Use of replicate calibration samples in analytical chemistry: uncertainties due to lack of knowledge of heteroscedasticity

Summary

Chemical and pharmaceutical assay procedures are subject to random and non-random measurement errors, and therefore the analytical result has to be reported as the most likely (probable) value. When taking account of the random errors, the analyst relies on textbook assurances that replicate instrumental responses (signals) for a given amount have an approximately normal (Gaussian) distribution, perhaps after a suitable transformation. While the assurances are not fully justified, we can presume that pharmaceutical and other regulatory requirements leave an adequate margin for such imponderable uncertainties. Nevertheless, it is good practice to attempt to understand all the factors that can be taken into account when estimating the most likely value for an assay.

A further pre-condition for the calculation of the most likely result is that, unless the variance of the instrumental response is constant within the working range, appropriate statistical weighting must be applied. Textbooks discuss the application of weighted regression to calibration curves, but the consequences of heteroscedasticity (non-constant variance) do not seem to have been fully presented for the case, of regulatory importance, in which a regression analysis as used a means of validating the routine use of single-point calibration.

In practice, single point calibrations are usually done in replicate. We show that failure to take heteroscedasticity into account can have a greater effect on the uncertainty budget of the analytical result than that allowable for the imprecision of the balance. We also show that the algorithms most commonly used for method calibration incorporate, implicitly, statistical weightings that are likely to be incorrect. Uncertainty due to heteroscedasticity can be reduced to acceptable levels by one or both of the following approaches:

1. Apply appropriate statistical weightings, which implies that the coefficient of heteroscedacity must be known approximately. An estimate of the coefficient can be obtained during method validation, though it may be sufficient to adopt a "consensus" value in the more usual situations. Difficulties may be encountered, however, notably when samples are difficult to weigh.
2. Specify a working range that is narrower than those currently given in pharmacopeias and other compendial texts. It seems likely that this will be achievable with the introduction of new weighing technology.

Finally, while chemical analysis is now almost entirely dependent of sophisticated computer-controlled automated instrumentation, we note that some aspects have not changed much since Lavoisier lost his head. In particular one would expect to see adaptive and iterative assay protocols which would eliminate the need for some possibly questionable but obligatory items of method validation protocols. This is more than an economic argument; no practicing analyst would consider a validation exercise to be representative of everyday conditions.

Introduction

Assays of drug substances, intermediates and some other fine chemicals are generally carried out using methods (typically chromatographic) that have to be calibrated by means of a reference substance. Textbooks and guidelines on statistics for chemical or pharmaceutical analysis frequently present calibration theory in terms of (inverse) regression analysis, whereby a calibration curve is prepared on or shortly before the day of the analysis, and the uncertainties of the results are calculated from the calibration data. Preparing and running at least six independent samples of a precious reference substance is time-consuming. A point not often discussed is that the effect of response drift on measurement uncertainty is increased when the number of standards is increased. We commented on the question of drift when presenting a new analytical method (Lee *et al*., 2003) because, although it was a major source of uncertainty, we were unable to find any relevant literature references. The subject is mentioned briefly in a IUPAC technical report that was published in 2002.

Normal practice for assays is to validate the procedure once or periodically, using the full calibration curve. This should be done by more than one analyst, using more than one instrument set. If the response is linear, a single calibration point may be used each time the method is applied, the uncertainty of the result being calculated from the validation data. In practice, the "single point" calibration standard is prepared in replicate.

We draw attention to a source of uncertainty associated with the use of replicate standards, which is encountered whenever the amounts taken are not identical. The cause is uncertainty concerning the dependence of measurement uncertainty on the amount of reference standard. Calculations show that there is a significant effect on the uncertainty of drug substance assays carried out according to commonly-applied protocols.

Analytical background

Most of the techniques used in pharmaceutical and other regulated analytical laboratories are chosen so that they can be relied upon to give a strictly linear response (straight line through the origin), enabling the use of single-point calibration. The examples given in the Eurachem Guide (2000) are of this kind, and linear least squares regression is only briefly mentioned in an appendix. Methods that have proved suitable for single-point calibration are frequently those for which the response to a known amount of pure reference substance could be calculated, at least in principle, from physical constants and measurable parameters. The initial signal is commonly a non-linear function of the amount, and if linearisation is performed by firmware or software the analyst may have no knowledge or control of the algorithm used.

In many methods, a known amount of a substance to be assayed is dissolved and introduced into a flow stream (flow injection analyser or chromatograph) in which the initial bolus is

dispersed to an extent that can vary (within limits) from day to day. In principle, and sometimes in practice, one could still perform an assay without calibration using, for example, a refractive index detector that has previously been accurately calibrated. However, practical difficulties and quantitative uncertainties generally render this approach impracticable. Instead, the complete instrumental setup is calibrated by introducing a sample of a reference substance of known purity, which may or may not be the same as the substance being assayed.

For the present discussion, the calibration runs are assumed to be separate from the assay runs (external standardisation). While linearity may be intrinsic to favoured methods, the linear working range of the equipment must be demonstrated during a separate validation exercise, the long-term applicability of which is ensured by an instrument qualification program. We will assume for the moment that any non-linearities detected during validation have a negligible impact on analytical results, and will return to this aspect later.

The straight-line calibration is characterised by its slope, possibly an intercept, and some information on the dispersion of the calibration data. A single-point calibration is characterised only by a response factor, which implies that there is prior knowledge that the intercept is zero; evidently, there is no within-day information on measurement errors.

The Eurachem Guide (2000) gives detailed procedures for estimating "bottom-up" the combined uncertainty from the numerous individual uncertainties of assays performed using 'single-point' calibration. The Guide does not discuss the situation common in regulated laboratories in which calibrations must be at least in duplicate. Probably, many analysts take the average of the two response factors or rely on an algorithm provided by a laboratory data system. We will show that an inappropriate choice of calibration algorithm can make a significant but undocumented difference to the uncertainty of the analytical result.

We will attempt to outline an explanation of the situation that can be followed by analysts who don't have too much statistical background.

