) \\ \sigma^2 &= (-2.5)^2 \times \frac{1}{6} + (-1.5)^2 \times \frac{1}{6} + (-0.5)^2 \times \frac{1}{6} + 0.5^2 \times \frac{1}{6} + 1.5^2 \times \frac{1}{6} + 2.5^2 \times \frac{1}{6} \\ &= \frac{17.5}{6} = 2.9167 \\ \sigma &= \sqrt{2.9167} \quad \Rightarrow \quad \sigma = 1.7078 \end{aligned}$$

In most analytical chemical experiments we have only continuous distribution functions, yet in some cases discrete distribution functions are important. For this reason we will learn about the distributions of both types. In most cases the 'true' distribution function corresponding to our measurements is not known. The

number of measurements done is usually too small to find the real distribution function. Here we will learn about some theoretical distribution functions and will use them as an approximation for the real situation. From the systematic point of view it might be more appropriate to start with discrete probability functions, however as continuous functions are more common in analytic chemistry, we will start with them.

3.4 Continuous distribution functions

3.4.1 *The normal distribution function*

The most common distribution function is the *normal* one, a distribution which fits most natural phenomena, although not all of them, when the sample is large enough and the random errors are sufficiently small. Most natural phenomena are symmetric about the mean, forming a bell-shape plot. For a function to be symmetrical about the mean μ, it should involve either $|x-\mu|$ or $(x-\mu)^n$, where n is an even integer, the simplest case being $n=2$. The most appropriate form of mathematical function should be proportional to $\exp[-c|x-\mu|]$ or $\exp\left[-c(x-\mu)^2\right]$. The inclusion of c is to allow for different widths of the distribution. After fixing c, the proportionality constant, A, can be determined by normalizing the distribution function, i.e requiring that $\int_{-\infty}^{\infty} f(x)\,dx = 1$. For the normal distribution, c is chosen as $c = 1/(2\sigma^2)$. Substituting $u = (x-\mu)/\sigma$, the normal distribution function becomes

$$f(u) = A\ \exp\left(-u^2/2\right)$$

This substitution of x by $(x-\mu)$ leads all normal distribution functions to be symmetric around the same number – the origin. The division of $(x-\mu)$ by σ leads all normal distribution functions to have the same width. Using $u = (x-\mu)/\sigma$ leads to $f(u)$ being the *standard normal distribution function*. The variable u is called the *standard variable*:

$$u = \frac{x-\mu}{\sigma} \tag{3.6a}$$

The value of the integral

$$\int_{-\infty}^{\infty} \exp\left(-u^2/2\right)\,du$$

can be found by the following method. The function is symmetrical about zero and hence the area under the function from $-\infty$ to 0 is equal to the area from 0 to ∞, hence $\int_{-\infty}^{\infty} = 2\int_0^{\infty}$

Let $I = \int_0^{\infty} \exp\left(-x^2/2\right)\,dx$, and by using the same notation for another variable, $I = \int_0^{\infty} \exp\left(-y^2\right)\,dy$, multiplication of the two integrals yields:

$$I^2 = \int_0^{\infty} \exp\left(-x^2/2\right)\,dx \times \int_0^{\infty} \exp\left(-y^2/2\right)\,dy$$

As x and y are independent variables and since $\exp\left(-x^2/2\right) \times \exp\left(-y^2/2\right) = \exp\left[-\left(x^2+y^2\right)/2\right]$

$$I^2 = \int_0^{\infty}\int_0^{\infty} \exp\left[-\left(x^2 + y^2\right)/2\right]\, \mathrm{d}x\, \mathrm{d}y$$

Let us change the cartesian coordinates to polar ones. In polar coordinates, $x^2 + y^2 = r^2$. The range of the variables x and y from 0 to ∞ means the first quadrant and hence r is going from 0 to infinity and θ from 0 to $\pi/2$. In polar coordinates, $\mathrm{d}x\ \mathrm{d}y = r\ \mathrm{d}r\ \mathrm{d}\theta$, and hence:

$$I^2 = \int_0^{\pi/2}\int_0^{\infty} \exp\left(-r^2/2\right) r\ \mathrm{d}r\ \mathrm{d}\theta$$

Calculating the inner integral yields:

$$\int_0^{\infty} \exp\left(-r^2/2\right) r\ \mathrm{d}r = \left[-\exp\left(-r^2/2\right)\right]_0^{\infty} = -(0 - 1) = 1$$

Hence:

$$I^2 = \int_0^{\pi/2} \mathrm{d}\theta = \frac{\pi}{2} \quad \Rightarrow \quad I = \frac{1}{2}\sqrt{(2\pi)}$$

