

---

# Statistics for food science II: chemical analysis data

---

*John A. Bower*

---

## The author

John A. Bower is a Lecturer in Food Science in the Department of Applied Consumer Studies, Queen Margaret College, Edinburgh, UK.

---

## Abstract

Introduces some simple statistics employed in analysing chemical analysis data. Describes measures of precision and accuracy and how the use of confidence intervals and repeatability can guide validation of data.

In part two of this series we look at some simple statistical measures which can be applied to data obtained by chemical analysis of foods. Procedures similar to these form part of an assessment of the validity of any analysis or measurement system, a topic of major importance in modern research and industry. The Department of Trade and Industry's VAM (valid analytical measurement) project centres on constant improvement of analytical standards. With increasing emphasis on quality assurance, safety and legal issues, many laboratories involved in chemical analysis of foods have sought a quality standard such as that provided by NAMAS (National Measurement Accreditation Service), guided by VAM advice. The statistical procedures are but part of the whole picture and they provide analysis techniques for the methodologies of validation, one of the key features being measurement of uncertainty in chemical analysis. Only a limited indication is given here for the general food scientist who is involved in chemical analysis – full protocol methods are detailed in the references.

Nowadays, chemical analysis can be very sophisticated but similar key elements of measurement still apply. For the purposes of illustrating the statistics, typical data from the determination of crude protein via Kjeldahl nitrogen, a method well known by food scientists, will be used. The results of two quality control determinations, at different levels of replication on the same food product, are shown in Table I. Assume that manual techniques were used throughout and that the most probable value was obtained from the production unit's knowledge of the ingredients.

## Were there any errors in the measurements during the analysis?

### Errors and measurement uncertainty

The term, "experimental error" is used extensively in student lab books to account for all manner of unexpected results. While this may be appropriate the error can be allocated to a number of possible sources which can usually be identified as discussed below.

*Gross* errors (e.g. a misread balance or grossly incorrect additions /omissions of reagents) are usually *accidental* in nature and with care they can be avoided. In the Kjeldahl analysis an obvious gross error would be seen if there was omission of the catalyst for one of the replicates. Rejection of that value could be considered and

Table I Quality control laboratory data for percentage of crude protein analysis on food product

Replicate number	Percentage of protein Analysis A	(N <sub>2</sub> × 6.25) Analysis B
1	7.3	8.4
2	8.5	9.1
3	–	8.7
4	–	8.2
Mean	7.9	8.6

Note: True/most probable value = 8.8 per cent

there are statistical tests for such “outlier” values. Thus these errors may not affect all measurements in a set and often can be easily detected. Other types of error occur even when the greatest care is taken.

*Systematic* errors (e.g. a balance which requires servicing and calibration, unrecognized faulty technique by the analyst, or a method-related systematic error) usually affect all the analyses in a similar manner. The systematic error effect is also known as bias and affects accuracy. Note that even if a balance is calibrated (i.e. set to weigh accurately using certified weights) it may still give an inaccurate reading if the balance model is unable to read beyond a certain level. Thus the lack of calibration is a determinate error, and can be changed, but the other is constant. Calibration improves accuracy and reduces or removes any bias which instruments may have. A blank determination is another aid to detection of a systematic error.

Another source of error is detected if the test sample was analysed more than once. Even if gross and systematic errors are absent, repeated measurements may show some variation. These are caused by *random* errors, e.g. small errors in weighing, use of volumetric devices and other analysis instrumentation. Even highly trained analysts using top of the range equipment may be unable to avoid random error. The random error effect in a series of measurements causes the individual results to fall on either side of the mean. They may be accidental in nature but are indeterminate as they are difficult to remove entirely. Random errors affect the precision of the analysis method.

These errors can occur at any stage of the analysis and accumulate to produce the overall error. Some errors augment one another whereas others may cancel one another out. The replicate values in Table I are all different and possible error sources could be deduced by examination of each stage of the Kjeldahl analysis. An estimate of error magnitude in the final results can now be calculated.

## Are the data items accurate and precise?

These terms were introduced in part I but they require further explanation.

### Accuracy and precision in measurement

Accuracy is the extent of agreement between the determined value and the true or most probable value; and precision is the extent of agreement among a series of measurements of the same quantity.

