Appendix A1: Analytical Errors and Statistical Treatment of Data

De Jesús M. A.; Padovani J.I. (2010); University of Puerto Rico; Mayagüez Campus; Department of Chemistry; P.O. Box 5000; Mayagüez P.R. 00681.

Any analytical method is subject to errors. In order to estimate the extent of these errors, a series of replicate analysis should be performed to test the reproducibility **(precision)** of its results. In a typical analytical method, the number of repetitions lies between two and six. On the other hand, the results of a specific determination are compared with those obtained for the same sample but using a different method, in order to have an idea of how close they are to the accepted value **(accuracy). The analytical chemist's role is to minimize these errors in order to keep the results within acceptable limits of accuracy.** In this section, we will discuss sources of error and their effects in a chemical analysis. We will also discuss the use of statistical methods to estimate the reliability of analytical data.

A. Errors in Experimental Data:

Errors that affect an analytical method can be classified in three categories: **random (indeterminate errors)**, **systematic (determinate errors), and outliers (gross errors)**. **Random errors** are associated to the ultimate limitations of physical measurements. As their name suggests, **these errors can be either positive or negative. They are always present and cannot be corrected.** Some sources of random errors are:

- Subjective interpolations between the markings while reading a buret
- Electrical noise generated by an instrument
- Changes in the earth's magnetic field
- Noise arising from the environment

The use of replicate analysis (replicate samples) is the best approach to minimize this type of error. **Replicate analysis brings about cancellation of uncertainties that have similar magnitudes but opposite effects**.

Unlike random errors, **systematic errors** have a definite value and magnitude, and have an identifiable source. This type of error biases the analytical method and affects all the data. Systematic errors can be **extensive** (dependent upon the amount of matter) or **intensive** (independent of the amount of matter). Systematic errors can be classified in three categories: **Instrumental, Method, and Personal errors.**

Instrumental Errors: these errors arise from imperfections in the design of the measuring devices. For example, in an UV-VIS spectrophotometer, the electrical components and circuits accumulate dust, thus increasing its electrical resistance. The instrument is also susceptible to temperature changes that produce variations in electrical conductivity. Non-electronic apparatus, e.g. burets and pipets, are also subject to errors. These may arise due to small imperfections in the glass or impurities adsorbed on its surface. This type of errors can be corrected by means of:

- Validation of the analytical method with standards of known concentration.
- Calibration against a blank sample that contains all the components of the matrix except the analyte.
- Certification of the reliability of the method by comparing it with other analytical methods.
- Analysis of the sample with different instruments or by other laboratories.

Method Errors: They are produced by the non-ideal behavior of the reagents and samples upon which the analysis is based. For example, the incompleteness of a reaction, and sample contamination or decomposition, may be considered as possible sources for this type of errors, which, in fact, are the most difficult to determine and correct.

Personal Errors: These errors arise from biases or physical limitations of the analyst. They may include differences to perceive colors and biases while reading the pointer in a scale or estimating the position of the meniscus in a volumetric flask. Practice and automation can reduce this type of errors.

On the other hand, **gross errors** differ from random and systematic errors since they rarely occur and have a specific direction. These errors lead to **outliers** or results that markedly differ from the rest in a replicate set of data. The use of statistical methods for the rejection of outliers will be discussed on the next section.

A. Statistical treatment of data:

Statistical analysis is used to determine the probability that a series of experimental results for a given population may be correct. It also serves as a tool to reject those results with a high probability of being incorrect. **Probability is a function associated with the tendency of an event to take place**. In its general form it is defined as:

$$P(x) = \frac{q}{r} \qquad (1)$$

where:

P (x) = is the probability that an event **x** may take place
q = is the number of times that the event x takes place
r = is the number of possible events

This equation has the advantage that it expresses the probability as a real number between zero and one. Probabilities of this type are known as **objective probabilities** since they can be determined within a confidence interval. But sometimes it is impossible to determine the possible number of events. This type of probabilities is known as **subjective probabilities.** They can be estimated from previous experiences and trends, as for example the determination of the exact number of students that will graduate on the next semester if a hurricane passes over Puerto Rico this year). Since a quantitative analysis is based upon objective probabilities, our discussion will be focused on them.

1. Analyzing Population Data (more than 20 replicate samples):

It is important to notice that a common statistical analysis is based upon the assumption that only random errors are present. A plot showing frequency (probability) of occurrence vs. number of events is called a probability distribution curve. The function that best represents the area under the curve on this type of plot is known as a probability density function. The most used probability function is the **Gaussian Distribution**, where the frequency of occurrence for a value, x, is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]} \qquad (2)$$

where:

x_i = individual value of the population
μ = mean value of the population
σ^2 = variance of the population

Empirically, it is found that the distribution of results for replicate analyses data for most quantitative analytical experiments approaches that of the Gaussian curve. This equation describes a symmetrical distribution about the average value and an exponential decrease or increase in the magnitude of these deviations **(Figure 1)**.

