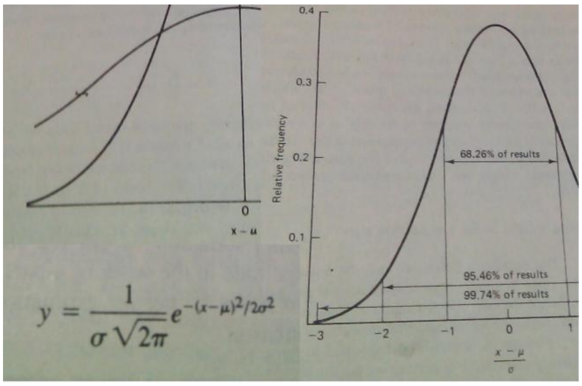
Unit I

Data Analysis



Science is based on quantitative data. They are generally derived from experimental measurements and may have some error. So it is important to study the error to get accurate value. Thus, here we discuss some methods used by analytical chemists in assessing the significance of experimental results.

Error

The term error is used to deal the numerical difference between a measured value and the true value. The true value of any quantity is something we never know, although we generally accept a value as being true when it is believed that the uncertainty in the value is less than the uncertainty in something else with which it is being compared. The percentage composition of a standard sample is certified by the National Institute of Standards and Technology (NIST), which may be regarded as standard or correct value.

Thus, error may be defined as the differences between the standard values and the results obtained by the new method are treated as error.

While discussing the term error we should learn more about their different types. Generally, error can be described by determinate and indeterminate error. Determinate errors are generally unidirectional with respect to the correct or true value. However, indeterminate errors lead to both high and low results with equal probability.

Determinate errors have been classified as methodic, operative and instrumental accordance with their origin.

Determinate errors can also be classified as constant and proportional.

Examples of sources of determinate errors are incorrectly calibrated instruments such as pH meter, balance, burette, etc.

Indeterminate errors cannot be attributed to any known cause, but they invariably attend measurements made by human beings. These errors cannot be corrected and hence are the ultimate limitation of the measurement.

Accuracy and Precision

A result can be treated as accurate when the value agrees closely with the true value of a measured quantity. A comparison is usually made on the basis of an inverse measure of the accuracy, i.e. the error. The absolute error is the difference between the experimental value and the true value. Suppose an analyst obtained a value of 24.24% copper in a sample which actually contains 24.14%, the absolute error is

$$24.24 - 24.14 = 0.10\%$$

It can be measured in percentage or in parts per thousand. **Relative error** is the absolute error divided by the true value. It is measured in percentage or parts per thousand Here the relative error is

$$(0.10/24.14) \times 100 = 0.41\%$$

Or, $$(0.10/24.14) \times 1000 = 4.14 \text{ ppt}$$

Precision is defined as the concordance of a series of measurement of the same quantity. It implies nothing about their relation to the true value. The term precise is commonly stated in terms of the **standard deviation, average deviation or range.**

The Normal Error Curve: Gaussian Distribution Curve

When a large number of replicate readings, at least 50 numbers are taken of a continuous variable, the results attained will usually be distributed about the mean in a roughly symmetrical manner. The mathematical model that best satisfies such a distribution of random error is called the normal Gaussian distribution.

Significance: Gaussian distribution curve is a bell-shaped that is symmetrical about the mean as shown in figure 1.

This curve satisfies the equation

$1/(\sigma\sqrt{2\pi})\ e^{-\,(x\,-\,\mu)2}/2\sigma^2$

Where, σ = standard deviation

μ = mean of total population

In Gaussian distribution about 68% of all values will fall within one standard deviation on either side of the mean, 95% will fall within two standard deviations and 99.7% within three standard deviations.