Response factors

Given that the analytical response has been shown during validation to be linear, the simple ratio of Equation 1 for the reference substance is equivalent to the slope of a full calibration curve if one had been prepared on the day:

Equation 1

$$Response\ factor = \frac{Response}{Amount}$$

The response may be a chromatographic peak area. The difference with respect to a full calibration curve is that the uncertainty of the analytical result is established entirely from the historical validation data. Traditionally, titrimetric methods are calibrated in terms of the titer,

reflecting an emphasis on the strength of the standard solution. The titer is the reciprocal of the response factor and the discussion that follows applies to it.

That is all that can be said about single-point calibration, except for one aspect that turns out to be less trivial than it may seem: in current analytical practice, it is rarely practicable to weigh the exact amount of standard that was specified, nor to select the amount of sample being assayed that gives a response close to those of the standards. Consequently, practically all chemical assays involve what amounts to an extrapolation, which is not allowed when results are juridically opposable. Regulatory authorities provide guidance, in accordance with long-standing practice, on suitable limits for the extrapolation. Such limits ought to be supported by method validation data, but no comprehensive analysis of the situation seems to have been published. In earlier times, the weighing of standards and samples was a tedious procedure, though rapidity was essential to minimise drift. Authors would indicate the acceptable range within which amounts need not be adjusted, typically 10% of the nominal amount. Nowadays, balances are easier to use, but samples may be smaller and more difficult to manipulate, and safety precautions are not always an aid to precision.

For regulatory analyses, such as those for the release of a drug substance or product, at least two independent calibration standards (separate weighings) are obligatory. This protects against blunders and provides a check that the repeatability of the response is not seriously inconsistent with the validation results. Often, the improvement in experimental uncertainty provided by replicates is also necessary. As explained above, it it not usually practicable to adjust the standards to bracket the expected value of the analytical result, and in any case the arguments that follow apply whether or not there is bracketing.

In the days before computer systems were introduced, it seemed natural and obvious to take the average of the response factors (or titers) for the two standards:

Equation 2

$$Response\ \ factor\ = \frac{1}{2}\left[\frac{Response_1}{Amount_1} + \frac{Response_2}{Amount_2}\right]$$

This practice can presumably be traced to Gauss, who did not presume that it is correct. He established that the average is the "most likely" value for a quantity that can't be known exactly because of random error, but only if the errors have certain statistical properties.

Data systems usually provide among their built-in functions a so-called unweighted linear least-squares regression fit to a straight line passing through the origin. While curve fitting is intrinsically an iterative process, there exists an analytical solution for this simple case, which can be written as follows for a duplicate calibration:

Equation 3

$$Response\ \ factor\ \left(Slope\right)\ =\ \frac{\left(Amount_1 \times Response_1\right) + \left(Amount_2 \times Response_2\right)}{\left(Amount_1\right)^2 + \left(Amount_2\right)^2}$$

As part of the blunder-proofing process, some quality systems require that it must be possible to verify these calculations using a (validated!) hand calculator. Consequently, analytical protocols and procedures may give Equation 2 while allowing the use of another algorithm provided by a data system, which may not be the same at all sites of a company. It may not be obvious that both Equations 2 and 3 are weighted linear least-squares regression fits to a straight line passing through the origin – and that both are usually wrong to some extent. The term "regression" was an unfortunate choice because it has no meaning in the present context. For our purpose its use implies that statistical methods designed for measurements on identical amounts of reference substance must be modified and extended when the amounts vary by an amount greater than a limit that has to be defined.

Discrepancies

Before presenting some theory, we will show by simulations that the difference between the algorithms is analytically significant; the analyst really does need to make a decision. It can be seen by inspection that equations 2 and 3 are the same if the two amounts are identical. To obtain some figures for purposes of illustration, we will consider a typical case of a chromatographic assay of a drug substance used to verify, for example, the uniformity of content of a batch of tablets. Although the pharmacopeias tend to avoid referring to confidence limits, the total uncertainty of the analytical results should not exceed 1.5%. Without entering into the details[1], this is unlikely to be achieved if the relative standard deviation (RSD) for independent measurements (separate weighings) on the reference substance is much greater than 0.7%, a stringent requirement given that the repeatability RSD of a chromatographic sample injector may be 0.5%. To facilitate the presentation (*Table* below), the discrepancies between the two calibration algorithms were calculated in terms of the expected analytical results for a fictitious case where the true assay is 100%, and on which the measurements are error-free. The conclusions do not depend on these conditions concerning the unknown. Calibration sample amounts were taken in the usually-allowed range of 90 to 110% of nominal, and calibration errors were 0, ±0.7% and ±1.4%, corresponding to zero, one and two times the typical standard deviation (expressed in conventional fashion as RSD).

1 Since acceptance limits are close to confidence limits, and analyses are expensive, there are rules for increasing the number of replicate assays if initial results are slightly out of specification.

Table. *Selected simulated assay results based on duplicate calibrations with a reference standard, and a compound being assayed known to be of 100% purity. The amounts of standard taken vary in the range 90 – 110% of nominal. The amounts of the samples of the unknown are irrelevant. The results are calculated according to the two algorithms given in the text above: A = Equation 2, B = Equation 3.*

Standard sample 1		**Standard sample 2**		**Assay results (%) True value = 100%**		**Discrepancy**
Amount (% of nominal)	**Error (%RSD)**	**Amount (% of nominal)**	**Error (%RSD)**	***A*: Average response factor**	***B*: "Unweighted" linear regression**	***A - B***
90	1.4	110	1.4	98.619	98.619	0.000
100	1.4	100	1.4	98.619	98.619	0.000
90	-1.4	90	1.4	100.000	100.000	0.000
105	-1.4	95	1.4	100.000	100.140	-0.140
110	-1.4	90	1.4	100.000	100.278	-0.278
110	1.4	90	-1.4	100.000	99.724	0.276
95	-0.7	105	1.4	99.651	99.547	0.104
90	-0.7	105	0.7	100.000	99.893	0.107
95	-0.7	105	0.7	100.000	99.930	0.070
110	-0.7	90	0	100.351	100.421	-0.070
90	0.7	110	1.4	98.961	98.893	0.068

It can be seen from the table that there is always a discrepancy unless the two calibration samples are equal or the calibration errors are equal. When both the amounts and the errors are at the extremes of these ranges, the discrepancy amounts to 0.28%, which is a very significant fraction of the tolerable experimental error. Discrepancies exceeding 0.1% occur even when the errors do not exceed one standard deviation. However, simulations (not shown) using normally-distributed random numbers for the measurement errors show that on most occasions the discrepancy does not exceed 0.1%. This is still high enough to call into question the validity of the calculations; it is also a reminder to check the numerical stability of the data system[2]. Attempting to calculate the statistical distribution of the discrepancy would peobably not be worthwhile, because there are indications that the distributions of "real-life" analytical measurements have relatively more values in the tails than would be expected for a normal distribution. In any case, significant discrepancies are bound to arise on many occasions in laboratories that carry out numerous analyses.

