The Sample, The Procedure, and The Laboratory

W. J. Youden
National Bureau of Standards, Washington 25, D. C.

IN THIS paper the viewpoint is taken that an analytical procedure has an inherent accuracy and precision. True enough, there must be an analyst in a laboratory to put the procedure to work and this implies to some analysts that an inseparable association exists between procedure and operator. A sample is also indispensable, yet there is no hesitation in sometimes attributing the variation in analytical results to a lack of homogeneity in the material furnishing the samples. At other times, often when a reasonable volume of a liquid is sampled, the aliquots used as samples can be considered identical in composition and any differences among the results cannot be charged to the samples.

The role of the analyst, or laboratory, may be revealed when two or more laboratories undertake determinations on samples drawn from the same stock of uniform material. In extreme cases the repeat determinations made by a laboratory cluster closely about the laboratory average without any intermingling of the results from one laboratory with the results from another laboratory. Figure 1 illustrates this point. The open circles represent the results from one laboratory and the solid circles the results reported by a second laboratory. Separation of the results from different laboratories is practically always present to some extent—that is, the separation between results from different laboratories is greater than would be anticipated, considering the agreement among the results obtained within a single laboratory. The reduction, or, if possible, the elimination of these interlaboratory differences is an everyday problem. Here is a major reason why busy analytical chemists turn to statistical techniques for help in resolving the complex of circumstances that surround analytical determinations.

Wrong Operations on Data

Often a study makes available a collection of analytical results obtained under a variety of circumstances. One wrong operation is to take the grand average of all the data and obtain the individual deviations from this average. It matters not whether the simple arithmetic average of these deviations (of course ignoring signs) is reported, or some more sophisticated quantity, such as the standard deviation, is computed. The quantity so reported is almost surely useless, if not downright misleading. Nor will matters be helped if the analyst happens to have available the theoretical or assumed true composition of the material and is able to measure his deviations from the true value. In fact, this usually makes matters worse. I am fully aware that these computations are very generally made, but they are made in the mistaken belief that the simplicity of the calculations ensures a meaningful result.

An illustrative example will clear the ground of erroneous operations on the data. The example is taken from some long ago microanalytical determinations of carbon reported by Power (*1*). Analyst H reported six determinations on pure ephedrine hydrochloride as follows:

59.09, 59.17, 59.27, 59.13, 59.10, 59.14
Av. 59.15

If the deviations are obtained by subtracting from these results the theoretical per cent of carbon, 59.55, the deviations are

−0.46, −0.38, −0.28, −0.42, −0.45, −0.41
Av. −0.40

We are immediately struck by the unvarying minus sign and the relative constancy of these large negative deviations. By accident, in this example, because all the deviations have the same sign, the average of these deviations (−0.40) is informative. It is, in fact, an estimate of the bias or systematic error in the results, and if the sign is retained, we have the *direction* of the bias. The average deviation is not always so kind as to furnish an estimate of the bias. When the signs of the deviations are not all the same,

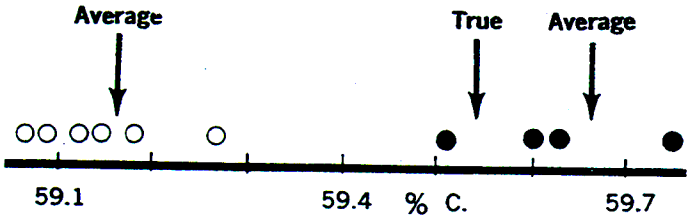
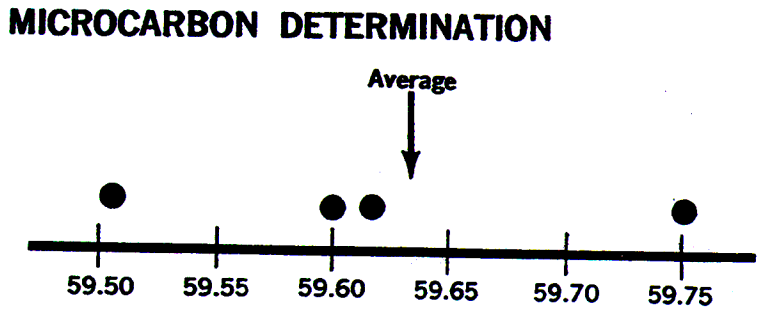


Figure 1

138-23A



Inclusion of extreme values may displace average unduly

the average of the absolute deviations no longer measures the bias—or anything else. Probability statements cannot be made about the above deviations because they all have the same sign. One could state that no matter how many determinations had been made, they would all have given negative deviations from the true composition.

