# Statistics and Chemometrics for Analytical Chemistry

James N. Miller, Jane C. Miller
and Robert D. Miller

Pearson

Seventh edition

# Statistics and Chemometrics for Analytical Chemistry

Seventh Edition

# 4 The quality of analytical measurements

## Major topics covered in this chapter

- Sampling
- Quality control
- Control charts
- Proficiency testing schemes
- Method performance studies
- Uncertainty
- Acceptance sampling
- Method validation

## 4.1 Introduction

Chapter 1 showed that in analytical science quantitative studies predominate, so estimates of the inevitable errors are essential. The results of almost all analyses are supplied to a customer or user, and these users must be satisfied as far as possible with the **quality** – the *fitness for purpose* – of the measurements. This has important implications for analytical practice. First, any assessment of the measurement errors must take into account the whole analytical process – including the sampling steps, which often contribute to the overall error very significantly. Second, the performance of the analyses undertaken in each laboratory must be checked internally on a regular basis, usually by applying them to standard or reference materials. Third, in many cases the results from different laboratories must be compared with each other, so that the users can be satisfied that the performance of the laboratories meets statutory, regulatory and other requirements. Finally, the analytical results must be supplied with a realistic estimate of their uncertainty, i.e. the range within which the true value of the quantity being measured should lie. These are the major topics discussed in this chapter: they all contribute to the important area of method validation, an essential process which is summarised in Section 4.15. The statistical methods used are often very simple, most of them being based on techniques described in Chapters 2 and 3. But their regular

application has been a major advance in analytical sciences in recent years, with a large improvement in the quality and acceptability of many analytical results. Some of the methods discussed here have broader applications. For example, the principles used to monitor the performance of a single analysis in a single laboratory over a period of time can also be applied to the monitoring of an industrial process.

## 4.2 Sampling

In most analyses we rely on chemical samples to give us information about a whole object. Unless the sampling stages of an analysis are considered carefully, the statistical methods discussed in this text may be invalidated, as the samples studied may not be properly representative of the whole object under study. For example, it is not possible to analyse all the water in a stream for a toxic pollutant, and it is not possible to analyse all the milk in a tanker lorry to see if it contains a prohibited steroid hormone. Sometimes a small sample has to be used because the analytical method is destructive, and we wish to preserve the remainder of the material. So in each case the sample studied must be taken in a way that ensures as far as possible that it is truly representative of the whole object.

To illustrate some aspects of sampling we can study the situation in which we have a large batch of tablets and wish to obtain an estimate for the mean weight of a tablet. Rather than weigh all the tablets, we take a few of them (say ten) and weigh each one. In this example the batch of tablets forms the *population* and the ten weighed tablets form a *sample* from this population (see Section 2.2). If the sample is to be used to deduce the properties of the population, it must be what is known statistically as a **random sample**, i.e. a sample taken in such a way that all the members of the population have an equal chance of inclusion. Only then will equations such as Eq. (2.7.1), which gives the confidence limits of the mean, be valid. Note that the term 'random' has, in the statistical sense, a different meaning from 'haphazard'. Although in practice an analyst might spread the tablets on a desk and attempt to pick a sample of ten in a haphazard fashion, such a method could conceal an unconscious bias. The best way to obtain a random sample is by the use of a random number table. Each member of the population is allocated a number in such a way that all the numbers have an equal number of digits e.g. 001, 002, 003, etc. Random numbers are then read off from a random number table (see Table A.8), starting at an arbitrary point to give, for example, 964, 176, etc., and the corresponding members of the population form the sample. An alternative (and much simpler) method which is sometimes used is to select the population members at regular intervals, for example to take every hundredth tablet off a production line. This is not wholly satisfactory, however, since there might be a coinciding periodicity in the weight of the tablets: in that situation taking every hundredth tablet, which is not truly random sampling, would not reveal the true extent of the variations in weight. Similarly, if the last few tablets in the batch were taken and there had been a gradual decrease in weight during its production, then this sample would give a wholly misleading value for the mean weight of the batch.

In the tablet example the population is made up of obvious discrete members that are nominally the same. Sampling from materials for which this is not true, such as rocks, powders, gases and liquids, is called **bulk sampling**. If a bulk material were perfectly homogeneous then only a small portion or **increment** would be needed to determine the properties of the bulk. In practice bulk materials are non-homogeneous

for a variety of reasons. For example, ores and sediments consist of macroscopic particles with different compositions, and these may not be uniformly distributed in the bulk. Fluids may be non-homogeneous on a molecular scale owing to concentration gradients. Such inhomogeneity can only be detected by taking a number of increments from different parts of the bulk. If possible this should be done randomly by considering the bulk as a collection of cells of equal size and selecting a sample of cells by using random numbers as described above.

