UNIT 3 EVALUATION OF ANALYTICAL DATA II

Structure

3.1 INTRODUCTION

A scientist always worries about the quality of analytical results. Whether these results are obtained directly by him or from someone else, he always thinks that the result should be good enough to use as a basis for action and should be of sufficiently good quality as need be. Any laboratory that performs analysis providing a basis for utility must have a solid quality assurance programme. The aspects which can answer the quality assurance are the assessing of analytical data. When assessing certain analytical data we are generally interested most in learning that upto what extent the results are reliable or how far they agree with the actual content of the component analyzed. Later on we wish to know whether statistically the same reliability is achieved in the analyses of different samples.

Statistically calculations are necessary to understand the significance of the data that are collected and, therefore, to set limitations on each step of analysis. The role of statistical calculations is to sharpen the analysts judgment concerning the effects of indeterminate errors. We can use statistical methods to evaluate the random errors which follow a normal distribution or Gaussian distribution. Advances in theoretical statistical methods and their application to industrial problems have given many

answers in a logical manner. The behaviour of most industrial plants is subject to variations caused due to multiple effects. That is, the individual results are subject to chance variations and in order to draw any worthwhile conclusions it is necessary to examine a set of data with a proper statistical approach. As more and more results are available the accuracy of estimation improves.

In this unit we shall examine the methods used by scientists in evaluating the significance of analytical data with the knowledge of normal distribution of errors in terms of probability. We shall be able to make use of the theory of probability to express the reliability of our data. We shall also see how the precision of measurements can be determined, how different sets of data may be compared with the help of different tests of significance. We shall also examine how to set the relation between dependent and independent variables of measurements through regression analysis and least square method. Finally we shall learn how to plot the data on a quality control chart.

Objectives

After studying this unit, you should be able to

- define some common terms of statistical calculations,
- understand the normal error curve,
- estimate the precision of analytical data,
- get a guideline concerning whether or not an outlying value in a set of replicate results should be retained or rejected,
- estimate whether the difference in two sets of data in experimental results is just by chance or there is some source of systematic errors in one of the sets,
- calculate the confidence limits and confidence interval, and
- analyse the data with the help of control charts.

3.2 SOME IMPORTANT TERMS

You have experienced that a single measurement can not be taken as an accurate result. Our confidence in an analytical result is increased by increasing the number of parallel determinations, known as *replicate determinations.* Determination of the number of times a measurement should be replicated in order to approach the value of experimental mean around the true mean with a certain degree of probability. That is, the more numerous the number of observations the more their results approach the truth. The replications are useful to us in two ways. First, a reliable central value of the set of analytical data should be evaluated. The mean is the most commonly used measure of the central value and the less common used measures are the median and the mode. Second, an analysis of the variation in results helps us to estimate the uncertainty associated with the central value of the data. You should note that a population is the collection of all measurements (very large number and to infinity to the analyst, while a sample is the subset of these measurements (finite number of measurements) selected from the population and we also call it as finite sample. Statisticians use Greek letters (such as μ, σ) for the population parameters, whereas English letters (such as $\bar{x}$, s) for finite samples known as statistics. The dual set of symbols is a valuable aid in discussing experimental data.

3.2.1 Mean

The *mean* is the most widely used measure of the central value. For a finite sample

(for n < 30) the mean known as *sample mean* is represented by $\bar{x}$ and is the arithmetic average of all the observations in the set of data;

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \frac{\sum x_i}{n} \quad \ldots (3.1)$$

where, x_1, x_2…, x_n are the replicate observations, x_i represents the individual value of x making up the set of n number of observations, the symbol $\sum x_i$ means summation of x individual values from i = 1 to i = n, i.e. to add up all of the individual values of x in the set of replicate analyses.

For the entire population of data or universe of data (the number of observations approaching infinity i.e. $N \rightarrow \infty$) the mean known as *population mean* is represented by μ and is given by

$$\mu = \frac{x_1 + x_2 + x_3 + ... x_N}{N} = \frac{\sum x_i}{N} \quad \ldots . (3.2)$$

where N has been used to denote the large number of observations. We will see later, after studying the probability distribution of data that the population mean μ is the most probable value and is taken to be the *true value* of the measured quantity.

3.2.2 Median

When quick measure of central value is to be decided and when gross errors are suspected the central tendency of a group of results can be expressed in terms of *median* by arranging the observations either in ascending or descending sequence. Median means the middle value. That is, there are equal numbers of results that are larger and smaller than the median. For an odd number of total observations (n is odd), the median is obtained directly as middle value. For a set of even numbered total observations (n is even), the median is the average of the middle pair of observations.

The median is a less efficient measure of central tendency than is the mean, but in some cases it may give an easy look of central tendency in the sequential arranged data particularly in dealing with small samples. Thus, for small number of observations, the median may be a better estimate of the central value (the true value). Statistically it can be shown that the median of 10 observations is as efficient conveying the information as is the mean from 7 observations.

The median is used advantageously when a set of analytical data contains a probable outlying result, a result that differs significantly from others in the set. An outlying result does not affect the median value since the outlying result lies on the extremes. On the other hand, an outlying result can have a significant effect on the mean of the set since it is included in the calculation of the mean.

3.2.3 Mode

The observation which occurs most frequently (i.e. with maximum frequency) in a series of observations is known as *mode*. It is yet another quick measure of central value if the number of observations is not too small. For example, the mode of the set of data: 12,6, 12,7, 12.9, 12,7, 12.6, 12.8, 13.0, 12.5, 12.6, the value 12.6 is the *mode* since this is occurring with maximum frequency (four times).

3.2.4 Deviation

The error of a measurement can not be stated if the true value of the quantity is not known. It is meaningful then to take the difference between a particular measured

value (observation) and the arithmetical mean of a series of measurements and this difference is called as its *deviation for apparent error*.
A deviation is generally taken without regard to sign. It is defined mathematically as,

$$d = |x - \bar{x}| \ldots \text{ or } D = |x - \mu| \quad \ldots (3.3)$$

where, d is the deviation of the observation x of a finite sample from its mean $\bar{x}$, D is the deviation of an individual measurement from the population mean μ, and $| |$ denotes that the difference is taken as absolute. The reproducibility of measurements is expressed in terms of various types of deviations.

3.2.5 Average Deviation

The average deviation (a.d.) or the mean deviation is the average of individual deviations:

$$\text{a.d.} = \frac{d_1 + d_2 + d_3 + \ldots + d_n}{n} = \frac{\sum d_i}{n} = \frac{\sum |x_i - \bar{x}|}{n} \quad \ldots (3.4)$$

where the symbols have their usual meanings.

The ratio of the average deviation to the mean is known as *Relative Average Deviation* (RAD) which can be expressed as *percent average deviation* when multiplied by 100.

$$\text{RAD} = \frac{\text{a.d.}}{\bar{x}} \quad \ldots (3.4)$$

$$\% \text{ a.d.} = \frac{\text{a.d.}}{\bar{x}} \times 100 \quad \ldots (3.5)$$

Historically the average deviation has been widely employed as the estimate of precision. However, it suffers from the disadvantage that the estimate of this statistics depends upon the number of measurements. The larger the number the better will be the estimate.

3.2.6 Probable Deviation

The probable deviation P is an older measure of precision and now only rarely used. It is defined as the deviation having a magnitude such that there are equal numbers of deviations greater and smaller than itself. In a set of large number of observations it is also known as *probable error*.

3.2.7 Range

The difference between the largest and smallest values in a set of measurements is known as the range. It tells the spread of data. The range is often used, with appropriate factors that depend on the number of measurements, as a quick statistics to a rough estimate of precision.

SAQ 1

Calculate the median for the data: 14.1, 13.8, 14.3, 13.6, 13.4 & 13.5.

……...

……...

……...

……...

……...

3.3 STANDARD DEVIATION

Standard deviation is the most important statistic to indicate the precision of an analysis. According to the International Union of Pure and Applied Chemistry (IUPAC) the symbol σ is used for population standard deviation and the symbol s is used for sample standard deviation.