2 Some old compilers and interpreters may still be in use.

Statistical background

It is perhaps not stated clearly enough in official guidelines that the role of an analyst is to report the statistically most likely result for a determination or an individual measurement.

The simulations above indicate that the two replicate single-point calibrations algorithms most likely to be used give different values for the response factor when random errors are present. The discrepancy is significant compared to other sources of uncertainty. Clearly, one or both algorithms fail to give the most likely value. We will show that the correct choice depends on knowledge of the dependence of random error on the amount of sample. Since the kind of regression analysis being used (calibration curve "forced" through the origin) is rarely appropriate in other fields of endeavour, the literature on the subject is sparse. The subject is difficult, and analytical chemists and other more-or-less statistically-challenged physical scientists will be relieved to be informed that even the USA National Institute of Science and Technology (NIST) can get it completely wrong (Knaub, 2009).

Fortunately, Knaub (2005, 2009) has provided a particularly clear account of the application of our subject to 'establishment survey' data; under normal circumstances expenditure on goods, services and energy tends towards zero as income does the same. Knaub states that Equations 2 and 3 above correspond to extreme cases of statistical properties that are rather unlikely to be encountered in practice. Perhaps surprisingly, cases that approach or even go beyond these extremes can in fact occur in analytical chemistry (see below), though very often the analyst lacks information to decide whether one or the other, or a different expression, would be appropriate.

Equation 3 defines the so-called unweighted least squares straight line through the origin. Although it is termed "unweighted", it represents a fairly uncommon situation in which the variance is independent of the amount and all the weights are unity. This could occur with a 'residue on ignition' test where the residue is small (less than 1 milligram), the main source of random variation being the change in apparent weight when the crucible (which may weigh 10 grams) is taken to red heat and then cooled. The analogous situation in spectroscopy or chromatography is when baseline noise is the dominant source of measurement uncertainty.

Analysts frequently encounter almost the opposite situation. This can arise in assays where weighing errors are small and the main random errors are volumetric in nature; the deviations are then largely multiplicative rather than additive. One then observes that the standard deviation is proportional to the amount, so that the relative standard deviation is constant. In practice there are rarely if ever enough data to estimate the true error distribution. If there is reason to assume nearly constant relative standard deviation, the calibration curve should be calculated using weighted regression, the weights being inversely proportional to the variance. It is reasonable with this kind of data to use the squares of the amounts as a surrogate for the variance.

For readers who don't follow argument provided in textbooks, we could explain that the theory

of "ordinary" (unweighted) regression assumes that the random errors in the responses are additive. Regression theory is a extension of the notion of the average (arithmetic mean). If measurements are made on identical replicate amounts, the average is the most likely value (if the errors have a normal or other suitable distribution), provided the errors are additive. The average is also the value from which the squares of the individual deviations have the smallest sum ("least mean squares"; easily demonstrated by geometry or calculus). Sometimes the random errors in a calibration curve are volumetric (for example due to injection error or thermal expansion of a liquid). They are therefore purely multiplicative, and one could develop a different specific approach to finding the most likely value for the slope. However, maximum likelihood theory shows that the whole continuum between purely additive and purely multiplicative errors can be covered in a unified way by applying suitable weighting to the "ordinary" least squares procedure for additive errors. Sometimes (see below), the types of errors might fall outside the range of this continuum; I do not know whether maximum likelihood theory applies in such cases.

The same argument about weighting applies to replicate single-point calibrations, which are just calibration curves with a small number of points corresponding to similar but not usually identical sample amounts. Using *w*, *x* and *y* for the weights, amounts and responses respectively, Equation 3 (for calibration with duplicate standards) becomes:

Equation 4

$$Slope = \frac{w_1 x_1 y_1 + w_2 x_2 y_2}{w_1 x_1^2 + w_2 x_2^2} \quad .$$

The weights, here inversely proportional to the squares of the amounts, have to be normalised:

$$w_1 = \frac{1}{x_1^2} \div \frac{1}{2}\left(\frac{1}{x_1^2} + \frac{1}{x_2^2}\right) = \frac{2 x_2^2}{x_1^2 + x_2^2} \quad \text{and}$$

$$w_2 = \frac{1}{x_2^2} \div \frac{1}{2}\left(\frac{1}{x_1^2} + \frac{1}{x_2^2}\right) = \frac{2 x_1^2}{x_1^2 + x_2^2} \quad .$$

After inserting these expressions into Equation 4 and performing a little high-school algebraic manipulation, the result is Equation 2, the expression for the average response factor. Knaub states this result as though it were common knowledge, but it is unlikely that all analytical chemists or pharmacists are aware of it. We recall that Equation 3 is just Equation 4 with the weights both set equal to 1; the un-normalised weights could have been written:

$$w_i = \frac{1}{x_i^0}$$

Different weightings are conveniently expressed by defining the coefficient of heteroscedasticity γ such that[3]:

3 Some authors use a different definition in which γ includes the multiplier 2.

$$w_i = \frac{1}{x_i^{2\gamma}} \quad .$$

In many of the usual kinds of situation, γ may range from 0 ("unweighted"; additive errors) to 1 (constant RSD; multiplicative errors), but will often be somewhere between these values. Possible exceptions will be discussed later in a separate section. As stated, there is not always enough information to decide on the appropriate weighting, but with assay methods involving volumetric operations it is likely that γ will often be closer to 1 than zero, unless there are particular difficulties with sample weighing. Many of the sources of error in liquid chromatography are volumetric. This makes nonsense of a statement by the NIST, cited by Knaub (2009), that [unequal] weighting should not be applied unless one knows the true weights.