It is important to note that with these terms the presence of one does not automatically imply the other: a high degree of precision does not imply accuracy and vice versa.

### Measures of accuracy

The degree of concordance with the true value can be calculated as the error of the mean (see “Statistics for food science I”) which can also be expressed as the relative error of the mean (REM):

$$EM = M - T \quad \% \text{ REM} = \frac{EM \times 100}{T}$$

where:

- EM = error of the mean
- T = “true” or “actual” value
- M = mean value.

As explained in part I, the true value may not be available for unknown samples, unless an independent analysis has been performed giving a confident estimate. If an indication only is required then a rough estimate can be given by “typical” values from text books and/or food product labels. In the food production situation (Table I), the true expected value can be calculated for quality control purposes from a knowledge of the chemical composition of the specified ingredients. Alternatively a standard or control material of known composition can be analysed along with the unknown under the same analysis conditions, thus enabling the above calculation. A suitable crude test material can be made up by the analyst, formulated from constituent chemicals or purified constituents. Another possibility used in some laboratories is use of a previously analysed material which is kept in stable storage and sampled along with the new samples. For more critical circumstances a CRM (certified reference material) would be required. While many RMs are available, not all common constituents such as nitrogen are found in the certified lists and the matrix (i.e. the physical and chemical “make-up”) of the RM may be different from the unknown food sample. There have been some

developments to answer these food specific requirements, e.g. the FAPAS (food analysis performance assessment scheme) initiative run by MAFF (Ministry of Agriculture, Fisheries, and Food) which has food product test materials for proximate analyses such as nitrogen protein. The use of a hierarchy of reference standards from secondary RMs to certified RMs and ultimately primary RMs forms part of the traceability chain for chemical composition instigated by VAM.

The presence of errors will affect the magnitude of the percentage REM obtained. Assuming the absence of gross and systematic errors then a percentage REM of zero is possible but unlikely due to random errors. Usually negative or positive percentage REM values are obtained representing results which are below or above the true value respectively. These statistics can now be calculated for the data of Table I. As could be easily deduced by inspection of the mean values, both analyses have underestimated percentage protein, and the magnitude of this is shown (Table II) by the negative percentage REMs. Analysis B has a greater agreement with the most probable value.

**Measures of variability (precision)**

The standard deviation (SD) and the mean absolute deviation (MD) introduced previously are measures of precision. These can be standardized as the percentage coefficient of variation (%CV; also known as the relative standard deviation) and the percentage relative mean deviation (% RMD) respectively:

$$\% CV = \frac{SD \times 100}{M}$$

$$\%RMD = \frac{MD \times 100}{M}$$

where

SD = standard deviation

M = mean

MD = mean deviation.

These measures are related (MD is approximately 0.8 times SD). Both are included here as MD is perhaps easier to understand and calculate. In the form above, erroneous com-

parisons between data sets possessing different measurement scales are avoided, e.g. an MD of ten for a mean of 10,000 gives a very low percentage RMD (0.1 per cent), but with the same MD for a mean of 100 the RMD is very high (10 per cent).

Two other related measures are important. Repeatability is the precision obtained when a method of analysis is repeated under the same conditions, i.e. by the same analyst using the same equipment, on the same sample material, etc. (also referred to as “within laboratory” or “within run” precision). The analyses in Table I can be assumed to have been done under repeatability conditions.

Reproducibility is the precision obtained when the same method of analysis is repeated on the same test material but under different conditions, i.e. a different analyst, different set of equipment or a different laboratory or even a different method (also known as “between run” or “between laboratory” precision).

Fuller statistical definitions of these terms, as would be required in interlaboratory proficiency testing schemes, are detailed in [1,2]. It is usual to find that repeatability conditions result in greater precision than those of reproducibility. In fact the poor reproducibility shown by different laboratories when analysing the same samples was one of the reasons for instigating the VAM project.

The magnitude of the percentage CV (or percentage MD) will range from zero upwards. “Perfect” precision would produce a CV percentage of zero and although this can occur, more commonly small values are obtained, caused by random error. Large percentage CV values may point to gross errors. Note that even if the method is perfectly precise, repeated values could still vary owing to inherent variation within the food material itself. Calculation of precision for the data of Table I shows (see Table III) that precision is relatively poor in set A (high %CV, %RMD values).

