

Chapter 6

Basic Statistics Concepts

Every physical measurement is subject to a degree of uncertainty. A result that is not particularly accurate may be of great use if the limits of the errors affecting it can be set with a high degree of certainty. Statistical methods are used to evaluate the probable errors in analytical measurements and interpretation of environmental data. This module discusses some basic concepts of statistics, which a chemist should be conversant with in order to evaluate the accuracy of the laboratory data and the estimation of the environmental characteristics based on the results of limited samples.

1 Frequency distribution

Results of 44 replicate analyses for hardness of a sample of water are given in Table 1.

Table 1. Results of 44 replicate analyses for hardness, mg/l as CaCO₃

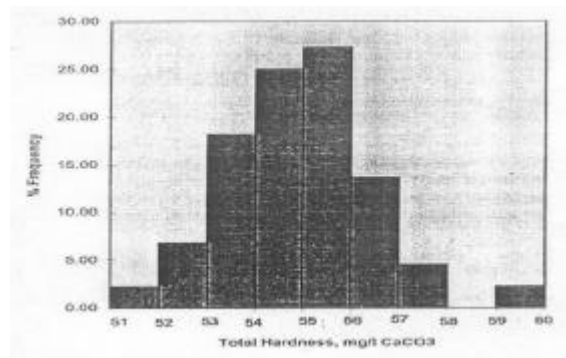
Class	Value	No. of Values in the Class	Frequency
(1)	(2)	(3)	(4)
51-52	51.2	1	0.0227
52-53	52.0,52.8,52.8	3	0.0682
53-54	53.1,53.1,53.3,53.4,53.5,53.6,53.6,53.9,	8	0.1818
54-55	54.0,54.1,54.3,54.3,54.4,54.5,54.5,54.6,54.7,54.7,54.9	11	0.2500
55-56	55.1,55.1,55.3,55.3,55.4,55.4,55.6,55.7,55.7,55.8,55.9,55.9	12	0.2727
56-57	56.2,56.3,56.7,56.8,56.9,56.9	6	0.1364
57-58	57.3,57.5	2	0.0454
58-59	-	0	
59-60	59.1	1	0.2227

The data are classified in 9 classes, column (1) and arranged in each class according to their magnitude, column (2). By convention an observation equal to the upper value of the class is counted in the next higher class. For example the value 52.0 is not placed in class 51-52 but in 52-53. The number of values in each class are given in column (3). The ratio of number of values in a class to the total number of observation is called frequency and is given in column (4).

A plot of data of column (1) vs. column (4) is called frequency distribution diagram and is shown in Figure 1. For clarity of presentation, the frequency value is multiplied by 100 and shown as %.

Note that if the number of observations is increased and the class interval is reduced, the histogram can be replaced by a continuous curve.

Figure1: Example of a frequency distribution diagram



Central Tendency

Arithmetic mean: The arithmetic mean, \bar{X} , of a set of data is calculated by adding all the observed values of variable (results of analyses) and dividing it by the total number of observations:

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n \tag{1}$$

where X_1, X_2, \dots, X_n are the observed values and n is the total number of observations.

The arithmetic mean is the most common measure of the central tendency. The mean value of the data of Table 1 was calculated as 54.9

Geometric mean: When there are a few very high values or very low, such as in the cases of bacteriological analysis, the arithmetic mean is not necessarily representative of the central tendency. In such cases the geometric mean, g , is used:

$$g = \sqrt[n]{X_1 \times X_2 \times X_3 \times \dots \times X_n} \tag{2}$$

Median: The median is the middle value of the set of data. If the sample size n is an odd number, one-half of the values exceed the median and one-half are less. When n is even, it is average of the two middle terms. For the data of Table 1, it is 54.8, which is the average of 54.7 and 54.9, the 22nd and the 23rd terms, when the data are arranged in ascending order.

Mode: The mode is the most commonly occurring value. It is not widely employed as it forms a poor basis for any further arithmetic calculations.