Power listed four of his own determinations that he considered acceptable. His results were 59.51, 59.75, 59.61, and 59.60, with an average of 59.62. Apparently Power avoided whatever circumstances led analyst H to his low results. The ten deviations that would be obtained by taking differences from the average of all ten results tell nothing useful. The deviations reflect a confused mixture of random errors and systematic errors. Even the average used clearly depends upon the relative numbers of determinations provided by the two analysts. If the theoretical composition is used, the deviations visibly consist of two groups with no intermingling. Statistical statements for such heterogeneous deviations are meaningless. It is more informative to state for each analyst the departure of his average from the theoretical composition, for each to give an estimate of his precision using the deviations from his own average.

When the magnitude of the systematic error is comparable to the random errors associated with precision, a predominance of the deviations from the true value will have the same sign. When a random error of opposite sign and somewhat larger than the systematic error comes along, the net result is to give a sign opposite to that shown by the majority. The best evaluation of the random errors exhibited by the above six results is obtained by using the deviations of the individual determinations from the average of all the results. The deviations, —0.06, 0.02, 0.12, —0.02, —0.05, and —0.01, must sum to zero and should show a reasonably equal partition between plus and minus signs.

The *estimate* of the standard deviations associated with the laboratory in which analyst H made his determinations is given by $s = \sqrt{\Sigma(\text{dev})^2/(n\text{-}1)}$ or 0.065. The estimate of the bias, —0.40, is about six times as large as s. A random deviation (either plus or minus) of this magnitude is extremely unlikely. Hence all the signs of the deviations are the same. As the ratio of the bias to the standard deviations gets smaller, there is more likelihood of a mixture of signs. Table I shows for various values of this ratio the expected division of the signs of the deviations from the true value.

This particular example was chosen to bring out clearly the two concepts of a systematic component of error and a random component of error. It may be, that in as clear cut a situation as this one, few would go astray. But it must be remembered that there is a continuum extending from very large obvious biases down to very small biases. The values computed from the data should correspond to meaningful chemical quantities. The separation of bias from random errors is indispensable to an efficient approach to the improvement of analytical procedures.

Statisticians have unwittingly contributed to the confusion when they remark that the divisor for the sum of the squared deviations must be one less than the number of measurements, because the deviations are measured from the average rather than the true value. The statistician and the chemist refer to quite different things when they speak of the true value. The chemist has in mind the actual correct composition. The statistician means the value that the average of the results would approach with an indefinite increase in the number of determinations made under the same conditions. In other words, the statistician's true value includes the systematic error, if any.

True Composition Unknown

If the true composition is not known, the estimation of the magnitude of a systematic error in the results is not so easy but in some situations not impossible. If the systematic error in the determinations is the same over a considerable range of sample weight (or volume), the systematic error may be estimated by plotting the actual measured quantity against the sample

weight. The measured quantity may be the weight of a dried precipitate or the milliliters used in the titration. Clearly if one sample weight is twice the weight of another sample, there should be twice as much precipitate or twice as many milliliters of reagent used. If there is a systematic error that is independent of the sample weight, all the results should be high (or low) by the same amount. A straight line fitted to the points will not go through the origin, as it ought to, but will intercept the y-axis. The intercept is an estimate of the systematic error. This device fails if the systematic error is proportional to the amount taken for analysis.

While it may be difficult to estimate the magnitude and sign of the systematic error, the demonstration that systematic errors are present is all too easy. If two laboratories report a number of analyses on the same material, any difference that can be established between the laboratory averages is evidence that one or the other or both sets of results are afflicted with a systematic error. It was shown above that any attempt to describe such joint collections of data by a single statistical unit is bound to be misleading.