From the random sample, the mean, $\bar{x}$, and variance, $s^2$, can be calculated. There are two contributions to $s^2$: the **sampling variance**, $\sigma_1^2$, due to differences between the members of the sample, e.g. the tablets having different weights, and the **measurement variance**, $\sigma_0^2$, e.g. the random errors in weighing each tablet. For a bulk material, where, for example, a concentration was being measured, the sampling variance, $\sigma_1^2$, would be due to differences in concentration for increments taken from different parts of the bulk, and the measurement variance, $\sigma_0^2$, would be due to random errors in the measurement of the concentration for each increment. The next section describes how these two contributions can be separated and estimated by using ANOVA. For bulk materials the sampling variance is dependent on the size of the increment relative to the scale of the inhomogeneities: as the increment size increases, the inhomogeneities tend to be averaged out and so the sampling variance decreases.

## 4.3 Separation and estimation of variances using ANOVA

Sections 3.8–3.10 described the use of one-way ANOVA to test for differences between means when there was a possible variation due to a fixed-effect factor in addition to the measurement error. We now consider the situation where the additional factor is a *random-effect factor*, the sampling variation. In this case one-way ANOVA is used to separate and estimate the two sources of variation. Table 4.1 shows the results of the purity testing of a barrelful of sodium chloride. Five increments, A–E, were taken from different parts of the barrel chosen at random, and four replicate analyses were performed on each increment. As noted above, there are two possible sources of variation: that due to the random error in the measurement of purity, given by the measurement variance, $\sigma_0^2$, and that due to real variations in the sodium chloride purity at different points in the barrel, given by the sampling variance, $\sigma_1^2$. Since the within-increment mean square does not depend on the increment mean (Section 3.9) it can be used to give an estimate of $\sigma_0^2$. The between-increment mean square *cannot* be used to estimate $\sigma_1^2$ directly, because the between-increment variation is caused both by the random error in measurement and by the possible variation in the purity. It can be shown that the between-increment mean square gives an estimate of $\sigma_0^2 + n\sigma_1^2$, where $n$ is the

**Table 4.1** Purity testing of sodium chloride

| Increment | Purity (%) | Mean |
|---|---|---|
| A | 98.8, 98.7, 98.9, 98.8 | 98.8 |
| B | 99.3, 98.7, 98.8, 99.2 | 99.0 |
| C | 98.3, 98.5, 98.8, 98.8 | 98.6 |
| D | 98.0, 97.7, 97.4, 97.3 | 97.6 |
| E | 99.3, 99.4, 99.9, 99.4 | 99.5 |

number of replicate measurements of each increment, in this case four. However, before an estimate of $\sigma_1^2$ is made, a test should be carried out to see whether it differs significantly from 0. This is done by comparing the within- and between-increment mean squares: if they do not differ significantly then $\sigma_1^2 = 0$ and both mean squares estimate $\sigma_0^2$. The one-way ANOVA output from Excel for this example is shown below.

| A | B | C | D | E |
|---|---|---|---|---|
| 98.8 | 99.3 | 98.3 | 98.0 | 99.3 |
| 98.7 | 98.7 | 98.5 | 97.7 | 99.4 |
| 98.9 | 98.8 | 98.8 | 97.4 | 99.9 |
| 98.8 | 99.2 | 98.8 | 97.3 | 99.4 |

One-Way Anova

SUMMARY

| *Groups* | *Count* | *Sum* | *Average* | *Variance* |
|---|---|---|---|---|
| Column 1 | 4 | 395.2 | 98.8 | 0.006667 |
| Column 2 | 4 | 396 | 99 | 0.086667 |
| Column 3 | 4 | 394.4 | 98.6 | 0.06 |
| Column 4 | 4 | 390.4 | 97.6 | 0.1 |
| Column 5 | 4 | 398 | 99.5 | 0.073333 |

| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
|---|---|---|---|---|---|---|
| Between-groups | 7.84 | 4 | 1.96 | 30 | 5.34E-07 | 3.056 |
| Within-groups | 0.98 | 15 | 0.0653 | | | |
| Total | 8.82 | 19 | | | | |

It shows that the between-increment mean square is greater than the within-increment mean square: the *F*-test (a *one-sided* test, see Section 3.9) shows that this difference is very significant, i.e. $\sigma_1^2$ is significantly greater than 0. The within-increment mean square, 0.0653, is the estimate of $\sigma_0^2$, so we can estimate $\sigma_1^2$ using:

$$\begin{aligned}\sigma_1^2 &= (\text{between-increment mean square} - \text{within-increment mean square})/n \\ &= (1.96 - 0.0653)/4 \\ &= 0.47\end{aligned}$$

## 4.4 Sampling strategy

If one analysis is made on each of $h$ increments the confidence limits of the mean are given by Eq. (2.7.1):

$$\mu = \bar{x} \pm t_{h-1} s/\sqrt{h} \qquad (4.4.1)$$

where $\bar{x}$ is the mean of the $h$ measurements and $s^2$ is the variance of the measurements. The total variance, $\sigma^2$, is estimated by $s^2$ and is the sum of the measurement and sampling variances, i.e. $\sigma_0^2 + \sigma_1^2$ (see Section 2.11.1): $\sigma^2/h$ (estimated by $s^2/h$) is the variance of the mean, $\bar{x}$. If the value for each increment is the mean of $n$ replicate measurements, then the variance of the mean is $(\sigma_0^2/n + \sigma_1^2)/h = \sigma_0^2/nh + \sigma_1^2/h$. Obviously, for maximum precision, we require the variance of the mean to be as small as possible. The term due to the measurement variance, $\sigma_0^2/nh$, can be reduced either by using a more precise method of analysis or by increasing $n$, the number of replicate measurements. However, there is no point in striving to make this measurement variance much less than (say) a tenth of the sampling variance, as any further reduction will not greatly improve the total variance, which is the sum of the two variances. Instead it is preferable to take a larger number of sample increments, since the confidence interval decreases with increasing $h$. If a preliminary sample has been used to estimate $s$, then the sample size required to achieve a given size of confidence interval can also be estimated (see Chapter 2, Example 2.6.1).