From this integral we can find the normalization factor of the distribution function:

$$\int_{-\infty}^{\infty} \mathrm{f}(u)\ \mathrm{d}u = 2A\ I \quad \Rightarrow \quad \int_{-\infty}^{\infty} \mathrm{f}(u)\ \mathrm{d}u = A\ \sqrt{(2\pi)} = 1$$

The normalization constant is $A = 1/\sqrt{(2\pi)}$ and the *standard normal distribution function* is:

$$\mathrm{f}(u) = \frac{1}{\sqrt{(2\pi)}}\ \exp\left(-u^2/2\right) \tag{3.7}$$

When changing from the standard variable u to the original variable x, the normalization constant also changes since $u = (x - \mu)/\sigma$ leads to $\mathrm{d}x = \sigma\ \mathrm{d}u$ and hence

$$\int_{-\infty}^{\infty} \mathrm{f}(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{f}(u)\ \sigma\ \mathrm{d}u = \sigma \int_{-\infty}^{\infty} \mathrm{f}(u)\ \mathrm{d}u = \sigma\ \sqrt{(2\pi)}$$

So the normalization constant is $1/\left(\sigma\ \sqrt{2\pi}\right)$.

The *normal distribution function* for the variable x is:

$$\mathrm{f}(x) = \frac{1}{\sigma\ \sqrt{(2\pi)}} \exp\left[-(x - \mu)^2/2\sigma^2\right] = \frac{1}{\sigma\ \sqrt{(2\pi)}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right] \tag{3.8}$$

It can be proven that for $\mathrm{f}(u)$ the mean is zero and the variance is equal to 1.

$$\mu(u) = \int_{-\infty}^{\infty} u\, f(u)\, du = \int_{-\infty}^{\infty} u\, \frac{\exp(-u^2/2)}{\sqrt{(2\pi)}}\, du = \frac{1}{\sqrt{(2\pi)}} \left[-\exp\left(\frac{-u^2}{2}\right) \right]_{-\infty}^{\infty}$$

$$= \frac{1}{\sqrt{(2\pi)}}(0 - 0) = 0$$

Using Equation (3.5):

$$\sigma^2 = \int_{-\infty}^{\infty} u^2\, f(u)\, dx - \mu^2 = 1 - 0 \quad \Rightarrow \quad \sigma^2(u) = 1$$

For $f(x)$, the mean is μ and the variance is equal to σ^2. The function $f(x)$ is a probability function, since $\int_{-\infty}^{\infty} f(x)\, dx = 1$, which means that $f(x)\, dx$ is the probability in a random experiment to receive a result which lies in the range x and $(x + dx)$.

The cumulative probability function $\phi(x)$ is defined as the integral of the probability function from $-\infty$ to x:

$$\phi(x) = \int_{-\infty}^{x} f(y)\, dy \tag{3.9}$$

For the normal distribution, the integral $\int_{-\infty}^{x} \exp(-y^2/2)\, dy$ cannot be calculated analytically, however it can be calculated numerically and has been tabulated. Table 3.1 gives the function $\phi(z)$, where z is the standard variable:

$$z = \frac{x - \mu}{\sigma} \tag{3.6b}$$

In using this table for calculating probabilities, it should always be remembered that *the variable should be first transformed to the standard variable*. Some examples of calculations using this table follows.

Given a population where the monthly expenditures of a family are *distributed normally*, with a mean of \$2000 and a standard deviation of \$800, calculate the percentage of the families who:

(**1**) Spend monthly less than \$3000.
(**2**) Spend monthly more than \$3500.
(**3**) Spend monthly less than \$1000.
(**4**) Spend monthly between \$1500 and \$2800.

Solution: The first step is to transform the nominal variable (x) to the standard one (z), using Equation (3.6):

$$x = 3000 \quad \Rightarrow \quad z = \frac{3000 - 2000}{800} = 1.25$$

$$x = 3500 \quad \Rightarrow \quad z = 1.875$$

$$x = 1000 \quad \Rightarrow \quad z = \frac{1000 - 2000}{800} = -1.25$$

$$x = 1500 \quad \Rightarrow \quad z = -0.625 \qquad x = 2800 \quad \Rightarrow \quad z = 1$$

$\phi(z)$ is found from Table 3.1.