The evaluation of analytical data is greatly simplified if it is assumed that the participating laboratories have the same precision. The basis for this assumption is that apparatus, equipment, and analyst training are highly standardized and of high quality. Weighings, titrations, instrument readings, and the like are likely to be made with about the same reproducibility. Usually if there are differences in apparatus or technique, these concern matters that do not contribute appreciably to the precision. Weighing errors, for example, are usually a minor consideration, so that little consequence comes from one laboratory using a balance with twice the sensitivity of the balance used in the other laboratory. Thoughtful consideration of the steps in an analytical procedure soon leads to the conclusion that differences between laboratories in regard to equipment, reagents, or in procedures are more likely to lead to systematic errors than to changes in precision.

The most obvious source of a systematic error is a deliberate or unwitting departure from the prescribed manner of carrying out the procedure. Chemists are individuals; they have their favorite precautions, short cuts, and prejudices.

Table I. Division of Plus and Minus Signs of Deviations from True Value Depends on Ratio of Systematic Error to Statistical Deviation

Systematic Error / Standard Deviation	Division of signs of Deviations, %	
2.0	97.7	2.3
1.5	93.3	6.7
1.2	88.5	11.5
1.0	84.1	15.9
0.8	78.8	21.2
0.6	72.6	27.4
0.4	65.5	34.5
0.2	57.9	42.1
0.0	50.0	50.0

If a chemist faithfully follows his own routine, his own analyses check each other extremely well. The same will be true for a chemist in another laboratory. His internal checks are no doubt just as good as those obtained in the first laboratory (same precision) but the results, as a group, may reflect the established practice of the laboratory. Similarly reagents in the two laboratories may be from different sources, or lots, or of different ages. All determinations run with a given set of reagents may show excellent internal agreement but average out at a value removed from the aver-

---

W. J. Youden, a statistical consultant at the National Bureau of Standards for the past 12 years, is an unusual combination of analytical chemist, chemical engineer, and statistician.

Although an Australian by birth, he came to the U. S. at an early age. He received his B.S. in chemical engineering from the University of Rochester (1921), and his Ph.D. in analytical chemistry from Columbia University (1924). His thesis concerned a new method for the gravimetric determination of zirconium. In 1937 he held a Rockefeller Fellowship at the University of London.

He joined the staff at the Boyce Thompson Institute for Plant Research in 1924. During the following 24 years, he did research on such topics as tobacco virus, isoelectric points, soil sampling, sugar analysis, seed treatment, pH methods, agricultural field trials, and greenhouse fumigation. As a result of some of those studies he became involved in statistical approaches and in particular to the design of experiments.

He put his statistical skills to work in a completely different area during World War II when he served as an operations analyst in the area of bombing accuracy with the Army Air Forces overseas (1942 to 1945). He also was an operations analyst for the Rand Corporation in 1947.

He joined NBS in 1948 as a statistical consultant. His major interests are the design and interpretation of experiments and the application of statistical techniques in analytical chemistry.

He has served as a visiting professor at the North Carolina State College (1951, 1954, 1955) and as a professor at the University of Chicago (1959). He has also given continuation lectures on the design of experiments for the Philadelphia and New York Sections of the ACS (1951 and 1954, respectively) and has been on seven speaking tours for the ACS and one for the Canadian Institute of Chemistry. He has served as a statistical consultant on several government boards, committees, and councils.

He has been a member of the ACS for 40 years. He is also a member of Sigma Xi, Phi Beta Kappa, and Phi Lambda Upsilon and is active on several ASTM committees. For his work on Youden squares, chain blocks, linked blocks, and partially replicated Latin squares, he has been honored by statisticians who have made him a Fellow of the American Statistical Association, a member of the International Statistics Institute, and a titular member of the Commission of Technology and Expression of Results of the Analytical Section of the International Union of Pure and Applied Chemistry.

age of determinations made with another set of reagents. Pieces of equipment may differ in their zero settings and introduce different biases without in any way altering the precision of the readings. Geographical location sometimes involves fairly persistent humidity differences between laboratories and this may be a reason for the difference between laboratory results.