Standard deviation

The data of Table 1 show that 52% of observations lie in the range of 54 - 56 and 84% in the range of 53-57. This tendency of observations to cluster (or not to cluster) around the mean value, 54.9, is measured by *standard deviation*, s . Standard deviation is calculated as:

$$s = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}}$$

$$\text{or } s = \sqrt{\frac{\sum X_i^2 - (\sum X_i)^2/N}{N}} \quad (3)$$

A small value of s signifies that most of the observations are close to the mean value. A large value indicates that the observed values are spread over a larger range. For example, if the data of Table 1 has more observations in 52-53 and 57-58 classes and correspondingly lesser number in the central classes 54-55 and 55-56, the frequency distribution diagram will be flatter and spread wider at the base and the value of s will be larger.

The standard deviation has the same units as the quantity measured. The standard deviation for the data of Table 1 was calculated as 1.545 mg/L.

The square of the standard deviation is called the *variance*. If the random distribution of data is due to r different reasons, the total variance is the sum of individual variance.

Normal Distribution

Most frequency distributions for *random* observations, when the total number of observations is very large tending to be the same as the population, N conform to normal distribution. The distribution is given by the theoretical equation:

$$y = \frac{e^{-(x_i - \mu)^2 / 2\sigma^2}}{\sigma\sqrt{2\pi}} \quad (4)$$

where y = frequency of observations / l = the mean, and σ = standard deviation. Note the distinction made in the notations for number of observations, mean and standard deviation for a sample of a limited set of data used earlier.

The standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum x_i^2 - (\sum X_i)^2/N}{N}} \quad (5)$$

For a normal distribution, 68.3 % of the total observations lie in the range $\mu \pm \sigma$, 95.5% in the range $\mu \pm 2\sigma$ and 99.7% in the range $\mu \pm 3\sigma$. This is illustrated in Figure 2.

The number of observations between any two limits of the variable can be equated to the area bounded by the frequency distribution curve, the ordinates at these limits and the x axis.

When the data set is small, more often than not, sample mean, \bar{X} will differ somewhat from the population mean, μ . Any error in \bar{X} causes a corresponding error in the standard deviation calculated with Equation 5. There is a tendency for the standard deviation value calculated to be small as the number of measurements becomes smaller. It can be shown that this bias can be largely eliminated by substituting the *degree of freedom* as $(n-1)$ in the calculation, Equation (3), which defines the standard deviation for a limited set of measurements.

The rationale for using $(n-1)$ is as follows. When μ is not known, we must extract two quantities, namely, \bar{X} and s from the set of data. The need to establish \bar{X} from the data removes one degree of freedom: that is if their signs are retained, the sum of individual

deviations must total zero; once $(n-1)$ deviations have been established, the final deviation is necessarily known. Thus only $(n-1)$ deviations provide independent measures for the standard deviation of the set.

2 Precision and Accuracy of Experimental Data

Classes of errors

The phenomena that are responsible for uncertainties in an analytical measurement can be divided in two broad categories: *determinate* or *system errors* and *indeterminate* or *random errors*. The total error is the sum of the two types of errors.

Determinate errors have assignable causes and are unidirectional. It may be possible to eliminate some of these, for example, errors because of improper calibration of the equipment, or presence of interfering substances in the sample. Errors inherent in the method, such as addition of a reagent in excess of theoretical requirement to cause an indicator to undergo colour change in a titration, may be difficult to eliminate.

Indeterminate errors are encountered whenever a measuring system is extended to its maximum sensitivity. The results fluctuate in a random manner about a mean value. The sources of these fluctuations can never be identified because they are made up of a myriad of instrumental, personal and method uncertainties that are individually so small that they can never be detected. For example, in the case of addition of an exact volume of reagent in an analysis through a pipette the uncertainties could be the time allowed for draining the pipette, the angle at which the pipette is held during delivery, temperature of the reagent, visual judgement of water level with respect to the graduation mark, etc. What is observed in the final result is then a summation of a very large number of minute unobservable uncertainties. The cumulative effect is likewise variable. Ordinarily they tend to cancel one another and thus exert a minimal effect. Occasionally, however, they act in concert to produce a relatively large positive or negative error.