Pertinent to these measures is the number of repeated measurements.

**Were the analyses done at an acceptable level of replication?**

The level of replication is an important consideration as it affects the statistical measures and the cost of the analysis in terms of time and personnel. In practice the costs can limit the degree of replication. For routine analyses with established techniques, modern instruments

Table II Accuracy measures for percentage of protein data

	Analysis A	Analysis B
Number of replicates	2	4
Mean (%)	7.9	8.6
% REM	-10.2	-2.3
<i>Note:</i> Most probable value = 8.8 per cent		

Table III Precision measure for percentage of protein data

	Analysis A	Analysis B
Number of replicates	2	4
Mean (%)	7.9	8.6
Range (%)	1.2	0.9
MD	0.60	0.30
SD	0.85	0.39
%RMD	7.60	3.50
%CV	10.74	4.55

and trained analysts, minimal replication may be common, except where the technique is very rapid and low in cost, e.g. as with modern nitrogen analysers based on the Dumas method (2.5 minutes per sample). Thus duplicate determinations or even a single one done along with a reference or standard analysis for the run may be typical. If a single determination is made there is no reference point for error detection. Statistically, the greater the number of determinations the more reliable or accurate the result. Whether or not a low level of replication is acceptable depends on several factors: the experience of the analyst and the laboratory itself; the method of analysis and its history with respect to the food in question; and the importance of the decisions which are to be based on the results.

Certainly, low levels of replication in isolation provide a weak basis for making confident decisions regarding the data obtained, e.g. a standard deviation based on only two values is an extremely shaky foundation on which to base further inferences. The difference in magnitude between the SD values (Table III) for two and four replicates, for data sets with similar ranges, illustrates this point. This does not, however, preclude the routine use of duplicates, as will be explained later.

The final consideration is how to use the calculated measures (Tables II and III) to answer questions concerning the acceptability of the obtained levels of precision and accuracy.

**Is the level of precision acceptable?**

**Acceptance level for precision**

The deviation of a set of replicates around the mean depends on the precision of the measurement system and on the degree of variability of the population from which the samples originate. If both are of a completely unknown nature then whether or not to accept a set of replicates cannot be decided easily. Some mea-

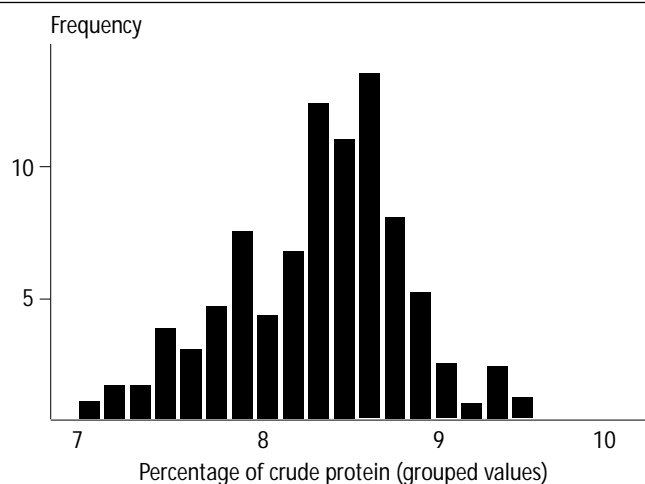
sure of variability must be established. This can be done by carrying out an initial set of a larger number of replicates than is envisaged for routine use, e.g. at least ten, or if appropriate, by proceeding with duplicate analyses without considering variability until a “data bank” of typical values has been established from which an estimate of deviation in the form of the standard deviation can be calculated, i.e. a comment concerning the “expected variation” for a set of replicates cannot be made until some measure of variability has been established.

Once this is available then an error estimate, known as a confidence interval (CI) can be calculated for the population mean of the measurement. It gives a region within which we are confident that the population mean will be located, with a specified probability or “certainty” level. This statistic can be used as an estimate of bias (accuracy) and the width of the interval gives another perspective on precision, as it emphasizes the effect of sample size.

To understand a confidence interval we need to appreciate the nature of a population distribution. Put simply, if we know how a population is “mapped out” then it can be used to make estimations based on samples taken from that population. Imagine that the food product (Table I) is analysed a very large number of times for crude protein content and grouped values are plotted on a histogram – then it is likely that a rough inverted cone shape would be obtained (see Figure 1).