The Eurachem Guide (2000) and IUPAC (2002) discuss heteroscedasticity only in the context of full calibration curves for the evaluation of linearity. Eurachem presents an alternative but less general way of reporting the dependence of variance on amount. A detailed laboratory guide by Eurachem on method validation (1998) contains the strangely restrictive proposition: 'If variance of replicates is proportional to concentration then use a weighted regression calculation rather than a nonweighted regression'. Perhaps the published text does not reflect its author's intention.

Method validation protocols may require linearity data to be evaluated for heteroscedasticity, in order to justify the use of unweighted regression analysis. With the narrow working range that is commonly validated for drug assays, a test for heteroscedasticity may indeed indicate that unweighted regression may legitimately be applied without, as it were, breaking the rules. However, this does not provide the most likely values for the measurement and the measurement uncertainty. The correct approach should be, therefore, to apply weighting according to available knowledge of the dependence of variation on amount. The difficulty is that even a full validation study does not provide enough data points for γ to be estimated with confidence. Incidentally, the same applies to the assumption that analytical measurements are normally distributed. We will return to this point, but for the moment it is sufficient to note that, while the foundations of statistics for (modern) analytical chemistry are not completely secure, we can accept some degree of compromise on the understanding that acceptance criteria leave an adequate margin for imponderable uncertainties.

Knaub (2005) proposes that when information is lacking, a compromise value of 0.5 could be used for γ in his field of research. This falls conveniently half-way between the extremes that analysts unwittingly adopt by using Equations 2 or 3 for the response factor. It results in the following simple expression for the response factor, which would replace Equations 2 or 3:

Equation 5

$$Response\ \ factor = \frac{Response_1 + Response_2}{Amount_1 + Amount_2} \quad .$$

This expression is the basis of the 'Classical Ratio Estimator' (CRE). Any discrepancies due to

unsuitable weighting fall half-way between those obtained (see *Table*) using "unweighted" regression and the average response factor. At a deeper level, the statistical properties of the classical ratio are advantageous with respect to the estimators of Equations 2 and 3.

We could quite reasonably finish at this point: Equation 5 should be used for replicate single-point calibrations unless there is evidence that a different formula is closer to reality. On the other hand, this could be an occasion to re-think the uncertain basis upon which we design our quantitative analyses and do the calculations.

Possible values of the coefficient of heteroscedasticity

We repeat the equation that defines the coefficient of heteroscedasticity:

$$w_i = \frac{1}{x_i^{2\gamma}} \quad .$$

The multiplier 2 in the exponent is a reminder that statistical weights are inversely proportional to the variance, and not to the standard deviation. Also, this convention is convenient in applications where values of γ are constrained to the range 0 (variance independent of the regressor *x*, which is the amount in our case) to 1.0 (multiplicative errors, constant RSD). Some authors do not apply the multiplier 2 and they define a different 'γ' which has twice the value of the coefficient defined and used here.

When discussing heteroscedasticity in the context of statistics for analytical chemistry, we need to bear in mind that γ is not necessarily constant within a range of measurements, nor need it be limited to values between 0 and 1. To illustrate, the response random errors of a photometer (light absorbance detector) for samples in solution, used alone or as a chromatographic detector may have up to 3 causes:

1. Noise due (for example) to fluctuations in the light source, dust in the light beam, and thermally-induced turbulence in the solvent. This dominates at low sample concentrations (low absorbance): $\gamma \approx 0$.
2. Sample-related errors, which become significant at higher sample concentrations. If these are entirely volumetric (negligible weighing errors), they are multiplicative: $\gamma \approx 1$.
3. At very high sample concentrations[4] the light energy reaching the detector may have become low enough for detector noise to contribute to the error budget[5]. If the other errors are mostly volumetric it is then possible to have $\gamma > 1$. This is a reminder that there is a potential trap for the unwary whenever the physical response of the detector is a non-linear function of the amount or concentration, and the signal is linearised within the "black box" of the instrument.

Chemical and pharmaceutical assays are, when possible, performed when the second source of error above dominates, so that γ is expected to be reasonably constant.

It is sometimes necessary to determine impurities and other trace analytes under conditions where the noise is significant at the lower end of the working range; γ may be zero for small amounts, and then increase. Method validation is of little value here since noise levels vary

4 Fractional light extinction increases exponentially with the concentration.
5 In pharmaceutical applications the response may become non linear before this happens.

from day to day, and it would be better to do nearly all the essential verifications each time the method is applied.

In inverse assays, the analyte is made to react with a known amount of a reagent, and then the amount of reagent remaining is determined. Consequently, the variance could decrease with increasing amounts, so that γ is negative. The technique is common in titrimetric analysis ('back titration'); it is said to lack precision because two standard reagents are involved, though in fact the penalty should be minimal with an appropriate experimental protocol. I don't know if heteroscedasticity has been studied in this situation.

I am not familiar with assays involving chemical or biochemical amplification, fluorescence quenching, etc., for which γ might be expected to have a wide range of values. For completeness, we note that the same could apply to physical systems involving feedback, and for which the stability is not constant with respect to the regressor. Possible examples are climatic studies, and servomechanisms such as the laboratory air conditioning system.

Determining the coefficient of heteroscedasticity

While Equation 5 is usually an improvement on the ordinary (unweighted) least squares solution of Equation 3, that is typically provided by chromatography data systems, the level of uncertainty due to ignorance of the statistical properties of the data may still seem excessive. The maximum calculation error, which reaches about 0.14% for the simulations presented here, is comparable to the allowable uncertainty attributable to the precision and accuracy of the balance (0.1%), though the error distributions will be different. Since it may not be feasible to reduce the allowable range of sample weights (usually ±10%), we should discuss whether it might be possible to characterise the heteroscedasticity more closely.