When the number of observations is very large (N $\rightarrow\infty$) the standard deviation known as *population standard deviation*, σ, which is used to express the precision of a population of data is given by the square root of the average of squares of deviations, thus,

$$\sigma = \sqrt{\frac{D_1^2 + D_2^2 + D_3^3 + ... + D_n^2}{N}} = \left[\frac{\sum D_i^2}{N}\right]^{1/2} = \left[\frac{\Sigma(x_i - \mu)^2}{N}\right]^{1/2} \quad .... (3.6)$$

where, x_i represents the individual observations, D_i the individual deviations, μ the population mean, N the number of observations, and the symbol ∑ denotes the summation for $i = 1$ to $i = N$ values.

For most of the cases in analytical chemistry a finite sample is considered where number of observations is finite (n < 30) and for finite sample the standard deviation known as *sample standard deviation*, s, is the square root of summation of squares of deviations divided by (n – 1) and is given by

$$s = \sqrt{\frac{d_1^2 + d_2^2 + d_3^3 + ... + dn^2}{n-1}} = \left[\frac{\Sigma d_i^2}{n-1}\right]^{1/2} = \left[\frac{\Sigma(x_i - \bar{x})^2}{n-1}\right]^{1/2} \quad ... (3.7)$$

Where (n – 1) is known as the degrees of freedom and other terms and symbols have their usual meaning. Note the divisor (n–1) is used to calculate s rather than the divisor N used to calculateσ. The value of s is only an estimate of σ and will more nearly approach σ as the number of measurements in an analysis increases.

Now let us see why the divisor (n – 1) is used in place of N. Obviously, for large values of N, it is immaterial whether N or (N – 1) is used, but for small number of observations n, which we take practically in analytical chemistry, the distinction is important and we should evaluate this.

At this stage only a qualitative approach will be considered to use (n – 1) as divisor. For a finite number of observations with a sample mean $\bar{x}$ there are n individual deviations (x_i – $\bar{x}$). If they are all added with regard to sign (both positive and negative) the sum of all deviations will come out to be zero. Therefore, only (n – 1) of the deviations are necessary to define n^{th} deviation and it removes one degree of freedom. That is, only (n – 1) independently variable deviations or (n – 1) degrees of freedom (υ) are necessary. We say that the divisor, (n – 1), is a reminder that there are only (n – 1) independent deviations from the mean.

To illustrate the calculation of standard deviation let us take an example.

Example 3.1

In an iron determination (taking 1 g sample every time) the following four replicate results were obtained: 29.8, 30.2, 28.6 and 29.7 mg iron. Calculate the standard deviation of the given data.

Solution

x_i (mg)	$\lvert(Xi - \bar{x})\rvert$	$(Xi - \bar{x})^2$	x_i^2
29.8	0.2	0.04	888.04
30.2	0.6	0.36	912.04
28.6	1.0	1.00	817.96
29.7	0.1	0.01	882.09
$\sum x_i = 118.3$		$\sum(x_i - \bar{x})^2 = 1.41$	$\sum(x_i)^2 = 3500.13$

$$\bar{x} = \frac{118.3}{4} = 29.6$$

i) Using Eq. (3.7) $s = \sqrt{\frac{\Sigma(Xi - \bar{x})^2}{n-1}} = \sqrt{\frac{141}{4-1}}$

Or $s = \sqrt{\frac{141}{3}} = \sqrt{0.47} = 0.69$ mg

3.3.1 Standard Deviation of the Mean

The standard deviation calculated as above by Eq. (3.7) or (3.20) refers to scatter in terms of standard deviation for a single measurement. We know that the arithmetic mean of a series of n measurements is more reliable (precise) than an individual observation. It can be shown statistically that the mean of n results is $\sqrt{n}$ times as reliable as any one of the individual results. Thus, the mean of 4 results is twice as reliable as a single measurement; likewise, the mean of 16 results is 4 times more reliable than any of the individual results etc. The precision is expressed in terms of deviation and less the deviation the more precise the result is. In other words, the deviation in the mean of a series of 4 observations is one-half that of a single observation, and the deviation in the mean of a series of 16 observations is one-fourth that of a single observation. That is, the deviation of the mean of series of n measurements is inversely proportional to the square root of n of the deviation of the individual values. Thus

$$\text{Deviation of the mean} = d_{mean} = \frac{d}{\sqrt{n}} \quad \ldots (3.9)$$

And the standard deviation of the mean (s_{mean}) is inversely proportional to the square root of n of the standard deviation of the individual values (s).

$$s_{mean} = \frac{s}{\sqrt{n}} \quad \ldots (3.10)$$

The standard deviation of the mean is sometimes referred to as the Standard Error.

3.3.2 Variance (V)

The term that is sometimes useful in statistics is the variance (V). This is the square of the standard deviation. The sample variance is given by

$$V_{sample} = s^2 = \frac{\sum d_i^2}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1} \quad \ldots (3.11)$$

And is an estimate of population variance σ^2,

$$V_{population} = \sigma^2, = \frac{\sum(x_i - \mu)^2}{N} \quad \ldots (3.12)$$

The variance is a less commonly used measure of precision mainly because of the drawback that it has the units of data squared. The standard deviation having the same units as the data is preferably employed as a measure of precision because it is easier to relate a measurement and its precision if they both have the same units. However, the advantage of using variance is that variances are added in many cases. We shall use this in determining the propagation of error.

The ratio of two variances is known as F function, that is $F = V_1/V_2$ where $V_1 > V_2$. The F function is used to compare the precision of two sets of measurements (section 3.9).

3.4 RELATIVE STANDARD DEVIATION AND COEFFICIENT OF VARIATION

It is worthwhile to quote standard deviation in relative terms rather than absolute. In doing so we can get a clear picture of data quality as compared to that by absolute standard deviation. The relative standard deviation, RSD, is calculated by dividing the standard deviation by the mean value of the analytical data. It is often denoted by the symbol sr and can also be expressed in suitable units as desired.

$$\text{RSD} = \text{sr} = \frac{s}{\bar{x}} \qquad \ldots (3.13)$$

Expressed in parts per thousand is,

$$\text{RSD} = \frac{s}{\bar{x}} \times 1000 \text{ ppt} \qquad \ldots (3.14)$$

Or in parts per million is,

$$\text{RSD (in ppm)} = \frac{s}{\bar{x}} \times 10^6 \text{ ppm.} \qquad \ldots (3.15)$$

In a special case (see below) when it is expressed as % RSD, it is known as the *coefficient of variation* (CV).

Often the analytical error is reasonably constant over the useful working range for the procedure. But problem arises when the analytical error depends on the quantity of the element present. For such observations sometimes the range of concentration can be divided into intervals and the standard deviation given for each interval. When the standard deviation turns out to be approximately proportional to the amount of the element present, it is meaningful to give precision compactly in percent by using the quantity coefficient of variation that is the standard deviation expressed as the percentage of the mean.

$$\text{CV} = \frac{s}{\bar{x}} \times 100 \qquad \ldots (3.16)$$

SAQ 2

Calculate the coefficient of variation and relative standard deviation in ppm for the data given in Example 3.1.

……...

……...

……...

……...

3.5 PRECISION OF COMPUTED RESULTS

The calculated result for certain analysis requires data from two or more independent set of analytical measurements, each of which has a random uncertainty and each of which contributes to the net precision of the final result. This is also known as *propagation of error* or *standard deviation* of calculated results. How such random uncertainties affect the net outcome of an analysis will be discussed in this section. Let us assume that the final result W is dependent on the experimental variables: a, b, c…, each of which fluctuates in a random and independent way. For this we can say that W is a function of a, b, c …, or

$$W = f(a, b, c \ldots\ldots) \quad \ldots (3.17)$$

Since the precision is best measured in terms of standard deviation, we have to calculate the standard deviation of a result which has been obtained by the calculation of two or more analytical data points, each of which has a separate standard deviation, we shall consider here at this level only two types of arithmetic operations: (i) the addition and subtraction, and (ii) the multiplication and division.