Finally there is an abundance of evidence that different laboratories have different systematic errors for a given procedure. Little convincing evidence exists of differences in precision. Of course each laboratory likes to believe that it does particularly precise work. Sometimes this belief is bolstered by a too enthusiastic culling of results and running of extra repetitions until a "satisfactory" agreement is obtained. Leaving aside any spurious apparent differences in precision generated in this manner, it seems fair to conclude that laboratories with equivalent equipment and personnel achieve about the same precision.

In any event, it takes a lot of determinations to make a convincing case for differences in precision. Suppose two laboratories each make ten determinations and an estimate is made of the standard deviation for each laboratory. One of the estimates of the standard deviations must be at least twice the other estimate to provide reasonable grounds for the suspicion that there is a real difference in the quality of the work. Suppose that one laboratory does regularly turn out work that has a standard deviation one half as large as that associated with the regular work of another laboratory. If each laboratory submits 20 repeat runs, there is only about a four out of five chance that this actual difference will be reflected convincingly enough in the data to warrant the conclusion that the laboratories differ in precision.

A more vivid illustration of the difficulties in the way of discriminating among laboratories is afforded by the following comments. We assume that six laboratories all have identical precision. The laboratories report five determinations apiece and the standard deviations are calculated. Then we should not be surprised if the ratio of the largest estimate to the smallest estimate of the standard deviation is as much as 5.4. Even if the estimates are based upon ten repeat determinations, the ratio may reach 2.8 purely from the chance distribution of the deviations. If ten, instead of six, laboratories participate, the ratios are 6.7 and 3.1. The nature of measurement is such that, even under the ideal conditions of assumed normality and absence of gross errors, any measure of precision is subject to large sampling variation. Unless there is clear evidence to the contrary, the best procedure is to combine, in the proper way, the several estimates of precision and award this value to all participating.

The combination of the estimates is easily effected by adding together the sums of the squared deviations available from the several sets of results and dividing by the sum of the divisors previously employed. The deviations for each set must be measured from the average of the laboratory (or group) from which the data originate. The six results by analyst H and the four results by Power give the following pooled estimate of the standard deviation:

$$s = \sqrt{\frac{0.0214 + 0.0295}{5 + 3}} = 0.080$$

The remarks about apparent and not real differences in precision also apply to different sets of data accumulated *within* one laboratory. Suppose that there are two sets of measurements, each made up of three repetitions. Perhaps these sets were made on different days. If the range, or spread, for one set is twice that of the other, one cannot conclude on this evidence alone that one set of measurements is more precise than the other or that more confidence may be placed in the average of the set with the smaller range. Assuming that, as far as the analyst knows, there was no change in the circumstances, there is no reason to expect a sudden real change in precision. The analyst should take the view that a given procedure, in competent hands, has an inherent precision which can be ascertained. Individual small sets of data will inevitably give estimates of the standard deviation that show considerable variation. This variation in the individual estimates of the standard deviation is natural, however surprising it may seem. Once sufficient repetitions have been accumulated, say 30 or more pairs of duplicates on samples not too widely spread in content of the element, an estimate of the standard deviation can be obtained that should be used in place of any estimate based on some small set of data. Of course, something can go wrong and sometimes does. There are statistical criteria for suspecting out of line results. If the difference between a pair of duplicates is exceptionally large, this is taken as evidence of a mishap. In that event additional determinations are in order.

Once it is accepted that differences in precision between laboratories can be forgotten because, if present, they are probably minor differences anyway, the way is open for a revealing examination of the data. In any event the evidence is conclusive that differences in the systematic errors are the major source of disagreement among laboratories. Certainly, if this were not the case, the whole edifice of standard samples would be without value. Obviously the use of a standard sample to check out a procedure can in no wise alter the *precision* of the analytical work. A standard sample may direct the attention of the analyst to the need to go over his procedure. Rarely will the measures taken make any difference in the agreement of check determinations. If poor agreement between duplicates were the real trouble, the analyst could use improved agreement between duplicates as a criterion of satisfactory results and dispense with standard samples. This is only saying what every analyst knows: Good agreement between duplicates is a necessary but not a sufficient condition for a good procedure.