While we can't apply a different value of γ to each assay monograph, a consensus value that would probably be somewhere between 0.5 (Equation 5) and 1.0 (Equation 2) could be envisaged for most kinds of assay (as opposed to trace analyses). This was discussed by Knaub (2011) on the basis of earlier work by K. R. W. Brewer, for the particularly difficult case of establishment surveys. Here, population statistics are estimated from relatively small population samples of variable size that may have missing data. In our branch of analytical chemistry, data sets are always complete (if anything goes wrong the complete procedure must be repeated), and historical data should be available to the extent that ever-changing information technology platforms allow.

However, there could be difficulties even for the most common case of chromatographic assays of powders or liquids. When the sample size is large compared to the uncertainty attributable to the balance, and the sample is easy to handle, the dominant sources of error are likely to be multiplicative (volumetric). It should be fairly straighforward to justify a consensus value of γ close to 1.0. The exact value is relatively unimportant insofar as the analyst will be able to weigh out amounts that are close to the nominal weight. By contrast, when the substance being assayed is sticky, hygroscopic, electrostatic or dangerous, the range of sample weights as well as the total measurement error are likely to be larger. While it is rarely helpful to attempt a full "bottom-up" analysis of assay measurement error, analysts will

by experience associate larger than normal[6] between-sample repeatability errors with sample-handling (weighing) difficulties . It is in such situations, where heteroscedasticity has a significant impact on the analytical result (because the nominal weight is difficult to achieve), that the value of γ will be less predictable (because the statistical properties of weighing errors are difficult to characterise).

Finally, we note that with a "consensus" value for γ other than 0.0, 0.5 or 1.0, the formula for the response factor will not reduce to simple expressions such as those of Equations 3, 5 and 2 respectively, because the exponents in the expressions for the normalised weightings are no longer integers.

Discussion and conclusion

According to a recent instruction by the United States Pharmacopeia (USP), the error (uncertainty) attributable to the balance used for weighing samples for assay must not exceed 0.1%. This could be taken as an indication of the desirable limit of uncertainty for other items of a list of contributions to the uncertainty budget, though it is not an absolute rule because the repeatability of chromatographs and titrators is not as good as this.

One could reasonably state that calculation errors due to rounding errors or to an inappropriate choice of algorithm should not exceed the error of the balance. When calibrations are done in duplicate according to current practice, the two algorithms likely to be in common use can show discrepancies of nearly 0.3% under the conditions we decribe (*Table*). The fundamental reason for the difficulty is that it is not usually possible to weigh exactly the nominal amount of reference standard. The variance of the analytical response is usually a function of the amount (heteroscedasticity) and this is reflected in the calibration function, whether or not the analyst is aware of it. Perhaps surprisingly, the subject seems rarely to have been discussed. It could be of historical interest because, quite possibly, a significant source of measurement uncertainty has been missed or ignored.

The two usual calibration formulae represent the limits usually encountered encountered with respect to heteroscedasticity. Both are mathematically equivalent to linear least squares fits to a straight line through the origin. With Equation 3 the statistical weights are equal (γ is zero), and with Equation 2 they are inversely proportional to the square of the sample amount (γ is unity).

As a mathematically simple compromise solution, one could use as response factor the 'classical ratio estimator' (CRE) of Equation 5, for which the coefficient of heteroscedasticity γ is ½. The maximum calculation error (*Table*) due to ignorance of the true value for the coefficient of heteroscedasticity is then halved. This may be sufficient, particularly if the working range can be narrowed a little. An essential requirement for regulatory methods is that calculations are explained and documented in such a way that those whose job it is to

6 Chromatographic assay protocols include verification or evaluation of within-sample instrument repeatability and of overall repeatability between replicate reference samples.

verify them will always obtain the same result. Theoretical rigour is of slightly less importance.

However, it could be argued that the maximum calculation error provided by the CRE still exceeds the allowable uncertainty for the balance[7], though its statistical distribution will be different. Consequently it may be useful to find out whether statistical analysis of numerous sets of real data could yield a more accurate "consensus" value for γ. The true value should be closer to 1.0 than 0.5 in many but not all cases. Suitable data sets are readily available in method validation reports, though it could be questioned whether they are representative of everyday practice. A serious limitation to this approach is that substances that are difficult to weigh are likely to yield data with variable and inconsistent statistical properties, and they may not be specifically flagged as such.

The difficulty in estimating heteroscedasticity could lead one to reflect on the fact that the statistical model for an assay procedure is never likely to be complete. Another potentially embarassing question that does not seem to have been anwered is whether modern assay data are normally distributed. This may be an occasion to float the idea, in an appendix, that if there is no reliable practically-applicable statistical model, we ought to find ways of making the confidence limits of reported results less dependent on models. Already, one weakness of our dependence on prior validation studies is that statistical data from such studies are not relied upon when estimating confidence limits (IUPAC, 2002).

Narrowing the working range

When samples are made up to an accurately known volume (as in flow-injection and chromatographic, but not titrimetric methods), the effects of sample weight variation could be compensated by adjusting the volume so as to obtain a predetermined concentration. This is impracticable using volumetric glassware, and until recently dispensing the solvent by weight was (apparantly) not permitted in pharmaceutical analysis, because the USP did not seem to have been aware of some rather basic physics.

The need to adapt volumes would be reduced or eliminated if the amount of sample weighed out could be kept within a range considerably narrower than ±10%. Such an improvement seems to be the only option for reducing the uncertainty due to uncertain heteroscedasticity with titrimetric methods, for which the sample solution volume is not critical. Some trials (unreported) with manual powder dispensers designed for high throughput chemical synthesis were unsuccessful. It would be interesting to know how closely a newly-designed balance (Mettler-Toledo) equipped with programmable dispensers for powders and solvents meets target weights, for a suitably wide variety of samples.

In some situations an alternative approach to adapting sample volumes would be to program the automatic sample injector to vary the volume injected. However, the accuracy of injectors operated in this mode may not be adequate for typical drug assays. Perhaps surprisingly, this technique does not appear to have been used for the determination of impurities (for which repeatability criteria are less stringent) by mass spectrometry, where the linear working range

7 The rule concerns the uncertainty established using standard weights, and not the complete operation ow weighing a sample.

may vary from day to day.