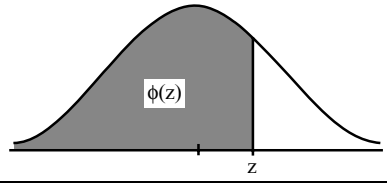

Table 3.1 Cumulative distribution function for the standard normal distribution.

z	0	1	2	3	4	5	6	7	8	9
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.568	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.625	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.692	0.695	0.699	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.806	0.809	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.832	0.834	0.837	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.866	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.902
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.935	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9650	0.9657	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9954	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Family (1): For $z = 1.25$, we look at row 1.2 and then at the number in column 5 (which is our third digit) and find 0.894. In some statistics tables the decimal point appears only in the column headed by zero, so in the column headed by 5 we would find 894, but we have to write 0. to the left of the number found as the probability is always less than one, i.e. 0.894. Finding $\phi(1.25) = 0.894$ means that the integral $\int_{-\infty}^{1.25} f(z)\,dz = 0.894$. It means that 0.894 from all the population lies in the range $[-\infty,\ 1.25]$. Hence 89.4% ($= 0.894 \times 100$) spends monthly less than \$3000.

Family (2): $z = 1.875$ cannot be found directly in Table 3.1, since the table gives only $\phi(1.87)$ and $\phi(1.88)$ in row 1.8 in the columns headed by 7 and 8. Since $\phi(1.87) = 0.9693$ and $\phi(1.88) = 0.9699$, then by interpolation $\phi(1.875) = 0.9696$.

Thus 96.96% ($= 0.9696 \times 100$) spends less than \$3500 and hence 0.04% ($100 - 96.96$) of the population spends monthly more than \$3500.

Notice that directly from the table we get only 'less than'. In order to obtain 'more than', the tabulated value has to be subtracted from the whole (1.0 in probability or 100% in population) population.

Family (3): For negative values of z we cannot find $\phi(z)$ in Table 3.1. Some books give tables which include negative values of z, but most books give a table with $\phi(z)$ only for positive values of z. In this case, $\phi(z)$ for negative values of z is calculated by the relation:

$$\phi(-z) = 1 - \phi(z) \tag{3.10}$$

This equation is explained by Fig. 3.1, where z has a positive value. Since the probability function is symmetrical about the zero, Equation (3.10) must hold.

We found earlier that $f(1.25) = 0.894$ and thus $\phi(-1.25) = 1 - 0.894 = 0.106$. Hence 0.106 or 10.6 % of the population spends less than \$1000.

When we are interested in the probability of having an expenditure between z_1 and z_2 we subtract the cumulative probability of those values (Fig. 3.2).

Hence the fraction of the population spending between \$1500 and \$2800 is:

$$\phi(1) - \phi(-0.625) = 0.841 - (1 - 0.734) = 0.841 - 0.266 = 0.575$$

i.e. 57.5 % of the population spend in this range.

In some statistics books, instead of tabulating the cumulative probability, tabulations of the normal curve area are given which take the lower boundary of the integral as 0 and not as $-\infty$, as can be seen in Table 3.2. In most cases the table gives a figure to indicate the boundaries of the integration. It is advisable to check if $\phi(0) = 0.5$, which means that the lower boundary is $-\infty$, or $\phi(0) = 0$, which means that the lower boundary is 0.

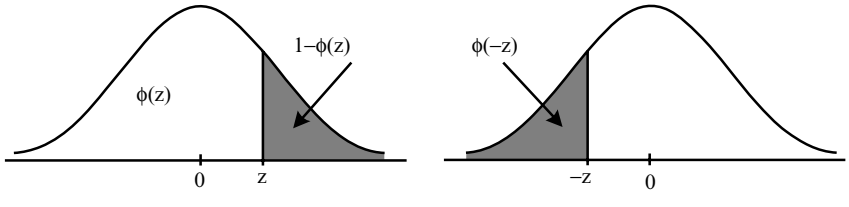

Fig. 3.1

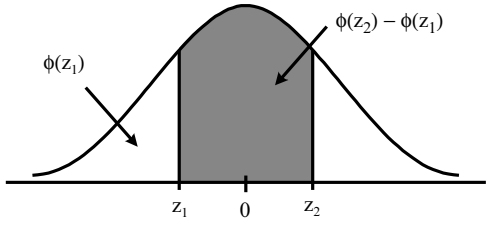

Fig. 3.2

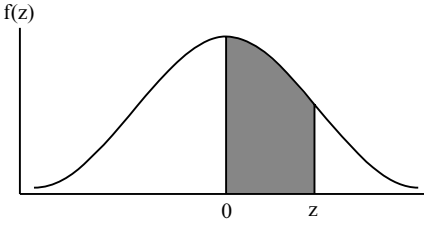

Table 3.2 Normal curve areas.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2707	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3437	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Not all random variables are distributed normally. There are other probability distribution functions. Many of them are derived from the *standard normal distribution* function which is called N(0, 1) i.e. normal distribution with mean $= \mu = 0$ and variance $= \sigma^2 = 1$. The most common distributions, other the normal one, are the chi-square (χ^2) distribution, the t (student's) distribution and the F distribution.

3.4.2 *Chi-square* (χ^2) *distribution*

The chi-square distribution function is derived from the normal distribution in the following way.