A possible sampling strategy with bulk material is to take $h$ increments and *blend* them before making $n$ replicate measurements. The variance of the mean of these replicate measurements is $\sigma_0^2/n + \sigma_1^2/h$. This total variance should be compared with that when each increment is analysed $n$ times and the increment means are averaged, the variance then being $\sigma_0^2/nh + \sigma_1^2/h$ (see above). Obviously the latter variance is the smaller, resulting in greater precision of the mean, but more measurements ($nh$ against $h$) are required. Knowledge of the values of $\sigma_0^2$ and $\sigma_1^2$ from previous experience, and the costs of sampling and analysis, can be used to calculate the relative costs of these sampling strategies. Improving the precision of the measurements to an unnecessary extent by increasing $n$ and/or $h$ will incur greater costs in terms of time, equipment use, etc., so in general the most economical scheme to give the required degree of precision will be used.

For bulk materials the sampling variance depends on the size of the increment relative to the scale of the inhomogeneities and decreases with increasing increment size. In some experiments it may be necessary to set an upper limit on the sampling variance so that changes in the mean can be detected. Preliminary measurements can be made to decide the minimum increment size required to give an acceptable level of sampling variance.

The heterogeneity of many materials encountered in analytical practice, and therefore the importance of using suitable sampling protocols, is closely related to the important topic of *sampling uncertainty*, which is considered in more detail in Section 4.13.

## 4.5 Introduction to quality control methods

If a laboratory is to produce analytical results of a quality that is acceptable to its clients, and allow it to perform well in proficiency tests or method performance studies Sections 4.11 and 4.12, it is obviously essential that its results should show excellent consistency from day to day. Checking for such consistency is complicated by the inevitable occurrence of random errors, so several statistical techniques have been

developed to show whether or not time-dependent trends are occurring in the results, alongside the random errors. These are referred to as **quality control** methods.

Suppose that a laboratory uses a chromatographic method for determining the level of a pesticide in fruits. The results may be used to determine whether a large batch of fruit is acceptable or not, and their quality is thus of great importance. The performance of the method will be checked at regular intervals by applying it, with a small number of replicate analyses, to a standard reference material (SRM), in which the pesticide level is certified by a regulatory authority. Alternatively an internal quality control (IQC) standard of known composition and high stability can be used. The SRM or IQC standard will probably be inserted at random into the sequence of materials analysed by the laboratory, so that the IQC materials are not separately identified to the laboratory staff, and are studied using exactly the same procedures as the routine samples. The known concentration of the pesticide in the SRM/IQC materials is the **target value** for the analysis, $\mu_0$. The laboratory needs to be able to stop and examine the analytical method quickly if it seems to be giving erroneous results. On the other hand, time and other resources will be wasted if the analyses are halted unnecessarily by false alarms, so the quality control methods should allow their continued use as long as they are working satisfactorily. If the values for the IQC samples do not show significant time-dependent trends, and if the random errors in the measurements are not too large, the analytical process is said to be *under control* or *in control*.

Quality control methods are also very widely used to monitor industrial processes. Again it is important to stop a process if its output falls outside certain limits, but it is equally important not to stop the process if it is working well. For example, the weights of pharmaceutical tablets coming off a production line can be monitored by taking small samples of tablets from time to time. The tablet weights are bound to fluctuate around the target value $\mu_0$ because of random errors, but if these random errors are not too large, and are not accompanied by time dependent trends, the process is under control.

## 4.6 Shewhart charts for mean values

In Chapter 2 we showed how the mean, $\bar{x}$, of a sample of measurements could be used to provide an estimate of the population mean, $\mu$, and how the sample standard deviation, $s$, provided an estimate of the population standard deviation, $\sigma$. Similar principles can be applied to quality control work, but with one important difference. Over a long period, the *population* standard deviation, $\sigma$, of the pesticide level in the fruit (or, in the second example above, of the tablet weights), will become known from experience. In quality control work, $\sigma$ is called the **process capability**. Equation (2.6.1) shows that 95% of the values of the sample mean, $\bar{x}$, lie within the interval $\mu \pm 1.96\sigma/\sqrt{n}$. Similarly 99.7% of the values of $\bar{x}$ lie in the interval $\mu \pm 2.97\sigma/\sqrt{n}$. In practice, the values of $z$ are often rounded to give the following equations for the construction of control charts.