Note on non-linear responses

Documents and guidelines emphasise strongly the need for a linear response if single-point calibration is to be relied upon. Frequently, when drafting method validation reports, it is necessary to argue that observed non-linearities may be neglected. The subject is important, because one of the most commonly used techniques, liquid chromatography with spectrophotometric detection, usually gives a response that falls off slightly at high absorbance values. Since the shape of the response is to some extent predictable from the spectra and the spectrometer bandpass, relevant calculations ought to be presented in the validation protocol.

The effects of a reasonably small degree of chromatographic non-linearity can be eliminated, however, if the concentrations of the calibration solutions can be made close enough to the nominal value. An additional requirement, that does not apply to uncertainties regarding heteroscedasticity, is that the expected response for the unknown be sufficently close to those of the standards. This can be arranged if the concentration of the solution of the unknown is adjusted accordingly, either by iteration or on the basis of prior knowledge of the likely analytical result.

A related question is whether the amounts of the two standards should deliberately be made different, with the aim of determining the straight line that is not forced through the origin. Generally, this would provide a closer curve fit. However, the situation regarding uncertainties would require study in order to avoid "sacrificing" a degree of freedom in order to correct a second-order deviation.

In the limit of accurate matching of standards and samples, there would be no need for a linear response; it would be sufficient to demonstrate that the response is adequately dependent on amount. Some chromatographic detectors in common use (particularly evaporative light scattering) have responses that are both unpredictable and non-linear. It is surprising, therefore, that the possibilities afforded by modern equipment for automating iterative techniques have not been widely exploited.

Finally, if the amount of standard could be closely matched to the expected result for each unknown, it would be possible to dissociate the two objectives of most quality control analyses: establishing compliance, and estimating the content. In the case of impurities, the test for compliance would amount to a limit test, which has less stringent validation criteria than a quantitative determination. One may need a system suitability test designed to verify that the response does depend on the amount (the detector is not saturated), but I am not aware of any discussion of that subject.

Note on response drift

As mentioned in the Introduction, response drift becomes more difficult to manage when more than one standard is run, and there is little published guidance on this matter. At the risk of drifting slightly off-topic, improving the monitoring and correction of drift should, therefore, be taken into account when discussing innovative approaches to the use of replicate standards.

In metrological procedures, for example for the calibration of temperature sensors, the effects of drift and fluctuations in the test conditions are compensated by taking repeated readings of the reference sensor and those under test; typically, the subsequent calculations involve the Classical Ratio (Equation 5). Drift is less often discussed in analytical chemistry and pharmacy, although it can be troublesome because of the strong effects of temperature variations on liquid volumes. A common policy with respect to quality assurance is to monitor drift using repeat injections of the unknowns or one of the standards, and simply to reject the analysis if drift exceeds a pre-defined amount. However, this approach does not provide the statistically most likely value for the analytical result, and it it may reflect an underlying attitude that drift is not supposed to exist. Ideally, measurable drift should, if it is regular, be corrected even if it is within specified limits.

Official guidance on suitable run sequences for assays with duplicate single-point calibration appears to be lacking. Such assays may concern one or a few samples (batch release), or a relatively large number (content uniformity, stability studies). An optimised automated sequence would have the calibration and system suitability runs intercalated with the unknowns, particularly if the vials of the unknowns are not to be recovered in the event that the instrument system is found to be out of specification.

Our example involving traditional 6-point calibration curves mentioned in the Introduction (Lee *et al*., 2003) serves to illustrate the practical limitations of current textbook methods applied to pharmaceutical analyses. We presented a method for the determination of trace impurities by gas chromatography-mass spectrometry. In this author's opinion, mass spectrometry does not meet the criteria described above (*Analytical background*) whereby a validation exercise can be relied upon to ensure long-term applicability with respect to the linear working range and other key parameters. We therefore prepared calibration curves on the day, according to standard procedures for inverse regression analysis. The aspect relevant to the present discussion is that the instrument response drifted fairly regularly during an injection sequence (calibration and a small number of determinations), by about 10%. Such drift is fairly typical for the technique, though not too well documented. It is to be expected because for operational reasons the ion source is not protected from the large amounts of solvent vapour that traverse the chromatography column at each injection. We chose to randomise the injection sequence, which led to uncertainties for the analytical results that were fairly large, though acceptable for a impurity determination. Clearly, much lower uncertainties could have been obtained, once reasonable linearity on the day has been established, by repeating the analysis with one or two standards more closely matched to the estimated amount of impurity. This would have been inconvenient because of the particular sample preparation method used, but with other analyses such iteration could be automated.

Acknowledgements

Thanks are due to James R. Knaub and to former colleagues Dominique Peter and Jean-Pierre Porziemsky for numerous helpful discussions.

APPENDIX

Validity of assay method validation procedures

Summary

Most chemical assay methods are calibrated, usually on the day, using a reference substance. Methods have to be validated to ensure that measurement errors comply with certain criteria throughout the working range. In practice, validation, a long and expensive procedure, does not provide enough statistical data to ensure that assay results are statistically the most likely values. It is proposed that the need for validation could be reduced or eliminated by means of adaptive or iterative calibration procedures, whereby the amount of reference substance is ajdusted to match the substance being assayed in order to considerably reduce the required working range.

Introduction

This appendix is concerned with quantitative analytical methods that are applied on many occasions, for example for 'release' of successive batches of a product or for stability studies. We confine the discussion mainly to determinations of a type frequently termed 'assays' where the substance (usually a single compound) being determined is the desired or main component, and the allowable statistical uncertainty is small (typically $U \approx 2\%$).

For economic reasons, the statistical properties and parameters of a method are investigated once only during a validation study, so that calibration and system suitability checks carried out each time the method is applied can be brief. In response to a certain amount of confusion dating from the widespread introduction of instrumental chromatographic and other techniques, validation procedures were formalised during the 1990s. Other essential features of current 'quality systems' are instrument 'qualification' procedures, and plethoric precisely-worded protocols.