Systematic Errors

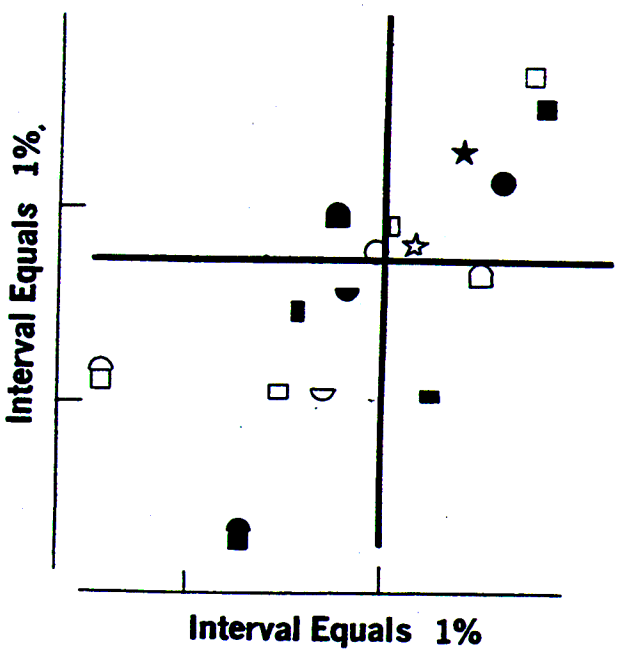
Just as a given analytical procedure may have a certain precision associated with it as a property of the over-all ensemble of operations

involved, so may the procedure itself be thought of as having a built-in systematic error. It is a common remark that this, or that, method tends to give high (or low) results. Obviously gravimetric procedures are vulnerable to low results if the precipitates are too soluble. Very often, in analytical procedures, a blank is specified and clearly this is intended to correct for a systematic error that would otherwise be present. The chemist's goal is to devise procedures that are inherently without any built-in systematic error or bias. It is usually considered sufficient to reduce the systematic error to the point where it is small relative to the precision error.

The systematic error of a procedure is a property of the procedure when performed as specified. Departures from the specified procedure may substantially modify the original bias. Sometimes a laboratory with the best intention of correcting a suspected bias may overshoot and even change the sign of the systematic error. In any event there is no question but that the procedure modifications and the equipment and reagents associated with each laboratory do result in a corresponding gamut of laboratory systematic errors that modify the basic systematic error of the procedure. Considerable advantage follows from accepting this picture of the structure of the systematic error. In the first place the true chemical composition may not be known. All that can be done then is to take as a working reference point the consensus of the participating laboratories. Individual laboratory systematic errors can, in fact, be measured only from this consensus reference point. A particular laboratory that is far out of line may be presumed to have departed from the accepted procedure in a unique way. In the absence of any other guide, the consensus of a reasonable number of laboratories may be taken to characterize the analytical procedure. After all, the laboratories are expected to follow the procedure. At a later date, an opportunity may arise to try the procedure on materials of known composition. Any discrepancy between the true composition and the consensus of the laboratories must be considered a defect in the procedure.

The essential point is that when this way of looking at the systematic error is "simplified" by concentrating attention directly on the difference between each laboratory's own average and the known composition, useful information is lost. Suppose that the systematic error for the procedure is positive and that one laboratory departs from the consensus by a nearly equal negative systematic error. This particular laboratory then has a practically perfect check with the true composition and therefore swears by the procedure. There are some omitted words here. The laboratory swears by the procedure *as carried out by that laboratory*. That does not advance matters at all unless we know, or can find out, in what respects this laboratory departed from the specified procedure. This may be a significant deliberate departure and ascertainable or it may be a chance departure dependent upon the reagents, apparatus, etc., that were used by this laboratory. In all fairness, each laboratory should be judged by its closeness to the consensus, if we have any confidence that the participating laboratories conscientiously tried to follow the procedure in every detail. The discrepancy between the consensus and the true value ought to be charged to the procedure.