3.5.1 Addition and Subtraction

The absolute standard deviation of a sum or difference is calculated by taking the square root of the sum of the squares of the individual absolute standard deviation. Consider the situation where a dependent quantity W is calculated from three quantities a, b and c in terms of sum and difference as follows:

$$W = a + b - C \quad \ldots (3.18)$$

Let us assume that the standard deviations of these quantities are s, s_a, s_b and s_c respectively. Since the variance of a sum or difference is equal to the sum of the individual variances, we have variance of W, s_w^2 as given by

$$s_w^2 = s_a^2 + s_b^2 + s_c^2 \quad \ldots (3.19)$$

Hence, the absolute standard deviation of a sum or difference is equal to the square root of the sum of the squares of the absolute standard deviations of the individual quantities irrespective of the quantity W being sum or difference of variables a, b and c. To illustrate let us consider the following example.

Example 3.2

Three quantities are to be summed up as y = a + b – c. The individual absolute standard deviations of the three quantities are given in parentheses. Calculate the standard deviation of the arithmetic operation and express the result of summation.

a = 2.50 (± 0.02), b = 3.10 (±0.03), c = 1.97 (±0.05)

Solution

y = 2.50 + 3.10 – 1.97 = 3.63

$$sy = \sqrt{s_a^2 + s_b^2 + s_c^2}$$

$$= \sqrt{(4 + 9 + 25) \times 10^{-4}}$$

$$= \sqrt{38} \times 10^{-2}$$

$$= 0.06$$

Thus, the standard deviation of the summation is 0.06, and the result should be represented as

$$y = 3.63 (\pm 0.06)$$

3.5.2 Multiplication and Division

In multiplication and division the precision is calculated in terms of relative standard deviation. Thus, the relative standard deviation of a product or quotient is determined by the relative standard deviations of the numbers forming the calculated result according to the following rule:

"For the products and quotients, the relative standard deviation of the result is equal to the square root of the sum of the squares of the relative standard deviations of the numbers making up the product or quotient."

Consider the case where the quantity W is calculated from three quantities a, b and c as $W = \frac{ab}{c}$. Let us assume that the relative standard deviations of these quantities are $\frac{s_w}{w}$, $\frac{s_a}{a}$. $\frac{s_b}{b}$ and $\frac{s_c}{c}$ respectively. According to the rule given above the relative standard deviation of the result is given by

$$\frac{s_w}{w} = \sqrt{\left(\frac{s_a}{a}\right)^2 + \left(\frac{s_b}{b}\right)^2 + \left(\frac{s_c}{c}\right)^2} \qquad \ldots (3.20)$$

Note that from relative standard deviation RSD the standard deviation s can be calculated as

$$\frac{s_w}{w} = \text{RSD, or } s_w = \text{RSD} \times W.$$

To understand the rule let us take an example.

Example 3.3

The result of three quantities a, b and c is to be calculated as $y = (a \times b)/c$. The individual standard deviations of each quantity are given in parenthesis: a = 50.23 (±0.07), b = 27.86 (± 0.05) and c = 0.1167 (± 0.0003). Calculate the standard deviation of the operation and express the calculated result with absolute uncertainty.

Solution

$$y = \frac{50.23\ (\pm\ 0.07)\ \times\ 27.86\ \ (\pm\ 0.05)}{0.1167\ (\pm\ 0.0003)}$$

First let us calculate the result without standard deviations,

$$\frac{50.23 \times 27.86}{0.1167} = 11991.497 = 11991$$

Let us now calculate, the relative standard deviation of y according to the rule of multiplication and division,

$$\frac{s_y}{y} = \sqrt{\left(\frac{0.07}{50.23}\right)^2 + \left(\frac{0.05}{27.86}\right)^2 + \left(\frac{0.0003}{0.1167}\right)^2}$$

$$\frac{s_y}{11991} = \sqrt{\left[1.9 \times 10^{-6} + 3.2 \times 10^{-6} + 6.6 \times 10^{-6}\right]}$$

$$\frac{s_y}{11991} = \sqrt{11.7 \times 10^{-6}} = 3.42 \times 10^{-3}$$

Or $s_y = 3.42 \times 10^{-3} \times 11,991 = 41$

The result is expressed as

$y = 11991 \pm 41$

3.6 GAUSSIAN DISTRIBUTION OF DATA

You learnt in section 3.3 that the precision of a given set of measurements can be ascertained from the value of standard deviation. However, this term does not indicate how the replicate results are distributed. Probability distributions are of fundamental importance to the use of statistics for judging the reliability of analytical data. You learnt in the Unit 2 that even after taking all the precautions to avoid systematic errors, the replicate results vary with each other by small to large extent. This variation in results is due to indeterminate or random errors. The word *error* is used somewhat loosely while talking for indeterminate error. Strictly we should speak of *deviation*, since the true value is not known. However, as we shall see below, in general convention the use of term error for indeterminate error should not create any confusion. The indeterminate errors can not be eliminated and are often the major source of uncertainty in a determination. The combined effect of the individual uncertainties, usually, causes replicate measurements to fluctuate randomly around the mean of the set.

What we say is all measurements contain random errors. The elimination of these errors is beyond the power of the observer. However, an analyst can estimate the uncertainty introduced by random errors. We can apply the laws of probability provided we have sufficient number of observations. These laws of probability or rules of chance in an empirical way are summarized as follows:

i) Small errors occur with high frequency,

ii) Large errors occur with low frequency, and

iii) Positive and negative errors occur with equal frequency, so that the arithmetic mean is the most probable value.

The above three rules of chance are applied for a very large number of observations which is known as the entire population or universe of data. You will see below that the rules can be verified graphically when frequency of errors are plotted as a function of their magnitudes. The plot so obtained is known as **NORMAL ERROR CURVE –** and describes the properties of universe of data. As an explanation consider the discussion given below.

If a large number of replicate readings are taken of a continuous variable, e.g. the percentage of iron in an iron ore, the results obtained will usually be distributed around the mean in a roughly symmetrical manner. When frequencies of observations (on y-axis) are plotted against values of observations (along x-axis), we get the distribution as shown in Fig.3.1 where the dots indicate the frequencies of respective values of observations. Such a distribution of random values (also of random errors) is called the normal frequency distribution. On drawing the boundaries of these frequencies we get a bell shaped curve that is symmetrical around the mean and is known as the normal distribution curve or a Gaussian Distribution Curve (as this follows the normal frequency curve of Gauss). You see that on either side of the curve there is an exponential decrease in frequency as the magnitude of error increases. The inflexion points of the curve are obtained when the result is $\pm 1\ \sigma$ of the mean μ. With this type of distribution about 68% of all values of the universe of data will fall within $\pm 1\ \sigma$ of the mean. The value of σ also indicates the precision. On comparison the values of σ on the curves ($x - \mu = \sigma$) the precision of two sets of data can also be compared. For example if the standard deviation (σ_1) of the first data set is ½ of the value of the standard deviation (σ_2) of the second data set as shown in Fig. 3.2 means that the precision of the first data set is twice than that of the second data set.