These arrangements work well in the sense that they enable acceptance limits to be set that come quite close to the (estimated) statistical uncertainty of the determination. They also facilitate dialog with the inspectorate without everyone concerned having to understand every detail of the underlying scientific rationale. However, taking into account the quality assurance overheads, method validation is both time-consuming and expensive. Consequently, the burden of validation may reinforce a tendency to restrict analytical protocols to a limited number of not-very-innovative techniques. This is one meaning of an expression commonly heard, which can be formulated as: 'quality destroys quality'.

I have shown in the main article that imperfect knowledge of heteroscedasticity is a source of uncertainty that seems to have been neglected. This recalls more long-standing concerns about

the validity of the commonly accepted presumption that analytical responses have normal distributions. These concerns are not just theoretical: they cast doubt on the notion that method validation data can be transposed to routine applications, because it becomes impossible to declare that the reported result is the most likely value. I am not aware of any published information on the costs that might be entailed by batch rejections that could have been ascribed to an underestimation of measurement uncertainties.

It may be worth reviewing the situation regarding method validation, with a view to finding a more efficient approach.

Purposes of an assay

Often, one may think of an assay as the determination of the amount of the substance that has a wanted function, either in an ingredient ('active principle', 'drug substance', etc.) or in a formulated product. If we wish pharmaceutical analytical practice to remain (or become) connected with normal statistical practice, assay results should be associated with their estimated confidence limits. This is not in fact current practice, and where confidence limits are discussed (IUPAC, 2002), they are not be derived from relevant experimental measurements such as those obtained during method validation.

There seems to be no published explanation for this peculiar practice, although one occasionally reads of difficulties in reconciling theoretical error budgets with real error dispersions. Possibly, there is some kind of unwritten understanding that the statistical properties of the measurements can not be established with sufficient confidence.

Our discussion of heteroscedasticity in the main article leads logically to the notion that one day it might become possible to develop adaptive or iterative automated techniques whereby the amounts of reference substances and substances being assayed that are taken for analysis are matched, so that all the analytical responses are nearly the same. Some possibilities are discussed. If this can be achieved, it would then be logical to draw a distinction between two possible purposes of an assay, because the statistical treatment would be different.

Frequently, we need primarily a binary decision about compliance. This is the case when a drug substance is expected to be pure to the extent that the total amount of any impurities is small compared to the uncertainty of the assay. In such situations, and if the amounts of standards and samples can be matched, it would be possible to apply non-parametric statistical methods that do not depend on knowledge of the dispersion of the errors. The approach would be similar to traditional limit tests for impurities, for which validation requirements are minimal.

On other occasions the assay can not be predicted so accurately, and we need a quantitative result in order, for example, to calculate the amount of drug substance to be included in a formulation. The question then is whether technical changes, including the adoption of adaptive or iterative methods, could provide a more certain and predictable estimation of the error distributions and confidence limits of the result. As proposed, automated weighing would help change long-standing practices that seem statistically rather uncertain.

Knowledge of statistics

Numerous textbooks on statistics for analytical chemistry have been published. However, a

fairly widespread practice within the pharmaceutical industry (at least) has been to discourage scientists from carrying out statistical analyses of their own data. Access to statistical software was restricted in some companies, partly because of this lack of trust, and partly because of software validation issues; statistics software is particularly vulnerable to difficulties with numerical stability. Practically every aspect of pharmaceutical development, as well some aspects of 'discovery' research is heavily dependent on support by professional statisticians.

As a consequence, analysts in this field, who may not always be at ease with statistics, have no opportunity to improve their skills. Method validation protocols seem to be designed for such end users and may, therefore, have something of a "check-box" feel.

Limitations of parametric statistics in analytical chemistry

When consulted about specific analytical problems, company statistians would express unease because there were not nearly enough data to justify the presumption that the error distribution is normal. The subject is rarely discussed in the literature, and I recall only one brief mention (in one of the editions in the book by Miller and Miller) of evidence that real data may have a relative excess of results in the tails of real distributions. Clearly this is unverifiable for any particular method validation, where running as few as 10 or 20 replicates may already present technical and organisational difficulties. On the other hand, assay results that fall by chance in the tails are the ones that can have economic consequences, and we have shown that the risk of reaching the wrong conclusion are aggravated if heteroscedasticity is not properly taken into account.

As far as I am aware, abundant production control data are treated as highly confidential and don't make their way into the literature.

The Eurachem (2000) and IUPAC (2002) documents place emphasis on the "bottom-up" evaluation of total measurement uncertainty by, as it were, dissecting out the numerous different individual sources of error. A practising analyst might prefer an approach whereby the experimental results of the method validation study can be used to estimate the uncertainty of subsequent reported assay results. This information is not obtained by following published guidelines on validation.

Individual errors do not need to follow a normal distribution too closely if they are numerous enough, and there is no dominant source of error. However, I am not convinced that these two conditions are always met.

Weighing errors may be considered to make a significant contribution to the uncertainty budget. The uncertainty of the balance reading has recently been defined (0.1% of the sample mass), but this should not be confused with the uncertainty of the complete unit operation of sample weighing. The difficulty with weighing out exactly the specified amount of sample is the main reason for writing the main article. An additional concern with the weighing of powder samples is that the weight of any powder that falls on the balance pan or becomes attached to the outside of the weighing vessel will be recorded; whatever the skill of the analysts, balances do sometimes need cleaning. By contrast, it is relatively difficult to think of general mechanisms whereby the recorded weight would be less than the amount that is subsequently transferred[8]. Consequently, the distribution of sample weights could possibly be skewed towards low values. Evaluation would be difficult because the feasible number of replicates is small; also, the estimation of the weighing error requires an analysis of variance

8 Hygroscopic substances are frequently encountered but that does not invalidate the general argument.

of the complete assay which generally has larger sources of error, and which in turn requires assumptions about the individual errors distributions.

Sample injection is a unit operation in chromatographic analysis that may make the biggest contribution to random error. Some years ago, while studying failure modes of a common type of injector, I noticed that asymmetric error distributions could be envisaged in normal operation. There was no opportunity to follow this up, and I am not aware of any relevant literature; when chromatographic (injection) repeatability is discussed, usually only the relative standard deviation for five or ten injections is given.