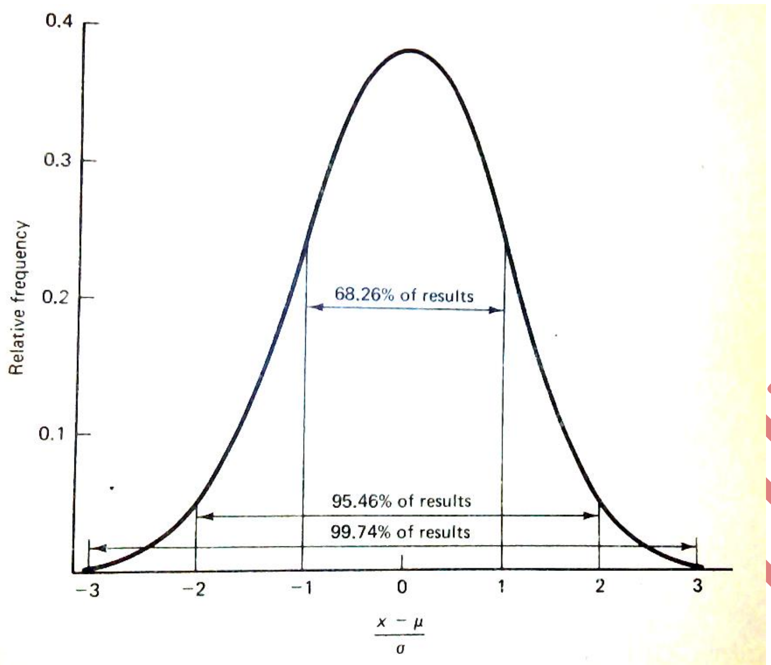


Fig. 1: Normal or Gaussian Distribution curve

Statistical Treatment of Finite Samples

The central tendency of a group of results is simply that value about which the individual results tend to "**cluster**". For an infinite population, it is μ, the mean of a such sample.

Mean

The *mean* of a finite number of measurements, x_1, x_2, x_3, x_4,……..,x_n is often designated $\bar{x}$ to distinguish it from μ. Of course, $\bar{x}$ approaches μ as a limit when the measured value approaches infinity.

$$\bar{x} = (x_1 + x_2 + x_3 + x_4 + \dots\dots + xn)n = \sum_{i=1}^{i=n} xi/n$$

It may be shown that mean of *n* results is $\sqrt{n}$ times as reliable as any of the individual results. The mean of 4 results is twice as reliable as 1 result in measuring central tendency. The mean of 9 results is three times as reliable, the mean of 25 results, five times as reliable etc.

Median

The *median* of an odd number of results is simply the middle value when the results are listed in order. However the median of an even number of results is the average of the tow middle ones. In a symmetrical distribution, the mean and the median are identical. The median is a less efficient measure of central tendency than is the mean.

Range

It is the difference between the largest and smallest values. Like the median, the range is sometimes useful in small statistics, but generally speaking it is an inefficient measure of variability.

Average deviation

The average deviation from the mean is often given in scientific papers as a measure of variability, although strictly it is not very significant from a statistical point of view. For a large group of data which is normally distributed, the average deviation approaches 0.8σ to calculate the average or mean deviation, one simply finds the differences between individual results and the mean, regardless of sign, adds these individual results, and by divides by the number of results

Average deviation, $\bar{d} = \sum_{i=1}^{i=n} \vdots\, xi - \bar{x}\ \vdots/n$

Relative standard deviation

Often the average deviation is expressed relative to the magnitude of the measured quantity, for example, as a percentage

Relative standard deviation (%), $\bar{d}/\bar{x}$ x 100 $=\{(\sum_{i=1}^{i=n} \vdots\, xi - \bar{x}\ \vdots/n)/\ \bar{x}\}$x100

in parts per thousand

Relative standard deviation (ppt), $\bar{d}/\bar{x}$ x 1000 $=\{(\sum_{i=1}^{i=n} \vdots\, xi - \bar{x}\ \vdots/n)/\ \bar{x}\}$x1000

Standard deviation

The standard deviation is much more meaningful statistically than is average deviation. The symbol *s* is used for the standard deviation of a finite number of values. The standard deviation, which may be thought as a root mean square deviation of values from their average.