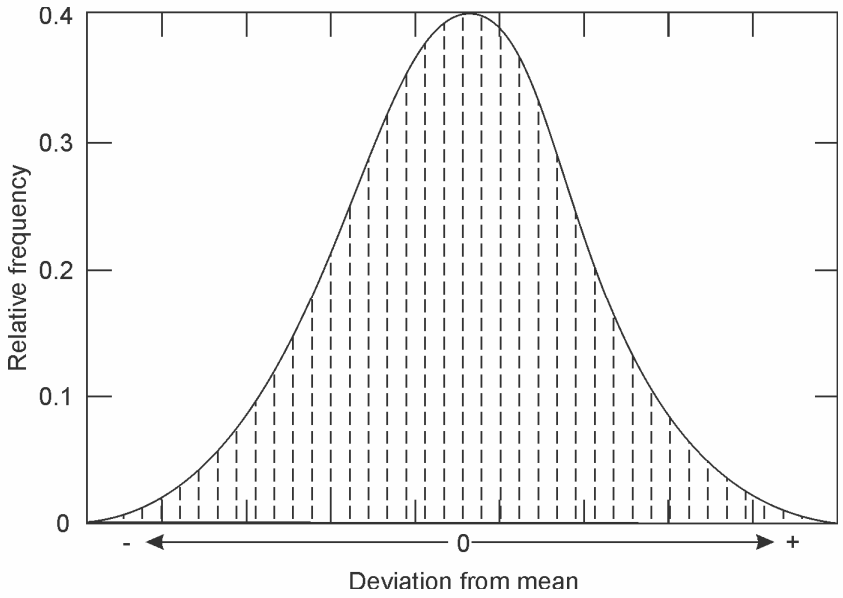

Fig. 3.1: Frequency distribution curve

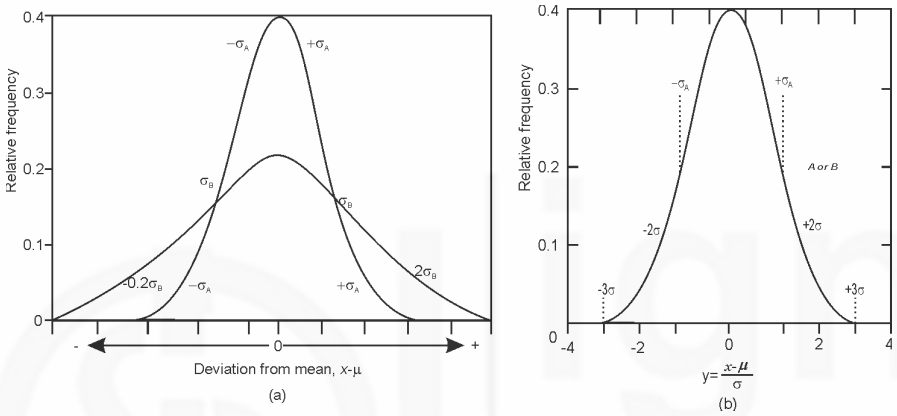

Fig. 3.2: Normal error curves (a) the abscissa is the deviation from the mean in the units of measurements (b) The abscissa is the deviation from the mean in units of σ

Mathematical Form of Probability Distribution

It is frequently possible to find a suitable mathematical model for the probability distribution of random errors. The distribution of random errors follows the normal frequency curve (of Gauss) and can be expressed by a differential Eq.,

$$\frac{dN}{N} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx \qquad \ldots (3.21)$$

where x the value of observation varies from $-\infty$ to $+\infty$, N stands for total number of observations, $\frac{dN}{N}$ is the fraction (i.e. frequency of the population with values in the interval x to $(x + dx)$; proportional, therefore, to the probability that a given value falls in this interval, μ is the average value of the entire population, σ is the standard deviation of the population. The differential Eq. is given by the two population parameters μ and σ. We can follow that $\left(\frac{x-\mu}{\sigma}\right)$ gives the deviation of an observation x from population mean μ in terms of σ and if we express this deviation by a variable y, it is possible to express the normal law of error Eq. (3.34) in terms of a single variable Eq. Thus,

$$\frac{|x-\mu|}{\sigma} = y \qquad \ldots (3.22)$$

Differentiating: $\frac{1}{\sigma} dx = dy$ … (3.23)

Substituting from Eqs. (3.35) and (3.36) into Eq. (3.34), we get the Eq.,

$$\frac{dN}{N} = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)\, dy = f(y)\, dy \qquad \ldots (3.24)$$

Which is plotted in Fig. 3.3.

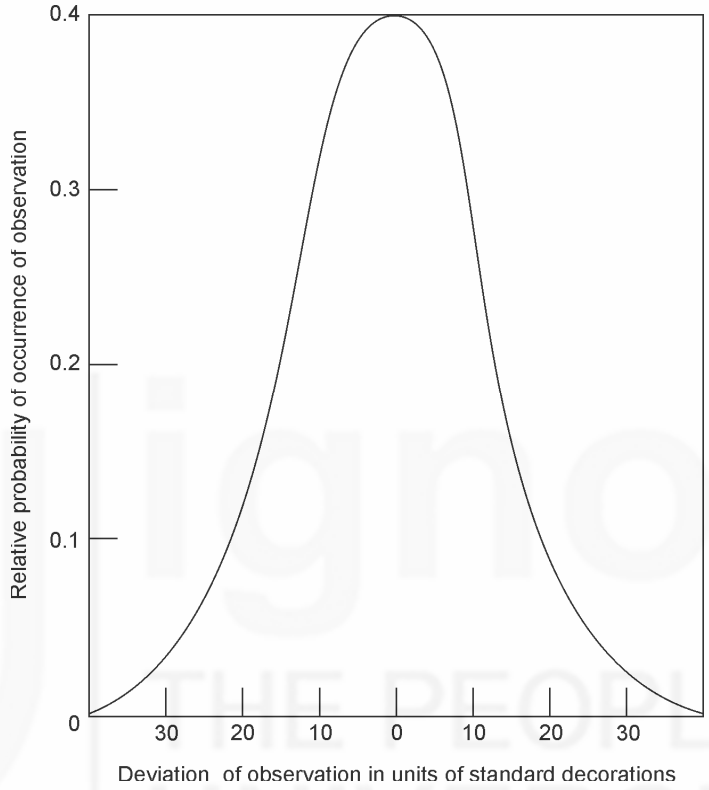

Fig. 3.3: Normal Error curve

From the differential Eq. (3.24) we can see that the maximum of f(y) will be obtained when y is equal to zero. That is, the probability of occurrence of observation being maximum for $y = (x - \mu)/\sigma = 0$, that is for $x = \mu$. *It shows that the average value is the most probable value of the population of data.* The total area under the curve of Fig. 3.3 can be obtained by integrating the Eq. (3.22) within the limits $-\infty$ to $+\infty$, which is

$$\frac{1}{2\sqrt{\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{y^2}{2}\right) dy = 1 \qquad \ldots (3.25)$$

Corresponding to a total probability of unity for the total number of observations lying from $-\infty$ to $+\infty$ (entire population). This also indicates that the *area* gives the fraction of the total population *(dN/N)* having magnitudes of y between these two values. In Table 3.1 are listed values of $\pm y$ and the deviation $|x - \mu|$ beyond which lie various fractions of the total area under the curve 3.3 to indicate the probabilities greater and smaller values than this deviation.

Table 3.1: Values of Error and Probability

Error $\|x - \mu\|$	$\pm y$	Probability of greater value	Probability of smaller value
0.674σ	0.674	0.50	0.50
1.00σ	1.00	0.317	0.683
1.96σ	1.96	0.05	0.95
2.00σ	2.00	0.045	0.955
2.576σ	2.576	0.01	0.99
3.00σ	3.00	0.0026	0.9974
3.29σ	3.29	0.001	0.999

According to the values of Table 3.1 you can see that for a normal distribution of errors the probability of an error greater than 0.674σ is 0.50 and also smaller than 0.674σ is 0.50. It means that 50% observations are such that their magnitude of error is less than $\pm 0.674\sigma$ and for rest 50% observations the magnitude of error is greater than $\pm 0.674\sigma$. The probability of an error smaller than $\pm 1\sigma$ is 0.683 or 68.3% and of an error greater than $\pm 1\sigma$ will be 31.7%. Similarly the probability of an error greater than 2σ will be 4.5% and of an error greater than 3σ is about 0.3% and so on. You follow from above that the area under certain limits in the probability distribution curve is a measure of probability of occurring observations under these limits.

SAQ 3

From the normal error curve find the probability of a result (i) lying between 0 and $+1\sigma$ of the mean μ, (ii) lying between $+1\sigma$ and $+ 2\sigma$ of the mean.

……..

……..

……..

……..

……..

……..

