STATISTICAL ASPECTS OF THE CEMENT TESTING PROGRAM*

By W. J. Youden[1]

Synopsis

A statistical technique has been developed for increasing the amount of information obtained in interlaboratory test programs. The main feature of the technique consists in a graphical presentation of the results reported for pairs of samples. The graphs make for easy identification of laboratory bias and provide a method for estimating the precision of the test procedures.

A rating system has been established to facilitate the examination of a large mass of data. The theoretical distribution of average scores was computed in order to provide a yardstick of judgment for the observed distribution of average scores.

When a reference sample is tested by a considerable number of laboratories, one consequence is to make it possible for each laboratory to find out how its own result compares with the results obtained by all the other laboratories. Unfortunately, relatively little can be learned from the collection of results reported for one reference standard. Examination of the results usually shows one or more absurd results that presumably are caused by slips in computing, typing, or misreading of an instrument. Generally these extreme values are fairly obvious and may safely be set aside. The next step is to obtain the average of all the remaining results and perhaps some measure of the scatter or error.

It should be remembered that a collection of results from different laboratories is not the same as a collection of repeat tests on one material by one laboratory. In the latter case any systematic error that may be associated with the laboratory's technique for conducting the test will be present in every result as a "constant" error. The observed scatter in these repeat determinations arises from *random* errors and reflects the *precision* of that laboratory. Any one result from that laboratory is a composite of the unknown true value, μ, the unknown constant error for that laboratory, L_i, and a random error e_i.

A collection of results, one from each of n laboratories testing the same material, will have the following composition

$$\mu + L_1 + e_1$$
$$\mu + L_2 + e_2$$
$$\vdots$$
$$\mu + L_n + e_n$$

The scatter exhibited by these quantities may arise from a complex of circumstances. Each laboratory will have its own systematic error, L. In addition

* Presented at a special session on Statistical Treatment of Interlaboratory Test Results, Including a New Graphical Method sponsored by Committee C-1 on Cement held on June 25, 1959 during the Sixty-second Annual Meeting of the Society.

[1] National Bureau of Standards, Washington, D. C.

the precision may vary from laboratory to laboratory so that each random error comes from a different normal distribution. Given a collection of results, one from each laboratory, there is no way whatever to unravel this complex of circumstances. Given a parallel set of results on a second sample, some progress can be made, if certain assumptions are made. These assumptions are (1) that each laboratory conveys its particular systematic error equally to both materials and (2) that for all practical purposes, differences in precision among the laboratories can be ignored.

By definition a systematic or "constant" error should persist in all results made by the laboratory, provided the materials are fairly similar and do not differ too much in the magnitude of the property tested. If these conditions do not hold and the systematic error varies with the type of material, the complications multiply. There is abundant evidence in the results reported by Blaine and Crandall (1)[2] that systematic errors do vary from laboratory to laboratory and that the systematic error of a laboratory is conveyed to a series of materials. Even more important the evidence is conclusive that when a laboratory does depart markedly from the concensus of the other laboratories, the systematic error is responsible in most cases. The identification of systematic errors as the major cause of divergent results makes the role of precision relatively minor. Perhaps this is fortunate because the detection of even substantial differences in precision between two laboratories requires a score or more of repeat determinations from each laboratory.

The collection of results, one from each of n laboratories, reflects a collection of n systematic errors, some small, some large. The magnitudes of these systematic errors may also be distributed in a manner approximating the normal. The fact that the gross results (which include both systematic and random errors) usually approximate the normal distributions supports this view. It is not surprising that among the systematic errors there may be a few exceptionally large ones. The results with large systematic errors contribute a substantial, sometimes even a major, portion of the sum of the squared deviations used in calculating the standard deviation. This inflation in the sum of the squared deviations gives a value for the standard deviation that is larger than it should be, larger than the one actually being achieved by possibly 95 per cent of the laboratories.

There would be some justification in resigning oneself to accepting this overlarge estimate of the standard deviation if the divergent results arose from chance or random errors. In that event, these large deviations would, in repeat studies, turn up from different laboratories. The study with a dozen cements shows this not to be the case. The same laboratories are identified with the large deviations, the sign of the deviation remaining the same. As soon as certain laboratories discover that they have, somehow, acquired rather large systematic errors, the way is open for corrective measures to be taken. In fairness to the test procedure, any appraisal of its merits should not include the work of laboratories that reflect persistent large systematic errors because these greatly inflate the estimate of the standard deviation. Finally, any progress toward an over-all improvement in testing is most easily achieved by a few laboratories taking corrective action.