Standard deviation, $s = \surd[\ \sum_{i=1}^{i=n} \vdots\, xi - \ \bar{x}\ \vdots^{2}/\ (n\text{-}1)$

If *n* is large (50 or more), then it is immaterial whether the term in the denominator is *n*-1 or *n*. When the standard deviation is expressed as a percentage of the mean, it is called the coefficient of variance, *υ*

$\upsilon = (s/\bar{x}\ )$ x 100

Variance

It is the square of standard deviation, which is designated as s^2. The variance is fundamentally more important in statistics than is *s* itself.

Variance, s^2 = (Standard deviation)^2

Exercises

Ex. 1: An iron core gives the following results during the Fe estimation as the value 7.08, 7.21, 7.12, 7.09, 7.16, 7.14, 7.07, 7.14, 7.18, 7.11. Calculate the mean, the standard deviation and coefficient of variance for the values.

Solution:

The mean, $\bar{x}$ = (7.08 + 7.21 + 7.12 + 7.09 + 7.16 + 7.14 + 7.07 + 7.14 + 7.18 + 7.11)/10

= 71.98/10 =7.13

Standard deviation, $s = \sqrt{}[\ \sum_{i=1}^{i=n} \vdots\ xi - \ \bar{x}\ \vdots^{2}/\ (n\text{-}1)$

$= \sqrt{}(7.08\text{-}7.13)^2 + (7.21\text{-}7.13)^2 + \ldots\ldots\ldots\ldots(7.18\text{-}7.13)^2 + (7.11\text{-}7.13)^2/9$

$= \sqrt{}181.98 \text{ x } 10^{-4}/9$

$= 4.49 \text{ x } 10^{-2} = 0.0449$

Coefficient of variance, $\upsilon = (s/\bar{x}\ ) \text{ x } 100$

= {0.045/7.13} x 100

= 0.63 %

Ex. 2: The normality of a solution is determined by four analysts. The results being 0.2041, 0.2049, 0.2039 and 0.2043. Calculate mean, median, range, average deviation, relative average deviation, standard deviation and coefficient of variance.

Solution:

The mean, $\bar{x}$ = (0.2041 + 0.2049 + 0.2039 + 0.2043)/4

= 0.2043

Median, M = (0.2041 +0.2043)/2

= 0.2042

Range, R = 0.2043 -0.2039

= 0.0010

Average deviation, $\bar{d} = \sum_{i=1}^{i=n} \vdots\ xi - \bar{x}\ \vdots/n$

= (0.0002) + (0.0006) + (0.0004) + (0.0000)/4

= 0.0003

Relative standard deviation (ppt), $\bar{d}/\bar{x}$ x 1000 = $\{(\sum_{i=1}^{i=n} \vdots\ xi - \bar{x}\ \vdots/n)/\ \bar{x}\}$x1000

= (0.0003/0.2043) x 1000

= 1.5 ppt

Standard deviation, $s = \sqrt{}[\ \sum_{i=1}^{i=n} \vdots\ xi - \ \bar{x}\ \vdots^{2}/\ (n\text{-}1)$

= 0.0004

Coefficient of variance, $\upsilon = (s/\bar{x}\ ) \text{ x } 100$

$= \{0.0004/.2043\} \times 100$

$= 0.2\ \%$

Analysis of results

Analysis of results can be made by two different ways. They are:

(i) Comparison of results

(ii) Reliability of results

Comparison of sets of values with either true value or with another set of values gives us the trick to determine whether the sets of values or the analytical procedure is accurate of precise.

There are two common methods

(a) Student's t -test

(b) Variance ration test or F –test

t –test: Student's t –test is used for small samples. It can also be used to test the difference between the mean of two sets of data's ($\bar{x}_1$ and $\bar{x}_2$). The purpose of the test is to compare the mean of samples with some standard value and to express some level of confidence in the significance of comparison.

The t –test is obtained as

$$\pm t = (\bar{x} + \mu)\sqrt{N}/s$$

Where, $\bar{x}$ is the mean value

μ is true value

N is number of determination

s is standard value

These calculated values is then compared with the sets of values obtained for different probabilities and *degree of freedom* from the given table

(*degree of freedom:* It may be defined as the number of individual observations that could be allowed to vary under the condition that $\bar{x}$ and s, once determined, be held constant.)