3.7 CONFIDENCE INTERVAL

From the above discussion you understand that mostly in chemical analysis, the true value of the population mean μ can not be determined because a very large number of measurements (approaching infinite number) would be required to calculate it. However, with the help of statistics a range can be established surrounding sample mean $\bar{x}$ within which the population mean μ is expected to lie with a certain degree of confidence based on probability distribution. Thus, "the range (or a numerical interval) around the mean $\bar{x}$ of a set of replicate analytical results within which the population mean μ can be expected to lie, with a certain degree of confidence (i.e., with a certain probability), is known as **Confidence Interval**. "The boundaries of this

range are called the **Confidence Limits**. "The likelihood that the true value falls within the range is called the **Probability or Confidence Level** and it is often expressed as a percentage" consider the example that in a set of iron determination it is 95% probable that the population mean μ lies in the limit ± 0.20% Fe of sample mean $\bar{x}$ = 11.30% Fe. It tells that the confidence interval is 11.10% to 11.50% Fe with 95% probability. The 95% confidence limit for μ = 11.30% ±0.20% Fe, and the confidence level is 95%.

The confidence interval is related to the standard deviation of the mean and its size depends on how well the sample standard deviation s estimates the population standard deviation σ. If s is a closer estimate of σ, the narrower is the confidence interval. To give confidence interval at the high confidence levels is one of the best ways of indicating reliability. We shall discuss below two cases for finding the confidence interval: (A) when σ is known or s is a good approximation of σ, and (B) when σ is not known.

3.7.1 Confidence Interval When σ is known or s is a Good Estimate of σ

In section 3.6 you have learned that the normal error curve can be expressed by an Eq. (3.24) in a single variable y where *y* is defined as $\pm y = (x - \mu)/\sigma$ given by Eq. (3.22). From the definition of *y* it follows that $(x - \mu) = \pm y\,\sigma$ with the help of normal error curve and the values of *y* with probability given in Table 3.1 we see that the areas represent probabilities for the absolute deviation $|x - \mu|$ to exceed the value of $y\,\sigma$. Since $y\,\sigma$ is the deviation of a single observation *x* from the population mean μ, we can express the probability in terms of *y* as,

$$\pm y\,\sigma = x - \mu$$

$$\text{Or} \quad \mu = x \pm y\,\sigma \qquad \ldots (3.26)$$

Again from Table 3.1, when y = 0.67 there are 50% chances that an observation will lie in the area having a lower deviation than ± .67 σ, & when y = 1.00 there is a 68.3% probability that a particular measured value has a deviation ± σ and so on. Thus, based on measuring a single value we can write for population mean μ (from Eq. 3.26) lying within limits.

$\mu = 0.67\,\sigma$ with 50% confidence (or 50% probability), … (3.27 – i)

$\mu = x \pm 1\,\sigma$ with 68.3% confidence … (3.27 – ii)

$\mu = x \pm 2\,\sigma$ with 95.5% confidence … (3.27 – iii)

$\mu = x \pm 3\,\sigma$ with 99.7% confidence … (3.27 – iv)

Otherwise in more useful way as

$\mu = x \pm 1.96\,\sigma$ with 95% confidence … (3.27 – v)

$\mu = x \pm 2.58\,\sigma$ with99% confidence … (3.27 – vi)

$\mu = x \pm 3.29\,\sigma$ with 99.9% confidence … (3.27 – vii)

However, it is not advisable to estimate the true mean from a single observation *x*. Instead we generally use the sample mean $\bar{x}$ to take the better estimate of population mean μ. We also know that for mean $\bar{x}$ of *n* observations the standard deviation of the mean is $\frac{1}{\sqrt{n}}$ times of the standard deviation of single observation. Then in terms of $\bar{x}$, the population mean μ lies within the limits.

$\mu = \bar{x} \pm \frac{y\sigma}{\sqrt{n}}$ with the confidence level corresponding to the value y.

That is, confidence interval for $\mu = \bar{x} \pm \frac{y\sigma}{\sqrt{n}}$ or from $\bar{x} - \frac{y\sigma}{\sqrt{n}}$ to $\bar{x} + \frac{y\sigma}{\sqrt{n}}$, and confidence limits can also be expressed with certain confidence level in the same manner.

In the modern practice it is usual to employ a confidence level of 95% for y = 1.96 or 99% for y = 2.58. Thus, we can say that the population mean μ lies within the limits.

$$\mu = \bar{x} \pm 1.96\ \sigma/\sqrt{n} \quad \text{with 95\% confidence} \qquad \ldots (3.28\ a).$$

It means that it is 95% probable that the population mean μ lies in the interval $\bar{x} - 1.96\ \sigma/\sqrt{n}$ to $\bar{x} + 1.96\sigma/\sqrt{n}$. In a similar way the population mean μ lies within the limits.

$$\mu = \bar{x} \pm 2.58\ \sigma/\sqrt{n} \quad \text{with 99\% confidence} \qquad \ldots (3.28\ b)$$

$$\mu = \bar{x} \pm 3.29\ \sigma/\sqrt{n} \quad \text{with 99.9\% confidence} \qquad \ldots (3.28\ c)$$

Consider the following example,

Example 3.4

Measurements of glucose levels in a patient suffering from diabetes gave the following results: 1.108, 1.100, 1.122, 1.088, 1.115, 1.099 and 1.075 g/L. Calculate the 95% confidence interval when σ = 0.019 g/L.

Solution

$$\bar{x} = \frac{1.108+1.100+1.122+1.088+1.115+1.099+1.075}{7}$$

$$= 1.101 \text{ g/L}$$

For 95% confidence the limits are $\pm \frac{1.96\sigma}{\sqrt{7}}$

$$= \pm \frac{1.96 \times 0.019}{2.646}$$

$$= \pm 0.01407$$

$$= \pm 0.014 \text{ g/L}$$

The population mean μ lies within the limits

$\mu = (1.101 \pm 0.014)$ g/L with 95% confidence

Thus, it is 95% probable that the population mean lies in the interval from 1.087 to 1.115 g/L.

3.7.2 Confidence Interval when σ is Not Known

In the usual practice the population standard deviation is not known and can only be approximated for a finite number of measurements by the sample standard deviation s. To overcome this difficulty another method is used, which is based on a statistical factor, "t" (also known as student t), that depends on the number of degrees of freedom and confidence level desired. The quantity t is defined as the difference between the two means divided by its standard deviation:

$$\pm t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} = \frac{(\bar{x} - \mu)\sqrt{n}}{s} \qquad \ldots (3.29)$$

Statisticians have compiled the values of t for the given degrees of freedom ($\nu = n - 1$) and for various confidence levels desired. For illustration some of the values of t are listed in Table 3.2.

Table 3.2: Values of t for υ degrees of freedom and various confidence levels.

υ	Confidence level		
	90%	95%	99%
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.500
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
∞	1.645	1.960	2.576

A simpler and approximate procedure, particularly useful for small numbers of observations (n < 10) to express confidence interval, is based on the range R (the difference between the largest and smallest values of observations). According to this to a certain degree of confidence the population mean μ lies within the limits

$$\mu = \bar{x} + C_n R \qquad \ldots (3.30)$$

where C_n is constant depending upon the number of observations and the desired confidence level. A few values of C_n are given in Table 3.3.

Table 3.3: Values of C_n for sample of n observations

N	Confidence level		
	90%	95%	99%
2	3.196	6.353	31.828
3	0.885	1.304	3.008
4	0.529	0.717	1.316
5	0.388	0.57	0.843
6	0.312	0.399	0.628

Note: All methods based on range should be used with caution, since the outlying results may cause uncertainty.

SAQ 4

A replicate analysis of potassium in blood serum yielded concentration of K^+ in mg/100 mL: 15.30, 15.85, 15.55 and 16.30. Calculate the 90% confidence interval for the set. Assume the value of C_n for 4 observations at 90% level = 0.53.

……...

……...

……...

……...

……...

3.8 CRITERIA FOR REJECTION OF DATA

Sometimes, in a set of analytical data there appears a value which is not fitting in the set as it looks at a wide difference from the rest of the values. Now the question arises how to decide whether to remove out a result which appears out of line with others when there are no known reasons to suspect it? The question is not of much importance if the number of replicate observations is large since a single value will have only a small effect upon the mean. But it is, of course, important when the number of replicate measurements is small, since here the divergent observation has a significant effect on the value of mean, while at the same time there are insufficient data to decide the fate of the suspected result. In a small set of data the decision for rejection or retention by the blind application of statistical test is no doubt an arbitrary decision.