Precision

The term precision is employed to describe the reproducibility of results. It can be defined as the agreement between the numerical values of two or more measurements that have been made in an identical fashion.

Absolute methods for expressing precision: The *absolute average deviation* from the mean is a common method for describing precision. The *spread* or *range* of a set of data is also a measure of precision and is simply the numerical difference of the highest and the lowest result. *Standard deviation*, described earlier, is a more significant measure of precision.

Example 1

From the data of replicate chloride analysis of a water sample given below, calculate the precision in terms of the average deviation from the mean:

Analysis	Chloride, mg/L	Abs. dev. from the Mean, mg/L
1	24.39	0.077
2	24.19	0.123
3	24.36	0.047
Total	72.94	0.247
	Mean = 24.313	ave. dev. = 0.247/3 = 0.082

Relative methods for expressing precision: It is frequently more informative to indicate the relative precision. The relative parameters are dimensionless and therefore can be used to compare two or more sets of data. Thus, for example, the *relative deviation* of analysis 1, in the above example, is $(0.077 \times 100)/24.313 = 0.32\%$.

Relative standard deviation or coefficient of variation is defined as:

$$CV = \frac{100s}{\bar{X}} \quad (6)$$

Example 2

Monitoring results at three sampling locations are listed below. Compare the sampling records in terms of variability.

	A	B	C
	40.0	19.9	37.0
	29.2	24.1	33.4
	18.6	22.1	36.1
	29.3	19.8	40.2
\bar{X}	29.28	21.48	36.68
s	8.74	2.05	2.80
CV	0.30	0.10	0.07

In terms of the coefficient of variation the concentration at A varies the most, followed by B and then C.

Accuracy

The term accuracy is used to describe the total error of the observation, which is the sum of the systematic and random errors. It denotes the nearness of the measurement to its accepted value.

In Example 1, if the accepted value for the chloride concentration is 24.35 mg/L, the absolute error of the mean is $24.31 - 24.35 = 0.04$ mg/L.

3. Propagation of errors

The indeterminate error or uncertainty is most commonly expressed as standard deviation. When the final result is computed from two or more data, each of which has an indeterminate error associated with it, the error of the result will depend on the arithmetic computations. The following examples illustrate the procedure to be followed:

Example 3:

For the sum

$$\begin{array}{r}
 + 0.50 (\pm 0.02) \\
 + 4.10 (\pm 0.03) \\
 - 1.97 (\pm 0.05) \\
 \hline
 + 2.63 (\pm ? ?)
 \end{array}$$

where the numbers in parentheses are the absolute standard deviations. Statistically the most probable standard deviation would be given by the square root of the sum of individual variances:

$$\begin{aligned}
 s &= \sqrt{(\pm 0.02)^2 + (\pm 0.03)^2 + (\pm 0.05)^2} \\
 &= \pm 0.078
 \end{aligned}$$

Therefore the result could be reported as:

$$2.63 (\pm 0.078)$$

Example 4:

The case when products and quotients are involved is illustrated in the following calculations:

$$\frac{4.10 (\pm 0.02) \times 0.0050 (\pm 0.0001)}{1.97 (\pm 0.04)} = 0.0104 (\pm ?)$$

In this case it is necessary to calculate the relative standard deviations:

$$(S_a)_r = \frac{\pm 0.02}{4.10} = \pm 0.0049$$

$$(S_b)_r = \frac{\pm 0.0001}{0.005} = \pm 0.020$$

$$(S_c)_r = \frac{\pm 0.04}{1.97} = \pm 0.020$$

$$(S_y)_r = \sqrt{(\pm 0.0049)^2 + (\pm 0.02)^2 + (\pm 0.02)^2} = \pm 0.029$$

The absolute standard deviation of the result will be

$$S_y = 0.0104 \times (\pm 0.029) = \pm (0.0003)$$

Therefore the uncertainty of the result can be written as:

$$0.0104 (\pm 0.0003)$$

For calculations that involve both sums and differences as well as products and quotients, the uncertainties associated with the former are evaluated first and then the latter following procedures illustrated in the two examples.