Increasing the number of points would have a smoothing effect on the shape and with a very large number a bell shaped curve would be obtained. Ultimately with an infinitely large number of values the curve would be smooth

Figure 1 Frequency distribution of 100 per cent crude protein content determinations



and would represent the population distribution for the measurement. Note that these measurements are all of the same constituent on the same material, using the same technique, etc. The curve shape would be typical of a normal distribution (see Figure 2) and similar distributions, “normal” in this case meaning “standard”. The curve has certain properties which allow powerful inferential statistics to be performed – the mean ( $\mu$ ), mode and median are centrally located; on either side of the centre the two “tails” are of such shape that more values are clustered towards the centre than at the edges; in terms of variation the proportion of the curve at one or more standard deviations ( $\sigma$ ) from the mean can be marked and measured.

It can be seen that when selecting a random sample from such a population, there is a higher probability of obtaining a percentage protein value within  $\pm 1$  standard deviation of the population mean than further away, as there is more area under that region of the curve. In fact approximately 68 per cent of all the values lie in this region, and approximately 95 per cent lie within  $\pm 2$  standard deviations.

Most chemical and physical measurements on food samples are likely to come from a normal population. Even if the parent population deviates from normality, statistical theory proves that the distribution of the means of samples from such a population will approximate to normality. Thus this distribution will also possess the above properties and provides the basis for determining the confidence interval for the population mean based on the sample mean.

Large sample sizes provide adequate estimates of the population parameters to allow calculation of the confidence interval using the

proportions described above. For small samples of the order likely to be used in chemical analysis a more appropriate distribution “standard” for making estimates is the *t*-distribution – it is similar in shape and characteristics to the normal distribution but is wider and flatter, having more “spread” (especially for small numbers of samples or replicates). Thus the interval will be wider, reflecting the increased uncertainty.

A measure of the degree of confidence must be specified and it is expressed on a probability scale of zero to 100 per cent, with 100 per cent representing absolute certainty. Unfortunately, choosing the 100 per cent level of confidence would result in an interval of very large width, unusable in practical situations. Usually the 95 or 99 per cent limit is selected, representing high degrees of confidence. The confidence interval limits are calculated using the *t*-value from the *t*-distribution based on the number of replicates:

$$95\% \text{ CI} = M \pm t \times SD/\sqrt{n}$$

where

$n$  = number of replicates.

The value of *t* is obtained from statistical tables and its magnitude depends on the confidence level and on the number of samples analysed (more specifically on the degrees of freedom, which is equal to the number of samples minus 1). Thus a high confidence level combined with low replication would maximize the *t*-value and the interval width and vice versa. These calculations can now be done for the data of Table I and are summarized in Table IV.

Thus, assuming no systematic error for Analysis B, the analyst would be confident that 95 per cent of the time, the population mean for percentage of crude protein content would lie between 8.0 and 9.2 per cent. The width of the interval can guide the acceptability of precision. Whether or not it is acceptable depends in turn on how confident the analyst is in the validity of the SD. In set A it is based on only two determinations and gives a confidence interval of 15.2 per cent owing to the large *t*-value and the large SD – in isolation such a wide interval would be

Figure 2 A normal distribution frequency curve

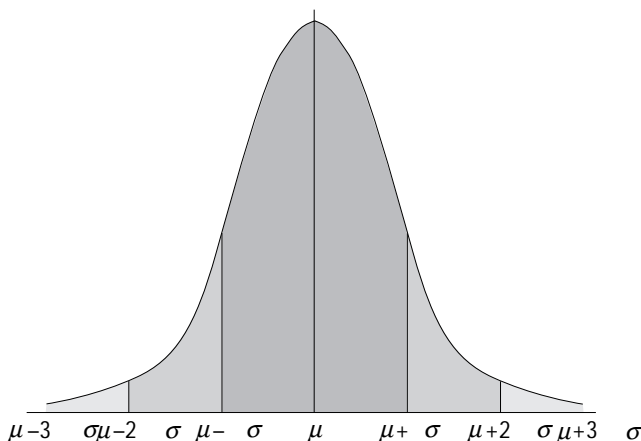


Table IV Confidence intervals for percentage of protein data

	Analysis A	Analysis B
Number of replicates	2	4
Mean (%)	7.9	8.6
Range (%)	1.2	0.9
SD	0.85	0.39
CI (95%)	0.3-15.5	8.0-9.2
<i>t</i> -value (95% confidence)	12.71	3.18