The criteria for rejection of an observation are based on the supposition that an outlier is due to some systematic source of error. If it is not, then it falls within the random error and should be retained. So many authors agree that the question of rejecting one divergent value from a small sample can not satisfactorily be answered. It is unfortunate fact that no universal rule can be invoked to settle the question of retention or rejection. Out of various rules we shall discuss below only two of the more widely recommended criteria for rejection of outlier: one is based on the average deviation and is called the "4d" rule, and the other which is based on the range is called the "Q" test.

3.8.1 The "4d" Rule

This test is based on the fact that 4 times of average deviation which is an estimate of ($4 \times$ a.d. $= 4 \times 0.80\ \sigma$) $= 3.2\ \sigma$ which lies in the probability distribution curve at a confidence level of 99.8% that a value of measurement lies within this limit. To apply this rule there must be at least 4 observations excluding the outlier (preferably 10 to 30 observations) and the following procedure is adopted:

i) Omit the doubtful value and find out the arithmetic mean of the rest.

ii) Find out the average deviation of rest of the values obtained after omitting the doubtful value. Call this as "d" and its four times as "4d" (i.e. 4× a.d.)

iii) Obtain the difference between the doubtful value and the mean calculated in (i) call this difference as z ($z = | xs - \bar{x} |$) where x_s is the suspected value and $\bar{x}$ is the mean.

iv) Criterion: If the difference z is greater than 4 times of average deviation calculated in (ii) the suspected value should be rejected otherwise retain it. That is, If $z > 4d$, then the doubtful value is rejected, and if $z \leq 4d$, then the doubtful value is retained.

3.8.2 The "Q" Test

Another criterion used to check the rejection of suspected result (in a set of 3 – 10 results) is the Q-test which is a simple and widely used statistical test. Q, the rejection quotient, is defined as the ratio of the divergence of the suspected value from its nearest neighbour to the range of the set of measured values. If the value of Q calculated is greater than the value of Q given in the table at the desired confidence level for the given number of observations, the suspected value is rejected. The Q test is applied as follows:

i) Arrange the observations either in the increasing or decreasing order. The lowest or the highest or both may be the doubtful values.

ii) Calculate the range, R = highest values – lowest value

iii) Find the difference between the suspected value and its nearest neighbour. Call this difference as Y.

iv) Calculate the rejection quotient, Q = Y/R and call it as $Q_{calculated}$.

v) Consult the table of Q (Table 3.4) for the given number of observations and call it as $Q_{tabular}$.

vi) Criterion: Compare Q_{calc} with Q_{tab}. If $Q_{calc.} > Q_{tab}$, then the suspected value is rejected.

Note: Since both lowest and highest values may be considered to be the suspected values, it is advisable to apply the test for both. Say in above if the lowest value was taken as the suspected value we have to extend the test as below.

vii) If the lowest value is rejected, then the range R' for remaining values is calculated and if the lowest value is not rejected, then the same range R is used and apply the above procedure steps (ii) to (vi) for the highest value.

viii) Repeat the process further if necessary. Some Q values are given in Table 3.4.

Table 3.4: Values of Rejection Quotient Q at 90% confidence level

Number of observations (n)	Q
3	0.94
4	0.76
5	0.64
6	0.56
7	0.51
8	0.47
9	0.44
10	0.41

The following example will illustrate application of the Q test.

Example 3.5

Apply Q-test to check the rejection of the highest value in the following results:

2.18, 2.19, 2.30, 2.15 and 2.20
R = 2.30 – 2.15 = 0.15
Y = 2.30 – 2.20 = 0.10
Q calc = Y/R = 0.10/0.15 = 0.67

From Table 3.4, for n = 5, $Q_{tab.} = 0.64$ $Q_{calc} > Q_{tab}$, therefore the value 2.30 is rejected at 90% confidence level.

SAQ 5

In replicate determination of iron the following results of percentage of iron were obtained. Should any of the results be rejected?
%Fe: 52.40, 52.47, 52.50, 52.51, and 52.46.

……...
……...
……...
……...

3.9 TESTS OF SIGNIFICANCE

In analytical chemistry we develop new methods of analysis and it is frequently desired to compare the results of a new method with those of an accepted (say from a past experience or a standard, e.g.; from the National Institute of Standards and Technology, NIST) method. The average values obtained from the two methods may show a difference. The question arises whether this difference is due to random fluctuation (indeterminate error) or directional fluctuation (systematic error). The answer is qualified by a degree of certainty involving the method what is known as NULL HYPOTHESIS which considers that there is no significant difference between two sets of data.

In null hypothesis procedure a comparison, of statistical parameters (based on mean or standard deviation, or some other property), is made between two sets of replicate measurements obtained by two different methods, one of them being the test method and the other usually being accepted method. With this comparison the value of the statistical parameter of test of significance is calculated and compared with the value of the parameter given in the statistical tables available. A simple examination of the two values (calculated & tabular) will show how large a difference needs to be in order to be considered for the limit of significant divergence. Thus, if there is not a statistically significant divergence, means the null hypothesis is valid or that there is no source of systematic error and the variation in results follows the law of random errors. And if there is a statistically significant divergence, means the null hypothesis is not valid and a source of systematic error is highly probable.

We shall discuss below three tests of significance: (i) the t-test which is based on comparison of two means, (ii) the F-test which is based on the comparison of two variances, and (iii) the chi^2 – test (χ^2 – test) which is given in terms of frequencies.

3.9.1 The t-test or Student's t Test

The t-test is used to test the null hypothesis that two means do not differ significantly. The application of t-test will be considered here only in a simple case.
When Accepted Mean Value is Known: Eq. (3.31) used to get the confidence limit is also applicable to the comparison of the finite sample mean $\bar{x}$ and the population mean μ. The quantity t is defined as

$$\pm t = (\bar{x} - \mu)\sqrt{n}/s \qquad \ldots (3.31)$$

where $\bar{x}$ = average value for the finite series

μ = population mean when the series has been carried to an infinite number of observations, or accepted value given by some national standards

s = standard deviation of the finite series, and

n = number of measurements in the finite series.

Values of t with reference to probability levels of 90, 95 and 99 percent are summarized in Table 3.2. In the table ν refers to the degrees of freedom.

In the procedure to apply the null hypothesis the quantity $\pm t = (\bar{x} - \mu)\sqrt{n}/s$ is calculated for the given observations and known as $t_{calculated}$. This value of t is compared with the corresponding value of t (t – tabular) found in table of t (Table 3.2) at the desired confidence level and corresponding to ν degrees of freedom of finite sample (ν = n – 1). On comparison,

i) If $t_{calc} < t_{tab}$, the null hypothesis is valid, there is no significant difference between the two means ($\bar{x}$ and μ), the variation in results is just by random errors and no systematic source of error is probable.

ii) If $t_{calc} > t_{tab}$, then the null hypothesis is incorrect, a significant difference between the two means is indicated and the difference is due to some source of systematic error in the values of finite series.

The above criteria indicate that the smaller the calculated t value, the more confident you are that there is no significant difference between the two means.

Suppose five observations obtained for the determination of atomic mass of cadmium were: 112.25, 112.36, 112.32, 112.21, 112.36. Does the mean of these values differ significantly from the NBS accepted value 112.41?

The test is applied as follows after calculating the required quantities

x	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
112.25	0.05	0.0025
112.36	0.06	0.0036
112.32	0.02	0.0004
112.21	0.09	0.0081
112.36	0.06	0.0036
∑x = 561.5		Sum = 0.0182

$\bar{x}$ = 561.5/5

= 112.30

s = (0.0182) ½ = 0.067 and n = 5

$\pm t = (\bar{x} - \mu)\sqrt{n}/s$

= (112.30 – 112.41) $\sqrt{5}$ /0.067

= – 0.11 × 2.236/0.067 = – 3.67

Or t calc = 3.67 (disregarding the negative sign)

From Table 3.2 at 99% probability level corresponding value of t for 4 degrees of freedom is 4.60. Thus t_{tab} = 4.60.