Do validation data represent the "real-life" situation?

Unlike routine pharmaceutical assays, validation involves numerous repetitive operations. One is likely to employ the most competent analysts and the most reliable equipment in order to avoid having to start over. Likewise, instrument components with limited lifetimes are likely to be changed for the validation exercise, and the instruments selected may be the ones that have been most recently serviced and validated ('qualified' in the jargon); I have not seen a protocol that requires analysts and instruments to be chosen by drawing lots. If a powder is known to be difficult to handle (hygroscopic or electrostatic), one may be tempted to postpone the exercise if the weather is unusually wet or dry.

Because analytical runs are likely to be longer than usual, the situation regarding operator familiarisation and fatigue may differ from the daily routine. Batch sizes and the duration of use of reagents and solutions such as chromatographic mobile phases may be increased.

In principle, methods should be tested using a fully representative selection of instruments, but this is not always practicable if, for example, one takes into account future purchases or company reorganisations and mergers. To illustrate, a critical factor with many liquid chromatographic methods is the performance of the spectrophotometric detector. Different models differ with respect to type of spectrometer, spectral bandwiths and available bandwith settings. The bandwidth is important for linearity, but it is not always specified in the method (and almost never in compendial monographs). Since linearity is also affected by optical imperfections (stray light) that can vary between instruments and with time, it ought really to be verified as part of the 'system suitability test' carried out each time the method is applied.

Mass spectrometry is used more often for impurity determinations than for the kinds of assay methods considered here. However, it should be mentioned here because a successful validation using this detection method does not reliably predict future performance, in particular with respect to the linear working range.

Method validation with adaptive or iterative assay procedures

The starting point for the main article was that, probably with most assays, the variance of the analytical response is a function of the sample amount (heteroscedasticity), and this is important because it not possible to weigh out exactly the specified amount of reference sample. Mathematical formulae for method calibration take heteroscedasticity into account, though analysts may not be aware of this. However, the degree of heteroscedasticity is not always known with sufficient accuracy, and it may vary from one application of a method to the next. Possible solutions are to adopt technology that allows sample weights to be more nearly identical, or to make adjustements subsequent to weighing in order that the amounts

introduced into the measuring instrument are nearly identical.

Such procedures would render linearity testing almost redundant. This test is probably the most difficult and time-consuming item in assay validation protocols, and it consumes large amounts of valuable reference substances. Moreover, the true response is very often detectably non-linear, and the analyst must provide a suitable argument for the validity of the method.

Balances with automatic powder and solvent dispensers were introduced only recently. This and other innovations are likely to affect the statistical properties of analytical data. Uncertainties due to non-linearity and heteroscedasticity will be reduced, and it may perhaps become possible to provide a more reliable estimate of the error budget.

Method validation vs in-use method evaluation

The main reason for having a separate method validation exercise is that the extensive testing can not be repeated every time a method is used. If we remove the need for rigorous linearity testing, this argument no longer applies; most of the other tests, notably repeatability, could be done during method development or during its practical application.

This would bring pharmaceutical practice more in line with the traditional textbook technique of longitudinal quality control charts. Normal pharmaceutical practice for routine assays is to use duplicate standards (at least) as a repeatability check, and sometimes to repeat chromatographic sample injections. It could be argued that longitudinal evaluation of duplicates should provide a more realistic assessment of repeatability than a more extensive once-for-all validation of repeatability.

Conclusion

Standard validation protocols were established some time after the introduction of instrumental assay methods, because it was necessary for analysts not too conversant with statistics to be able to satisfy regulatory authorities that routine methods are under control. It is currently impracticable to conduct all the necessary verifications each time a method is applied.

Transposing validation data to subsequent individual analyses implies that the statistical properties of the data have been characterised. We have shown that even if the heteroscedasticity of the validation data is known, this is not routinely take into account. Furthermore, there appears to be no reference in official documents to evidence that, under current practice, measurement errors are normally distributed; there are never enough data to verify the distribution for a given analysis. This may be reflected in the observation that validation reports do not always provide an estimation of the statistical uncertainty of a routine determination. We have presented additional reasons to doubt whether extensive and expensive validation exercises indicate the performance of methods under normal conditions.

We propose that new technology be introduced to enable the working range of routine analyses be made narrower than at present. An advantage would be that results would be less dependent on the statistical properties of the meaurement errors. Above all, the need for method validation would be largely eliminated, and it should be possible to distribute any remaining validation tests between the method development studies, system suitability

procedures and retrospective in-use statistical analyses.

References

Eurachem (1998). The Fitness for Purpose of Analytical Methods A Laboratory Guide to Method Validation and Related Topics. Eurachem/Citac, *http://www.eurachem.org/index.php/publications/guides/mv*

Eurachem (2000). Quantifying uncertainty in analytical measurement. Second edition. Eurachem/Citac, *http://www.measurementuncertainty.org.*

IUPAC Technical Report (2002). Harmonized guidelines for single laboratory validation of methods of analysis. *Pure Appl. Chem.* 2002, 74, 835-855.

Knaub (2005). The Classical Ratio Estimator. James R. Knaub*, Interstat* 2005, 0510004. *http://interstat.statjournals.net/YEAR/2005/articles/0510004.pdf*

Knaub (2009). Properties of Weighted Least Squares Regression for Cutoff Sampling in Establishment Surveys. James R. Knaub*, Interstat* 2009, 0912003. *http://interstat.statjournals.net/YEAR/2009/articles/0912003.pdf*

Knaub (2011) Ken Brewer and the coefficient of heteroscedasticity as used in sample survey inference. James R. Knaub, *Pak. J. Statist.* 2011, 27, 397-406.

Lee et al. (2003). Determination of polar alkylating agents as thiocyanate/isothiocyanate derivatives by reaction headspace gas chromatography. Christopher R. Lee, Florence Guivarch, Céline Nguyen Van Dau, Dominique Tessier and Ante M. Krstulovic. *Analyst*, 2003, 128, 857-863.

Miller (1991). Basic statistical methods for analytical chemistry. Part 2. Calibration and regression methods. J. N. Miller*, Analyst*, 1991, 116, 3-14.