The consequence of this point of view is that laboratories close to the consensus deserve pats on their backs. A laboratory whose result departs from the consensus should be called to account even when it *happens* to check the true composition. If the laboratory deliberately departed from the procedure it should share this knowledge, and also simultaneously admit that it did not adhere to the agreement to test the procedure as given. If every laboratory departs capriciously from the procedure as specified, then the whole business of interlaboratory testing might as well be forgotten because no single version of the procedure can be tried



Persistence of systematic errors is shown in two series of analyses run by the same 8 laboratories. Each laboratory is shown by a different symbol. The solid symbols refer to the first series and the open symbols the second series

out. If the laboratory has no reasonable explanation to offer for the good check it got, when the consensus of all was clearly not a check, there seems no more reason to congratulate this laboratory than a laboratory that had an equally large deviation from the consensus but in the opposite direction. After all, if chance is operating in the events that introduce laboratory systematic errors, maybe the chances of a plus or negative systematic error are not too different. So one laboratory, judged by the true composition, looks very good, another very bad when perhaps both laboratories have substantial defects in their reagents or apparatus.

When all, or nearly all, the results from a particular laboratory deviate in the same direction from the known composition, the evidence of a systematic error in the results is unmistakable. The advantage of remembering the possible, and likely, composite character of the systematic error, lies in the steps that may be taken to achieve better results. The procedure may require modification. Certain laboratories may need to mend their ways. The desired end is one where all the laboratories cluster closely about their consensus combined with close agreement of the consensus with the known composition. In fact, it can hardly be maintained that an agreed upon procedure exists unless the laboratories can achieve good agreement among themselves around *some* value. Once this stage has been reached, it will improve the chances of successfully locating the cause and remedy for a discrepancy between consensus and true value.

Separation of Systematic and Random Errors

Very few data suffice to demonstrate the presence of individual systematic errors for laboratories and to provide an estimate of their common precision (*2–4*). Two fairly similar materials, not very different in percentage of the element to be determined, will be required. These conditions are stipulated because the precision as well as the systematic error may depend on the per cent of element present and possibly be changed if interfering substances are present. Only one determination is necessary on each material by each of a number of laboratories. If duplicates are run, the averages will be used. Let the materials be designated X and Y. The laboratories are numbered 1 to n, and the results symbolized as $x_1, y_1; x_2, y_2; \ldots; x_n, y_n$. A pair of coordinate axes should be drawn on a piece of graph paper. A scale of values is laid off on the x-axis covering the range from the lowest value reported for X to the largest result. Using exactly the same unit, the scale of values on the y-axis must cover the range from the lowest value for Y to the highest result. Usually the scale is so enlarged that the smallest division on the graph paper corresponds to one unit in the last place of the values reported.

The pair of values furnished by a laboratory determines the location of a point on the graph paper. There will be as many points as there are participating laboratories. A horizontal line is located through the average (consensus) of the values reported for Y and a vertical line drawn through the average of the values reported for X. These two lines divide the graph paper into four quadrants. The pair of deviations from the averages, associated with a laboratory, must be either ++, +−, −+, or −−, and these correspond to the four quadrants just formed. If plus and minus deviations from the average of each material are equally likely, then the four combinations, ++, +−, −+, and −−, are equally probable so that, in theory, equal numbers of points should fall in the four quadrants. This distribution of the points would not be changed even if the laboratories did have different precision, because the signs, and not the magnitudes, of the deviations determine the quadrant getting the point.

Examination of scores of such charts has shown in almost every chart an unequal division of points among the quadrants. Two of the quadrants, the upper right corresponding to ++, and the lower left, corresponding to −−, contain a majority of the points. The explanation for such a departure from theory is immediate. If a laboratory does have a systematic error, this error, by definition, appears in both the result for X and the result for Y. While the random errors may be of opposite sign, the deviations will be converted to the same sign if a large enough systematic error is added to, or subtracted from, each random error. The results reported by the laboratories show only the net remaining after random and systematic errors have been combined. The signs give the show away and the surplus of points in the ++ and −− quadrants is graphic testimony of the presence of systematic errors.

Analysts like to dream of a world in which only random errors exist, and small ones at that. Consider the contrary world where perfect precision exists but each laboratory has persistent individual systematic errors. This would mean that if a laboratory's result for X is higher by 0.10% than the consensus for material X, then on material Y it will be exactly 0.10% higher than the consensus for Y—exactly the same amount higher on both materials because of perfect precision (sampling errors assumed not present). In this contrary world all the points would lie precisely on a 45° line passing through the point where the horizontal and vertical lines intersect. Perfect location of all points on such a line has not been observed, but some distressingly near approximations have been encountered.