Comparing two *t* values we see that

$t_{calc} < t_{tab}$ at 99% confidence level.

It is concluded that the null hypothesis is valid at 99% probability level and there is no significant difference between the two means. The variation is due to indeterminate errors.

3.9.2 F-Test

The F test serves to show whether the precision of two different methods is the same within specified probability limits. It is applied in terms of variance ratio. The F value which is the ratio of two variances in question is determined by the relation

$$F = \frac{V_1}{V_2} = \frac{s_1^2}{s_2^2} \quad \ldots (3.32)$$

Placing of the larger of the two variances in numerator ($(s_1^2 > s_2^2)$), so that the value of F is always greater than unity. The value of *F* determined by the use of Eq. (3.51) for experimental variances s_1^2 *and* s_2^2 is known as $F_{calculated}$.

Statisticians have compiled tables of F values for various significance levels for various degrees of freedom ($\upsilon_1 = n_1 - 1$, $\upsilon_2 = n_2 - 1$), where n_1 is the number of observations for the set of larger variance (i.e. larger standard deviation). Table 3.5 is a brief F table (two sided) for the 95% confidence level. The value of F obtained from such a table is called as F tabular.

To test null hypothesis for the two sets of data by F test, the calculated value of F is compared with the corresponding tabular value of F. On comparison (i) If $F_{calc} < F_{tab}$, that is if the experimental F is smaller than the corresponding tabular value of F, then no statistically significant difference is indicated between s_1 and s_2 (i.e. between two sets of data), and the null hypothesis is valid, and (ii) If $F_{calc} > F_{tab}$, then s_1 is significantly greater than s_2, and the null hypothesis is not valid.

Table 3.5: Values of F at 95% confidence level

υ_2	υ_1					
	2	3	4	5	6	∞
2	19.00	19.16	19.25	19.30	19.33	19.50
3	9.55	9.28	9.12	9.01	8.94	8.53
4	6.94	6.59	6.39	6.26	6.16	5.63
5	5.79	5.41	5.19	5.05	4.95	4.36
6	5.14	4.76	4.53	4.39	4.28	3.67
∞	3.00	2.00	2.37	2.21	2.10	1.00

To illustrate "F" test suppose that two series of observations are made one of 4 observations of standard deviation equal to 0.02 and another of 6 observations of standard deviation equal to 0.04. We have to test whether there is significant difference between the two standard deviations.
For the condition of application of F test we have to consider the greater standard deviation that is, 0.04 as s_1 and the smaller 0.02 as s_2. Hence $\upsilon_1 = 6 - 1 = 5$, and $\upsilon_2 = 4 - 1 = 3$
F is calculated as

$$F_{calc} = \frac{V_1}{V_2} = \frac{s_1^2}{s_2^2} = \frac{(0.04)^2}{(0.02)^2} = 4.0$$

Corresponding tabular value of F for $\upsilon_1 = 5$ and $\upsilon_2 = 3$ from Table 3.5 at 95% confidence level is F tab = 9.01.

An examination of two F values you find that $F_{calc} < F_{tab}$, therefore, it is concluded that the null hypothesis is valid and there is no statistically significant difference between the standard deviations of the two sets of data or statistically no significant difference is observed between the precisions of the two sets of data.

3.9.3 The χ^2 (chi-square) Test

Chi-square test is applied to study the behaviour of data if the theoretical behaviour can be expressed quantitatively in terms of expected frequencies. The test is applied to check the number bias (if any) for a particular digit in instrument reading. A number

bias varies considerably from observer to observer and also depends on the number of sub-division of instrument readings. Naturally, for a limited number of observations, a certain fluctuation is statistically expected. However, if the fluctuation, is such that is not governed by probability, can be due to a number bias. Such a number bias can actually impose a limitation upon the accuracy of readings by an individual. The chi-square test informs us about such a number bias. This test gives the comparison of a number of frequency distribution. The quantity chi-square is defined by

$$\chi^2 = \sum(f_i - F_i)^2/Fi \qquad \ldots (3.33)$$

where f_i = observed frequency

F_i = expected frequency

The calculated Chi-square applying Eq. (3.33) is compared with the values of chi-square given in the table for the corresponding degrees of freedom. In Table 3.6 some critical values of chi-square are listed. On comparison of the two,

i) if $x^2_{calc} > x^2_{tab}$, there is a number bias,

ii) if $x^2_{calc} > x^2_{tab}$, there is no number bias, and fluctuation is by chance.

Table 3.6: Some values of Chi-square

υ	Confidence Level	
	95%	99%
1	3.84	6.63
2	5.99	9.21
3	7.81	11.30
4	9.49	13.30
5	11.10	15.10

To illustrate let us consider the tossing of a coin for 500 times. The expected results or theoretical results should give expected frequency of 250 times *head* and 250 times *tail*. It is rare that these results are obtained exactly. Suppose 270 times we get *head.* Then 250 is the expected frequency F_i, and 270 is the observed frequency f_i. From Eq. (3.52) we get

$$\chi^2 = (270 - 250)^2/250$$
$$= (20)^2 / 250 \quad = 1.6$$

Thus, $\chi^2_{calc} = 1.6$

From Table 3.6 for (v = 2 – 1 = 1) one degree of freedom at 95% of confidence level χ^2 tab is 3.84. On comparison we get,

$$\chi^2_{calc} < \chi^2_{tab}.$$

It is concluded that there is no number bias and the fluctuation is by chance.
As a second example if we were to examine the last digit of the students burette readings estimated to 0.01 mL, we would find a considerable number bias in favour of the last digits as 0 and 5.

3.10 CONTROL CHARTS

In order to keep the track of day by day performance of the production process that is, for quality control, the industrial analytical laboratories often use the control chart

technique. By plotting a sequence of points in order, a continuous record of the quality characteristic is available. This technique readily indicates whether or not a result lies within certain confidence limits, and the trends in data or sudden lack of precision can be made evident so that the causes may be sought.

The control charts usually have 3 or 5 horizontal lines; that is, a central line and either one or two pairs of limit lines, the *inner* and the *outer* control limits. The central line represents some standard value. The inner and outer pair of lines corresponds to the confidence limits for the entries made on these charts. Usually, a pair of inner limit lines of deviation 2σ is drawn on both sides of the central line representing a confidence level of 95 percent. Thus, the probability of an observation falling outside this limit is 5 percent, that is, only 1 in 20, and a tendency for a greater scatter would indicate that the precision is not under control of 95 percent confidence. It is common practice in most industries to set inner control limits of $\pm 2\ \sigma$ as warning limits and outer control limits of $\pm 3\ \sigma$ (Fig. 3.6). The outer control limits correspond to a confidence level of 99.7 percent, or a probability of 0.003 that a point will fall outside these limits.

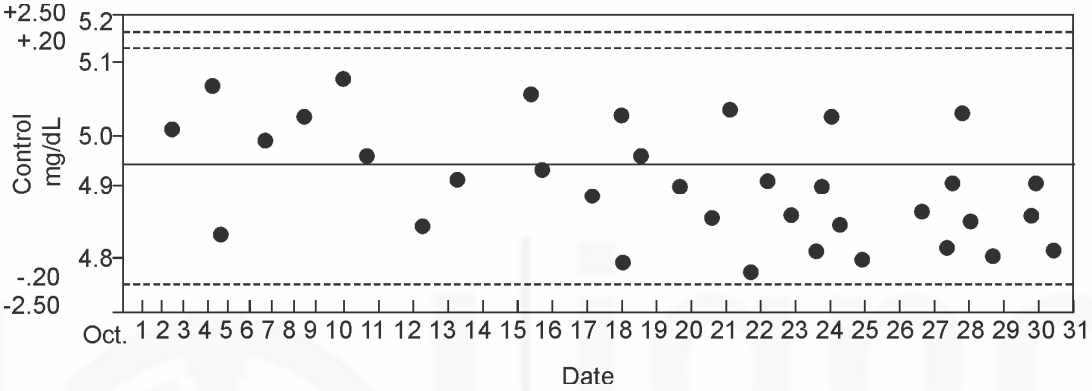

Fig. 3.4: Control Chart

The entries which are usually averages of three or more observations are plotted serially on the control chart one after the other. The plotted points that lie inside the particular pair of limit lines indicate result within control with the given degree of confidence whereas the plotted points that fall outside the particular pair of limit lines indicate falling out of control for the given degree of confidence. Special attention should be paid to one sided deviation from the control limits, because systematic errors often cause deviation in one direction.