Table V Repeatability for crude protein by Kjeldahl nitrogen

	Analysis A	Analysis B	Analysis C
Number of replicates	2	4	10
SD	0.85	0.39	0.35
Range (%)	1.2	0.9	–
<i>t</i> -value (95% confidence)	12.71	3.18	2.26
Repeatability (95 per cent) value (based on Analysis C) = 1.1 per cent			

unacceptable. Analysis B, using four replicates, cuts the interval to 1.2 per cent, obviously more acceptable. This fact seems to condemn low replication but if a confident SD is established initially by a larger number of replicates, or on a series of a least six duplicate analyses [1], then this can be used to calculate a useful statistic, a form of repeatability [1,2] for subsequent analysis with two replicates:

$$r = t \times (\sqrt{2}) \times SD$$

where

*r* = estimated variability or repeatability which must not be exceeded;

*t* = value from table based on the larger original number of initial analyses;

SD = standard deviation of original number of repeat determinations under repeatability conditions.

Assuming such circumstances, an additional analysis based on ten crude protein determinations is given below (Table V) along with the calculated *r* statistic.

The *t*-value is smaller as it is based on the original ten determinations. Thus we would expect duplicate crude protein determinations to differ by less than 1.1 per cent, so although Analysis A looks more favourable now it could still be rejected on these grounds. Indeed, in the author's experience of the manual Kjeldahl technique on a range of food products, the precision of Analysis B (or better) is more typical and it is likely that a gross error has occurred in Analysis A.

Following the above procedure now gives a more definite guide to accepting the level of precision.

### Is the level of accuracy acceptable?

#### Acceptable level of accuracy

The magnitude of the percentage of REM or the EM will decide this, but how large should it be before it is regarded as unacceptable? If it is based on comparison with typical values then these can vary by up to 10 per cent or

more and this must be borne in mind when gauging accuracy via crude methods. Similarly "most probable" estimates are also approximations. Analysis B (Table II) is within 10 per cent of the estimated true value whereas Analysis A exceeds this limit. The confidence interval detailed above as a precision check can also be used for accuracy, provided that a confident measure of the SD was obtained – if the expected value lies within the interval then this is acceptable. In the example (Table II), both analyses achieve this level of acceptance, but the Analysis A result is rejected because of the very large interval. Determination of crude protein is a proximate technique, and accuracy much beyond that of Analysis B may be an unrealistic target. For a certified RM a similar procedure can be applied – the determined interval should contain the certified analyte value. Additionally an interval will be quoted on the certificate – the mean value obtained by the analyst should lie within this interval. This is a more stringent test as the certificate interval is likely to be narrower. In either case, if the determined mean is outwith the interval then the result can be viewed as inaccurate and this may indicate the presence of a systematic error.

Depending on circumstances, there is some leeway in the decision-making process and individual analysts can decide on acceptable proximity to the precision and accuracy levels. Textbooks on analytical methods may not quote figures for acceptable accuracy and precision. Often it is left to the experience and knowledge of the analyst.

### References

- 1 Association of Public Analysts, *A Protocol for Analytical Quality Assurance in Public Analysts' Laboratories*, Association of Public Analysts, London, 1986.
- 2 Calcutt, R. and Boddy, R., *Statistics for Analytical Chemists*, Chapman & Hall, London, 1983.

### Further reading

- Miller, J.C. and Miller, J. N., *Statistics for Analytical Chemistry*, 3rd ed., Ellis Horwood, Chichester, 1993.
- Rees, D.G., *Essential Statistics*, 2nd ed., Chapman & Hall, London, 1991.
- Ryan, P. and O'Donoghue-Ryan, F., "Balancing values", *Laboratory Practice*, Vol. 38 No. 7, 1989, pp. 23-7.
- Ryan, P. and O'Donoghue-Ryan, F., "Weighing up accuracy and precision", *Laboratory Practice*, Vol. 38 No. 6, 1989, pp. 29-33.