THE PROBLEM OF DIFFERENTIATING AMONG LABORATORIES

Suppose we have 100 laboratories, all without systematic errors, that is, only

[2] The boldface numbers in parentheses refer to the list of references appended to this paper.

random or precision type errors. First imagine that all have the same precision in conducting the three-day tensile strength test. If the tensile strength is taken as 440 psi and the standard deviation for all the laboratories is 40 psi, it is an easy matter to calculate the expected distribution of the test results. Statistical tables show that four laboratories may be high by 70 psi or more, and four laboratories low by 70 or more psi.

TABLE I.—PREDICTED DISTRIBUTION OF RESULTS REPORTED BY 100 LABORATORIES. ASSUMPTION A: ALL LABORATORIES HAVE STANDARD DEVIATION OF 40 PSI. ASSUMPTION B: ONE-THIRD OF THE LABORATORIES HAVE STANDARD DEVIATION OF 30; ONE-THIRD A STANDARD DEVIATION OF 40; AND ONE-THIRD A STANDARD DEVIATION OF 50.

Tensile Strength, psi	Predicted Number of Laboratories	
	Assumption A	Assumption B
510.0 and more	4.0	4.3
490.0–509.9	6.5	6.1
470.0–489.9	12.1	11.6
450.0–469.9	17.5	17.8
430.0–449.9	19.8	20.5
410.0–429.9	17.5	17.8
390.0–409.9	12.1	11.6
370.0–389.9	6.5	6.1
less than 370.0	4.0	4.3
	100.0	100.1

The second column in Table I shows the expected number of laboratories obtaining results in various ranges of breaking strengths. When the results are reported back to the laboratories, those close to the average will undoubtedly feel rather pleased. No grounds whatever exist for this self-congratulation because repeat tests will find the participating laboratories shifting up and down the scale in a chance manner. Remember that the table was calculated on the assumption that *all* the laboratories really have the same precision (and systematic errors were absent.)

Now imagine an alternative situation whereby one third of the laboratories have a standard deviation of 30 psi, another third a standard deviation of 40 psi, and the remaining third a standard deviation of 50 psi. Again the expected distribution of laboratories can be calculated and it is clear from the third column in Table I that the scatter of the results is more or less indistinguishable from that found when all the laboratories had a standard deviation of 40 psi. But it may be noted that under assumption B there is a slightly greater concentration of results in the vicinity of the mode, and in the extreme "tails," with corresponding deficiencies at moderate deviations. Such greater concentration near the mode, balanced by larger "tails," is a characteristic of distributions composed of results of unequal precision, and in practical work may usually be taken as an indication of the presence of results of unequal reliability. Unfortunately, there is no trustworthy way of identifying in such a distribution of single results which correspond to the higher precisions and which to the lower. Laboratories close to the average may be happy about it, but no way exists to justify their satisfaction using only this one set of results.

If repeated test results are available, then the way is opened to differentiate among the groups. Ultimately membership in one of the 30, 40, and 50 psi standard deviation categories could be established. But even if as many as 25 repetitions of the test are run by each laboratory, the discrimination is rather poor. For example, one out of five of the 40 psi standard deviation laboratories will be credited with a standard deviation of 35 or less and hence be classified as belonging to the 30 psi group. And a second one of the five will have the bad

luck to have its estimate be 45 or more and be put down as belonging to the 50 psi group. Discrimination among laboratories as to precision is not easy. On the other hand, constant errors very soon reveal themselves. Thus, a constant error equal to the standard deviation will lead to a preponderance of deviations of the same sign, the expected proportion being 0.84.

THE TWO-SAMPLE PROGRAM

The prevailing presence of systematic errors in test results can be demonstrated if a number of laboratories perform only one test on each of two test materials. A simple graphical representation (2) of the n pairs of results obtained from the n participating laboratories suffices to show that systematic errors are present. On a piece of graph paper draw the usual x and y axes. On the x axis lay off a scale of values covering the range of values found for one of the materials. Lay off a similar scale using the same unit on the y axis. The pair of values for the two materials reported by a laboratory may now be used as coordinates to determine a point representing the work of that laboratory. A point is plotted for each laboratory. The collection of points is then divided by a horizontal line so that half the points are above the line and half below. A vertical line is also drawn dividing the points half to the left of the line and half to the right. These two lines partition the graph paper into four quadrants. The upper right quadrant corresponds to a region where points represent values greater than the median for both materials. The lower left quadrant is a region where the points represent values less than the median for both materials. The other two quadrants provide for values, one of which is greater than the median and one below the median. The two quadrants correspond to the two ways this can happen.

The median is less disturbed by extreme values and is convenient because counting is quicker than computing the averages. The intersecting lines may be drawn through the point determined by the average for each material if that is preferred. Usually this makes an imperceptible difference in the location of the lines. When there is a difference in location, it comes about from the inclusion of one or more "wild" values in the averages and the median values are preferable.