3.11 SUMMARY

Statistical analysis is necessary to understand the significance of analytical data. In this unit you have studied the methods used by scientists in evaluating the significance of analytical data with the knowledge of normal distribution of errors in terms of probability. Replicate determinations should be made in order to approach the value of experimental mean around the true mean with a certain degree of probability. To indicate the precision of an analysis the most important statistical parameter is the standard deviation. A clearer picture of data quality is sometimes obtained by relative standard deviation. Precision of the calculated results is estimated by the calculation of standard deviation by taking care of uncertainties of different sets of data used in computation of results.

You have learnt here that the study of probability distributions is of fundamental importance to the use of statistics for judging the reliability of analytical data. What we say is all measurements contain random errors which follow the normal law of

error. It can be expressed by a differential Eq. The area under various limits of the normal error curve can be calculated by the integration of the differential Eq. The area is the measure of probability that an observation lies in these limits. Various confidence limits are used to get the confidence interval. The rejection of an outlying observation can be ascertained by the suitable statistical methods.

In analytical chemistry it is frequently desired to compare the results of two different methods, one of them is the test method and the other is usually an accepted method. This is done by null hypothesis using the tests of significance: the t-test which is based on the comparison of two means, the F-test which is based on the comparison of two variances, and the chi-square test which is given in terms of frequencies. In many analyses the relation between a physical quantity measured and concentration on plotting gives a straight line. The method of least square is used to obtain the best straight line for which the sum of the squares of deviations of the points from the line is minimum. For quality control, the industrial analytical laboratories often use the control chart technique which tells the trends in data of certain analysis.

3.12 TERMINAL QUESTIONS

1. Consider the following set of replicate measurements of an analyte: 0.792, 0.794, 0.813 and 0.900 g. The true value is 0.830 g. Calculate (a) mean (b) median (c) range (d) standard deviation (e) coefficient of variation (f) absolute error of the mean (g) the relative error of the mean in parts per thousand. Consider no observation is rejected.

2. Calculate the uncertainty of the operation $y = a/b$. The individual uncertainty (as standard deviation) of each quantity is given in parenthesis: $a = 36.2 (\pm 0.4)$; $b = 27.1 (\pm 0.6)$. Express the calculated result with absolute uncertainty.

3. An analyst got the percent alcohol content in a blood sample: 0.084, 0.089 and 0.079. Calculate the 95% confidence limit for the mean assuming t = ± 4.30 for two degrees of freedom and 95% confidence.

3.13 ANSWERS

Self Assessment Questions

1. Arranging sequentially the observations: 13.4, 13.5, 13,6, 13.8, 14.1, 14.3 the median will be the mean of 3rd & 4th observation.
   Median = (13.6 + 13.8) / 2 = 13.7

2. $\bar{x}$ = 29.6 mg, s = 0.69 mg
   CV = (0.69/29.6) × 100 = 2.33%
   RSD (ppm) = (0.69/29.6) × 10^6 = 2.33×10^4 ppm.

3. i) Fraction of area under the normal error curve lying between – 1 σ to + 1 σ = 0.683
   Therefore, probability of a result lying between 0 and + 1 σ is = 0.683/2 = 0.342 or 34.2%

   ii) Similarly probability between 0 and + 2 σ = 0.954/2 = 0.477
   Therefore, probability between + 1 σ and + 2 σ will be
   = 0.477 – 0.342 = 0.135 or 13.5%

4. Range R = 16.30 – 15.30 = 1.00
   $\bar{x}$ = (15.30 + 15.85 +15.55 + 16.30)/4 = 15.75
   μ = $\bar{x}$ ± Cn R = 15.75 ± 0.53 × 1.00

15.75 – 0.53 = 15.22 & 15.75 + 0.53 = 16.28

Therefore, with 90% confidence the population mean μ lies between 15.22 and 16.28 mg/100 mL.

5. The value 52.40 is not fitting well in the other observation, therefore, this value is the suspected value. Omitting this value, the arithmetic mean of the rest is,

$$\bar{x} = \frac{52.48 + 52.50 + 52.51 + 52.46}{4} = 52.49$$

Average deviation = $\frac{0.01 + 0.01 + 0.02 + 0.03}{4} = 0.07/4$

and 4 × a.d. = 0.07

Difference between suspected value and the mean of the rest = 52.49 – 52.40 = 0.09 which is greater than 4 × a.d. Hence the suspected value 52.40 should be rejected according to 4d rule.

Terminal Questions

1. a) Mean $= \frac{\Sigma x_i}{n} = \frac{0.792 + 0.794 + 0.813 + 0.900}{4} = \frac{3.299}{4}$

$= 0.824525$

Rounding up to three decimals the mean is

$\bar{x} = 0.825$ g

b) Median $= \frac{0.794 + 0.813}{2} = 0.8035$

= 0.804 g (Rounding off to 3 decimals)

c) Range = 0.900 – 0.792 = 0.108 g

d)

$\lvert x_i - \bar{x} \rvert = d_i$	$(x_i - \bar{x})^2 = d_i$
0.825 – 0.792 = .033	1089×10^{-6}
0.825 – 0.794 = 0.031	961×10^{-6}
0.825 – 0.813 = 0.012	144×10^{-6}
0.900 – 0.825 = 0.075	5625×10^{-6}
	$\Sigma d_i = 7819 \times 10^{-6}$

$$s = \left(\frac{\Sigma d_i^{\,2}}{n-1}\right)^{1/2} = \left(\frac{7819}{3} \times 10^{-6}\right)^{1/2} = 51.05 \times 10^{-3} \text{ g}$$

Standard deviation s = 0.051 g.

e) Coefficient of variation CV = $(s/\bar{x}) \times 100$

= (0.051 / 0.825) × 100 = 6.18%

f) Absolute error of the mean = Mean – True value

= 0.825 – 0.830 = – 0.005 g

g) Relative error of the mean = (– 0.005/0.830) × 1000

= – 6.02 ppt

2. First let us calculate the result without standard deviations.

36.2/27.1 = 1.335 = 1.34

Now to get the relative standard deviation of y, let us apply the rule of multiplication and division. Thus,

$$\left(\frac{s_y}{y}\right)^2 = \left(\frac{s_a}{a}\right)^2 + \left(\frac{s_b}{b}\right)^2$$

$= (0.4/36.2)^2 + (0.6/27.1)^2$

$= 122 \times 10 - 6 + 490 \times 10 - 6 = 612 \times 10 - 6$

sy/y = $(612 \times 10 - 6)1/2 = 0.0247$

sy/y = $0.0247 \times 1.34 = 0.33$

Uncertainty of the operation = ± 0.03

The result is, y = 1.34 ±0.03

3. Mean $\bar{x} = (0.084 + 0.089 + 0.079)/3 = 0.084$

$$s = \sqrt{\frac{(0.000)^2 + (0.005)^2 + (0.005)^2}{3-1}} = 0.005$$

95% confidence limit $\mu = \bar{x} \pm ts/\sqrt{n}$

$\mu = 0.084 \pm (4.3 \times 0.005) / \sqrt{3}$

$= 0.084 \pm 0.012$

Some Useful Books

1. *Analytical Chemistry* by Cary D. Christian , John Wiley and sons.

2. *Basic concepts of Analytical Chemistry* by S.M. Khopkar, New Age International Publishers.

3. *Vogel's Textbook of Quantitative Chemical Analysis* by J. Menham, R.C. Denney, J.D. Barnes and M.J.K. Thomas, 6th Edn, Low Price Edition, Pearson Education Ltd, New Delhi (2000).

INDEX