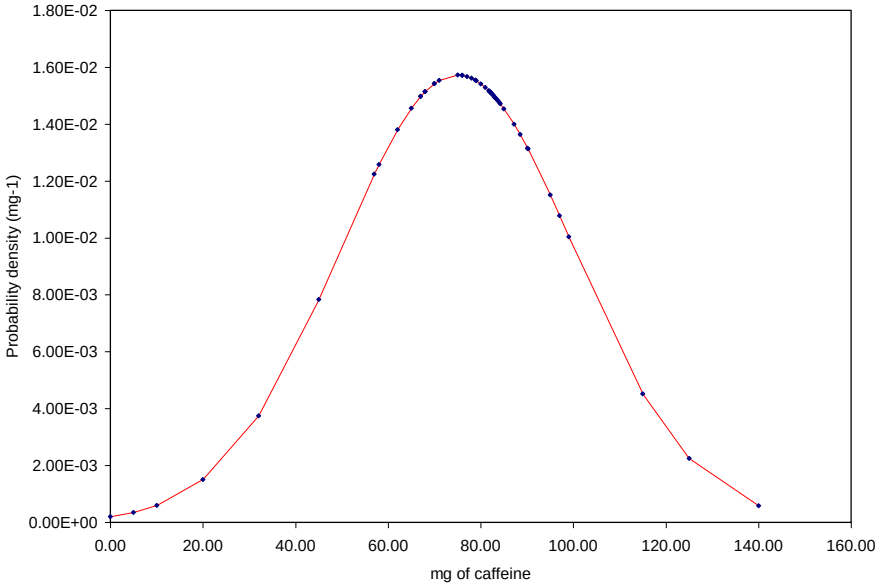


Figure 1: XY Scatter Chart of a Normal (Gaussian) distribution curve for the caffeine content in a commercial drug. In this type of chart the mean occurs at the center (single maximum value) of the curve. For a given interval, the area under the curve is directly proportional to the probability density of the event. A small value of the standard deviation (σ) narrows the curve, while a variation in the mean (μ) displaces the curve along the x-axis.

To obtain the average value of the population, or **population mean**, μ, the sum of all replicate values is divided by N, the total number of values in the set, with the provision that N approaches infinity.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \int_{-\infty}^{\infty} x f(x) dx \qquad (3)$$

In the **absence** of systematic errors the population mean is also the true value for the measured quantity.

The **variance, σ**, is the sum of the squares of the deviations from the mean divided by the total number of values N in a set:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \qquad (4)$$

where the deviations from the mean, d_i , are defined by :

$$d_i = X_i - \mu \qquad (5)$$

2. Analyzing a replicate set of data (less than 20 samples):

Performing an analysis with a large number of samples is impractical in terms of costs and time. Therefore, from three to six determinations (replicates) are typically performed in a chemical analysis. The results obtained are statistically evaluated by modifying the aforementioned equations for a finite set of measurements (see next section).
In a replicate analysis the average value of a series of determinations is expressed as the **sample mean,** which is defined as:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (6)$$

where $\overline{X}$ is the average of a series of determinations.

Another term used in the statistical treatment of data is the **median**. The median **is the middle value for a set of ordered data for which half of the data is larger in value and half is smaller.** For an odd number of results, the median can be evaluated directly, whereas for an even number of results, the median is the average of the two values that lie in the middle of the set.

Example 1: Calculate the mean and the median for the following data:
40.1, 40.2, 40.3, 40.5, 40.7, 40.9

$$\textbf{mean} = \overline{X} = \frac{40.1 + 40.2 + 40.3 + 40.5 + 40.7 + 40.9}{6} = \underline{40.5}$$

$$\textbf{median} = \frac{40.3 + 40.5}{2} = \underline{40.4}$$

a. Determining the precision in a replicate set of data:

In quantitative analysis the reliability of a method is described in terms of **precision** (the closeness of data to other data that have been obtained in exactly the same way) and accuracy (the closeness of a result to its true or accepted value). The precision is described based upon the deviation of the mean:

$$d_i = X_i - \overline{X} \qquad (7)$$

Five terms are commonly used to describe the precision in a replicate analysis: relative average deviation, standard deviation, relative standard deviation, variance and coefficient of variation. These terms are expressed as:

Relative average deviation:
$$\overline{d_i} = \frac{\dfrac{\sum_{i=1}^{N} X_i - \overline{X}}{N}}{\overline{X}} \times 10^{p} \qquad (8)$$

Where p is the power used to express the result as percent, p=2, parts per thousand, p=3, or parts per million, p=6.

