A simple example of sampling uncertainty evaluation

In analytical chemistry, the test sample is usually only part of the system for which information is required. It is not possible to analyze numerous samples drawn from a population. Hence, one has to ensure a small number of samples taken are representative and assume that the results of the analysis can be taken as the answer for the whole.

Analysis of Variance (ANOVA) can be used to check this assumption, and to determine the variation in the test portions chosen and the contribution to the variation of the measurement process.

The total analysis variance is given by

$$\sigma^2_{measurement} = \sigma^2_{sampling} + \sigma^2_{analysis} \qquad \ldots[1]$$

where $\sigma^2_{sampling}$ is the variance due to actual differences between the samples and $\sigma^2_{analysis}$is the variance in making the measurement in the laboratory.

From equation [1], it is obvious that sampling uncertainty has to be evaluated together with the analytical uncertainty for a complete measurement uncertainty evaluation. Without carrying out repeated testing on the samples given, one will not be able to assess the repeatability or precision of the analyte concentration in the laboratory sample. If only a single analysis is carried out on each of the samples drawn from a population, the end result shows the sampling precision only.

The following example demonstrates the basic principle of one-way (or one-factor) ANOVA and how it is applied to evaluate the overall measurement uncertainty covering both sampling and analysis uncertainties. In this case, possible contributions of biasness in both sampling and analysis are not considered. If necessary, variances of these biases determined separately can be added to the equation [1].

Example

Three composite samples were taken at the top, middle and bottom of a grain silo during loading for analysis. The laboratory samples were sub-sampled for the determination of Kjeldahl nitrogen in 4 replicates and reported as the % crude protein by multiplying a conversion factor of 5.71. The results are given in table 1 below.

Table 1: Analysis of grain taken from different levels in a grain silo

	Silo position		
Repeat #	Top	Middle	Bottom
1	12.3	13.4	13.2
2	12.7	12.8	13.5
3	11.8	13.6	13.1
4	12.2	13.0	12.9

(1) Does the sampling procedure have a significant effect on the results at the 95% confidence level?
(2) If so, what are the standard uncertainties (expressed as standard uncertainties) in sampling and in analysis?
(3) What would the standard uncertainty of measurement expected if we were to make single measurements taken at random from anywhere in the silo?

Solution

Using factor as 'sampling position', the test data can be treated by a one-way ANOVA method which can be carried out either by its basic principles or by Excel spreadsheet.

Calculations by first principle

The average and standard deviation of each level sample are tabulated in Table 2, keeping reasonable number of decimal points for accurate calculations.

Table 2: The means and standard deviations of samples

	Silo position		
	Top	Middle	Bottom
Mean $\bar{x}_i$	12.250	13.200	13.175
Std Dev, s_i	0.3697	0.3651	0.2500

Upon calculation, the mean of the sample means, $\bar{\bar{x}}$ was found to be 12.875 with standard deviation $s_{\bar{x}}$= 0.5414.

Use equation [2] to calculate the sum of squares (between-samples), SS_b:

$$SS_b = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2 \qquad \ldots [2]$$

where n_i is the number of repeats for each sample.

Hence, SS_b = 2.345 with 2 degrees of freedom, df_b (i.e. 3 samples – 1), leading to mean square (between-samples), $MS_b = SS_b/df_b$ = 1.1725.

Use equation [3] to calculate the sum of squares (within-sample), SS_w:

$$SS_w = \sum(n_i - 1)s_i^2 \qquad \ldots [3]$$

Upon calculation, SS_w = 0.9975 with 9 degrees of freedom, df_w (3 samples x 4 repeats – 3 samples), leading to mean square (within-sample), $MS_w = SS_w/df_w$ = 0.1108.

Note: An alternative way to calculate the *df* for within-sample is first to find the total degrees of freedom of whole set of data which is 3 samples x 4 repeats *minus* 1, giving *df* of 11, and then minus the degrees of freedom for between samples which is 2. The end result is the same.

So, in summary, we have:

Between-sample

SS_b = 2.345; df_b = 2; MS_b = 1.1725

Within-sample

SS_w = 0.9975; df_w = 9; MS_w = 0.1108

To check the significance of between-sample variation against the within-sample variation, the F-statistic test was carried out as follows:

$F = \frac{MS_b}{MS_w}$ = 10.579 which is larger than the critical F value of 4.256 at α = 0.05, df_b = 2, df_w = 9, indicating that the variance in the sampling does have a significant effect on the overall measurement variance at 95% level.

Now, the within-sample mean square allows estimation of the repeatability of the measurement and so,

$$s_w = s_r = \sqrt{(0.1108)} = 0.33 \text{ \% protein}$$

Note that the parameter s_r is an estimate of $\sigma_{analysis}$, the standard deviation of the laboratory analysis.

The between samples mean square (MS_b) is an estimate of the combination of the analysis variance and the variance due to the different sampling positions, i.e., $MS_b = MS_w + n \times MS_{sampling}$ where n = 4 repeats in this case.

Therefore, $s_{sampling} = \sqrt{MS_{sampling}} = \sqrt{\frac{MS_b - MS_w}{n}} = 0.515$

It follows that the variance of a single analysis is given by

$$\sigma^2_{measurement} = \sigma^2_{sampling} + \sigma^2_{analysis} = 0.515^2 + 0.333^2 = 0.376$$

and the combined standard uncertainty expressed as combined standard deviation $u_{measurement} = \sigma_{measurement}$= 0.61(3)% protein.

Since the overall mean of this exercise $\bar{\bar{x}}$ was 12.88% protein, the reporting format is:

12.88 $\pm$ 2(0.613) or 12.88 $\pm$ 1.23% protein with a coverage factor of 2

ANOVA calculations by MS Excel spreadsheet

Results of the above one-way ANOVA calculations by basic principles can be easily verified by using the Data Analysis Toolpak software in the MS Excel spreadsheet. The subsequent data analyses for uncertainties are the same as above. The outputs of its single-factor ANOVA on the data given in Table 1 are as follows:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Top	4	49	12.25	0.136666667
Middle	4	52.8	13.2	0.133333333
Bottom	4	52.7	13.175	0.0625

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.345	2	1.1725	10.579	0.004	4.256
Within Groups	0.997	9	0.11083			
Total	3.343	11				