Most interlaboratory studies yield plots that are intermediate in character between the two extremes of equal numbers of points in the four quadrants and all the points in the ++ and −− quadrants. The points scatter in an approximate ellipse whose long axis is the 45° line through the point corresponding to the averages for X and Y. The larger the systematic errors, relative to the precision error, the more elongated and thinner the ellipse will be. When the points do straggle more or less closely along the 45° line, the evidence for an unsatisfactory procedure is conclusive. Possibly the procedure is in-

adequately described and is so vulnerable to individual interpretation that, as a group, the laboratories are having trouble. On the other hand, if a substantial majority of the points are clustered in a fairly broad ellipse with only a few points far out along the 45° line (either in the ++ or −− quadrants), there is a strong suspicion that the more remote laboratories have their own unique way of making the determinations.

An excuse often advanced by a laboratory with an out of line result is the claim that it got a nonrepresentative sample. This claim is considerably weakened when the laboratory's point is far out and near the line, because now the laboratory has to claim nonrepresentative sample for both materials, and, furthermore, departing in the same way. An even stronger objection can be put forward against this claim. If the materials sampled are not uniform, then, in taking the samples of X, half of the samples will be high and half low. This is also true for material Y. The two samples sent, quite blind, to a laboratory may be high in both (++); high in X, low in Y (+−); low in X, high in Y (−+); or low in both (−−). All combinations are equally likely, so that if the lack of uniformity of the stocks is sufficient to dominate over the systematic errors, then the points should be equally distributed among the quadrants. The argument is now turned in reverse and a lack of equal distribution among the quadrants considered evidence that sample variation is not the problem.

There is a possible ambiguity. Either very poor precision or nonuniform material may lead to an equal distribution among the quadrants. The allocation of samples may be modified to resolve this ambiguity if desired, but the event has not been observed, so means to distinguish between these causes will not be given.

Earlier mention was made that in the event of perfect precision the points would lie exactly on the 45° line. Random errors displace the points from the line. The perpendiculars from each plotted point to the 45° line are a means of estimating the precision of the procedure as revealed by the combined results from the participating laboratories. Designate the lengths of the perpendiculars by $p_1, p_2, \ldots, p_n$. Then an estimate of the common standard deviation is given by

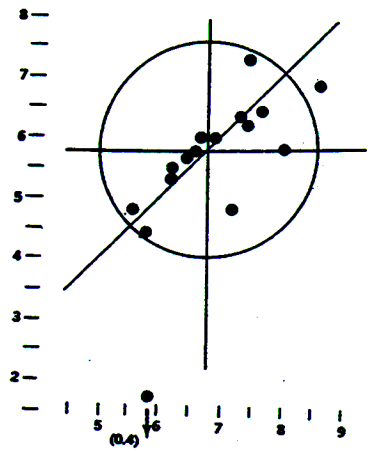
$$s = \sqrt{\Sigma p^2/(n-1)}$$

Some readers may be interested to show that this formula is equivalent to

$$\sqrt{\frac{\Sigma d^2 - n\bar{d}^2}{2(n-1)}}$$

Here each d is the difference between the result reported for X by a laboratory and the result reported for Y by the same laboratory. The algebraic average of these differences gives $\bar{d}$.

Each laboratory provides a perpendicular. Measure the distance along the 45° line from the foot of the perpendicular to the point corresponding to the averages for the two materials. This distance, divided by the $\sqrt{2}$, gives the best estimate of the systematic error of the laboratory measured relative to the consensus of all the laboratories. If the true compositions of the materials are known, they may be used to plot a point. The distance along the 45° line to the true point divided by $\sqrt{2}$ gives an estimate of the systematic error of the procedure as used by the participating laboratories.