Standard deviation:
$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N-1}} \qquad (9)$$

The standard deviation measures how closely the results are clustered about the mean value, or their average deviation of the data from it's mean value. Statistically, it defines the bounds about the mean within which 67% of the data in a Gaussian distribution can be expected. Note that when **s** is calculated, one degree of freedom is lost. Since **μ** is unknown, two quantities have to be determined from the set of replicate data, $\overline{X}$ and **s**. As one degree of freedom is used to determine $\overline{X}$, it must be subtracted in the determination of **s**.

NOTE: Standard deviations cannot be added, nor subtracted.

Relative Standard Deviation: $$RSD = \frac{s}{\overline{X}} \times 10^{p} \quad (10)$$

RSD provides an insight of the magnitude of the deviation relative to the average value. It is very useful in the construction of calibration curves and in instrumental analysis.

Variance: $$s^2 = \frac{\sum_{i=1}^{N} X_i - \overline{X}^2}{N-1} \quad (11)$$

It is particularly useful to determine the propagation of uncertainties.

Coefficient of Variation or Percent Relative Standard Deviation:

$$CV = \frac{s}{\overline{X}} \times 100 \quad (12)$$

It is equivalent to the RSD expressed as a percent.

Note from equations 7-12, that values close to zero denote high precision, while larger values denote poor precision.

Example 2: A chemist obtained the following results (in mg) for the determination of acetaminophen for Lot Number 44178 of Panadol gel caps: 500, 498,502, 497and 499. Calculate the mean, the median, the standard deviation, the variance and the percent relative standard deviation.

a. Determine the mean:

$$\overline{X} = \frac{497 + 498 + 499 + 500 + 502}{5} = 499.2 = \underline{499\,mg}$$

b. Determine the median:

median = 499 mg

c. Determine the standard deviation:

$$s = \sqrt{\frac{(497-499.2)^2 + (498-499.2)^2 + (499-499.2)^2 + (500-499.2)^2 + (502-499.2)^2}{5-1}}$$

s= ±1.92

d. Determine the variance:

variance = s^2 = **3.70**

e. Determine the RSD in % (CV):

$$RSD = \frac{1.923538}{499.2} \times 10^{2} = \underline{0.385\%}$$

b. Determining the accuracy of a replicate set of data:

The **absolute** and the **relative errors** can be used to determine the accuracy of a replicate set of data. The **absolute error is the difference between the mean value ($\overline{X}$) and the accepted value (μ):**

Absolute Error: $$AE = \overline{X} - \mu \quad (13)$$

NOTE: The absolute error bears a sign, indicating the direction of the deviation (positive or negative).

The relative error is the absolute error divided by the accepted value (μ) elevated to the corresponding power, in order to express it as percent, ppt, ppm, etc.:

Relative Error: $$RE = \frac{\overline{X} - \mu}{\mu} \times 10^{\,p} \quad (14)$$

NOTE: The relative error is the magnitude of the absolute error divided by the accepted value. Therefore, the relative error provides a notion of the magnitude of the error but not its direction.

Example 3: Using the results of Example 2, determine the absolute error and the relative error (in ppt) if the accepted value for the acetaminophen content in the drug is 500 mg.

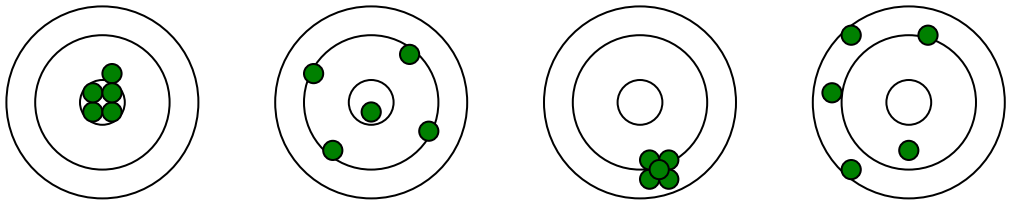
Since the average content of acetaminophen 499.2 mg:

$$AE = 499.2\ mg - 500\ mg = \underline{-0.800\ mg}$$

$$RE = \frac{499.2 - 500}{500} \times 10^{\,3} = \underline{-1.60\ ppt}$$

Note that the negative sign indicates that the experimental mean is below the expected value.

It is important to note that good precision does not necessarily imply good accuracy, and vice versa. Remember that statistical treatment of data assumes that results are only affected by random errors. Therefore, the presence of systematic and gross errors may alter either the precision or the accuracy of the analysis **(Figure 2).**



A B C D

Figure 2: Pattern of darts on a dartboard used to illustrate the effects of errors in the precision and accuracy of a quantitative analysis: A. good precision and accuracy; B. poor precision but good accuracy; C. poor accuracy but good precision; D. poor precision and accuracy.

REFERENCE:

- Harvey D. ***Modern Analytical Chemistry*****, 1st ed**. Chapter 4; McGraw-Hill, NY, **2000.**