Confidence interval of the mean

By rearranging the above equation, we obtain the confidence interval of the mean or confidence limits

$$\mu = \bar{x} \pm t.s/\sqrt{N}$$

We can use this equation to estimate the probability that the population mean, μ, lies within a certain region centred at $\bar{x}$, the experimental mean of our measurements.

Exercises

Ex.3: If $\bar{x}$ of twelve determination is 8.37 and a true value, μ =7.91, say whether or not these results is significant if the standard deviation, s is 0.17.

Solution.

The t value is

$$t = [(\bar{x} - \mu)\sqrt{N}]/s$$

$$= [(8.37 - 7.91)\sqrt{12}]/0.17$$

$$= 9.37$$

	20%	10%	5%	2%	1%
Probability (α)	0.20	0.10	0.05	0.02	0.01
Given t value	1.363	1.796	2.20	2.72	3.71

The calculated value for t is 9.37. Comparison of this value with true value, it implies that the calculated t value is highly significant. The t value tells us that the probability of obtaining the difference of 0.46 is less than 1 in 100.

Ex. 4: For the estimation of iron from a sample following results are obtained:

8.43, 8.41, 8.4, 8.32, 8.34, 8.24

Find out the mean standard deviation coefficient of variance and say whether or not these results are significant if the true value is 8.37

F-test: This test is used to compare the precision of two sets of datas. It is calculated as

$F = S_A^2/ S_B^2$

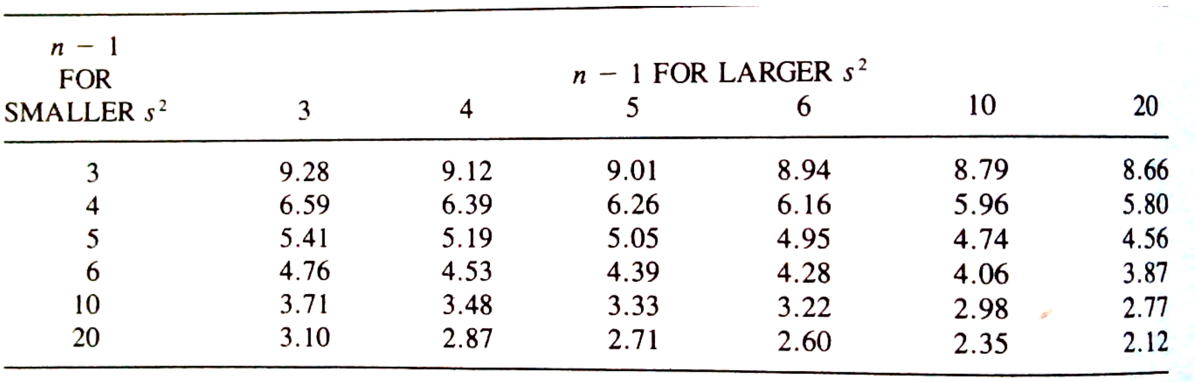
Where S_A is standard deviation for one method

S_B is standard deviation of another method

Generally, S_A>S_B

The value obtained for ‘F’ is then check for its significance against values in the F table. If the calculated F value can relate with the lower probability then the two sets of data’s is highly significant.

The F test may be used to determine the validity of the sample t test described here, but it may also be interest in its own right to determine whether two analytical procedures yield significantly different precision.



$n-1$ FOR SMALLER s^2	$n-1$ FOR LARGER s^2					
	3	4	5	6	10	20
3	9.28	9.12	9.01	8.94	8.79	8.66
4	6.59	6.39	6.26	6.16	5.96	5.80
5	5.41	5.19	5.05	4.95	4.74	4.56
6	4.76	4.53	4.39	4.28	4.06	3.87
10	3.71	3.48	3.33	3.22	2.98	2.77
20	3.10	2.87	2.71	2.60	2.35	2.12

Source: Quantitative Analysis, 6th edn, R.A. Day *et al.*