The extensive range of systematic errors noted in results by a large number of laboratories all analyzing the same sample of phthalic anhydride indicates the possibility of a faulty procedure

Number of Laboratories Required

The small amount of work called for from each laboratory should make it easier to enlarge the number of participating laboratories over the usual handful. Much can be said in favor of a large number of participating laboratories. Information regarding the prevalence of systematic errors can be obtained only by having enough laboratories to reveal them and to estimate fairly, by their consensus, the systematic error of the procedure. There is another easy way to enlarge the number of points. An additional pair of different materials, still rather similar to the first pair, are sent to the same laboratories. The results are used to prepare a second graph. The second graph is placed on the top of the first graph, so that the horizontal and vertical lines are coincident and all the points transferred to one graph. This merely gives a common consensus point. As the true compositions have not been used, the absolute values are not involved. If the true compositions are known, the common graph is prepared by plotting the true point on each graph and superimposing these points. The axes are kept parallel. Laboratory numbers should be attached to the points. If a laboratory has both its points far out along the 45° line, the conclusion is obvious to all concerned.

The whole process should be repeated with materials having very different per cent values of the element to be determined. A separate estimate of the precision is proper and should be made. Indeed the systematic error of the procedure may change and possibly that of individual laboratories. The range of per cent and types of materials that require study depend on the analytical chemistry involved.

Discussion

The economy of effort achieved by the elimination of duplicates and other ramifications such as an elaborate schedule of operators, days, etc., is considerable. More important, the rather spurious yardstick of parallel duplicates by the

same operator is discarded. Parallel duplicates are favored indeed. Whatever the attendant circumstances, these duplicates have everything in their favor as far as showing agreement is concerned. Just what use can be made of such a yardstick? Nearly every practical comparison involves determinations carried out under less uniform conditions than a pair of parallel duplicates. Even the single analyses on the two materials are likely to be run together, so that there is the same criticism to be directed against using these to estimate precision. The two materials would be better run at least on different days. Figure 2 shows a plot of potassium determinations by 14 laboratories on two samples of fertilizer. The two samples were run a month apart, so that the estimate of precision is realistic. The clear evidence of individual systematic errors in materials run a month apart shows the persistence of systematic errors.

The estimate of precision proposed here is usually optimistic. A laboratory runs two materials, no doubt under parallel conditions. The two results provide an estimate of the *difference* between the two materials. When the difference is taken between the two results, any common effects drop out, so that the difference is in large measure freed of any consequences of the particular set of circumstances existing when this pair of determinations was made. Every laboratory provides an estimate of the difference and the estimate of the precision is based upon the concordance of these several estimates of the difference between the two compounds. The most that can be said in support of this scheme is that, unlike duplicates on one material, the laboratories do not know the difference between the two materials. There is no protection against a laboratory that runs two or more determinations on each material and reports the averages of these under the label that they are single determinations. Eventually, if over a number of times, a given laboratory always has a point unusually close to the 45° line, it might reasonably be asked to disclose how it consistently achieves a precision so much better than other laboratories.

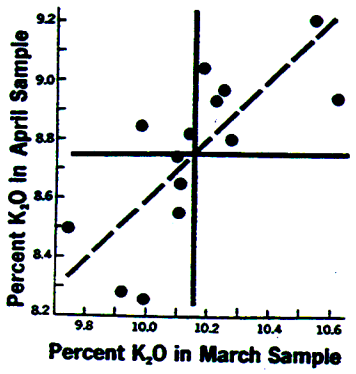


Figure 2

Very careful efforts on analytical work are associated with atomic weight determinations and with the work on standard samples or reference materials. The approach here is chemical rather than statistical. Using every iota of available chemical information elaborate precautions are taken to eliminate, or correct for, every possible source of systematic error. Comparatively little dependence is placed upon repeat determinations. Here the chemist supplies his own testimony to support the position taken in this paper. Systematic errors are the real headache. If enough care is taken, or alternative procedures are employed, the systematic error can be greatly reduced. By such means atomic weights and standard samples gain acceptance. In the ordinary work of analytical chemistry, most of these precautions are not feasible. Nevertheless the goal of general agreement among laboratories, using a procedure with a very small bias, is the task of the analytical laboratories. To achieve their goal, the laboratories must get the right kind of data and interpret them properly.

Literature Cited

(1) Power, F. W., ANAL. CHEM. 11, 660 (1939).
(2) Youden, W. J., *Ind. Eng. Chem.* 50, 83 A (August); 91 A (October); 77 A (December, 1958).
(3) Youden, W. J., *Ind. Quality Control* 15, No. 11, 24 (1959).
(4) Youden, W. J., *Technometrics* 1, 409 (1959).