Results greater than the median are considered to give plus deviations and results less than the median are taken to give negative deviations. The four quadrants thus are identified with the four possible paired combinations (++, +−, −+, −−) of deviations from the averages. If only random errors of the precision type are present in the results, the points may be expected to be distributed equally among the four quadrants because positive and negative deviations are equally likely. If systematic errors are present, the effect is to shift the points into the upper right and lower left quadrants. No matter where a point is located in the +− or −+ quadrants the addition (or subtraction) of a sufficiently large constant error will shift the point into a ++ (or −−) quadrant. The argument is then applied in reverse. If points are found predominantly in the ++ or −− quadrants, the presence of systematic errors is established.

It is instructive to consider results that are without systematic errors. The points would then be clustered compactly in a circular pattern centered on the intersection of the two lines. If systematic errors of varying magnitudes, both plus and minus, are now added to the results, the points move outward from the center more or less closely along a

45-deg line drawn through the intersection of the two lines. In fact, points will be off this 45-deg line solely because of the precision error that scattered them in the original compact circle.

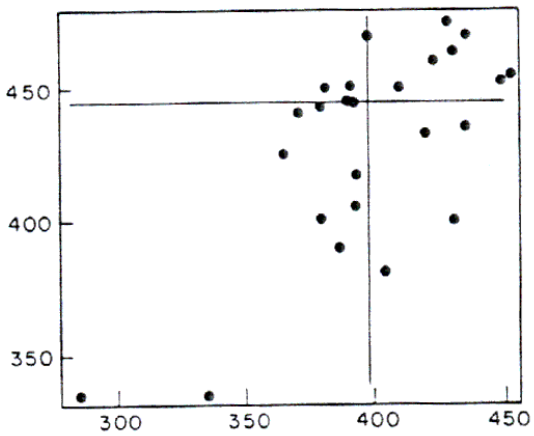

FIG. 1.—Tensile Strength, psi.

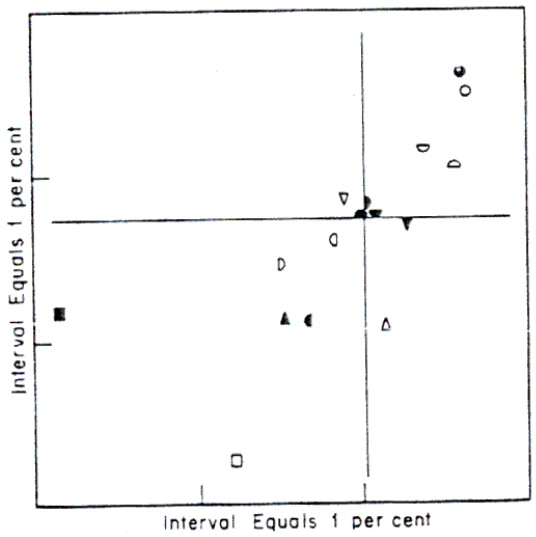

FIG. 2.—Calcium Oxide in Cement.

Figure 1 shows a plot of seven-day tensile strengths reported by 25 laboratories on two cements. An impressive majority of the points lie in the ++ and −− quadrants. The most extreme points are displaced along the 45-deg line. Certainly the two lower points provide a strong hint to two of the laboratories that they are troubled with large negative systematic errors. The possible contention of nonuniform material and sample troubles can be dealt with quickly. The materials must be satisfactorily uniform, otherwise there would exist high and low samples. High and low samples will be equally frequent. When the pairs of samples are sent out, each laboratory would receive one of the four sample combinations, ++, +−, −+, −−. These four combinations are all equally likely and all four quadrants would then be about equally populated with points. This equality of distribution is not present so that it is concluded that the material is reasonably uniform.

Figure 2 shows the results from eight laboratories reporting on calcium oxide in cement on two sets of paired samples. The solid symbols identify the laboratory results on the first pair of materials analysed. Two months later another pair of materials were analysed and the hollow symbols were plotted on another piece of graph paper using the appropriate range of values but the same unit interval for 1 per cent of calcium oxide. The two graphs were then superimposed so that the quadrant forming lines were coincident and all the points plotted on one graph. The configuration of the points, not the absolute values, is the important aspect of the plot. Notice how the points congregate in the ++ and −− quadrants. Clearly the circle laboratory and the square laboratory run true to form. The evidence is convincing that the circle laboratory gets results that are consistently higher than those obtained by other laboratories and the square laboratory gets low results. All this is established with just four results from each laboratory. The 45-deg line provides a reasonably good fit to the points so that some of the other laboratories undoubtedly have smaller systematic error that are, in part, obscured by the random precision error.

Practically all tests gave patterns which showed a generally elliptical character with the long axis of the ellipse at 45-deg through the point plotted for the median (or average) values for the two materials. The degree of eccentricity of the ellipse is informative. The more elongated ellipses indicate that the test procedure is very vulnerable to individual interpretations or modifications that introduce systematic errors into the test procedure. A more careful specification of the test procedure or perhaps some change in the procedure may be needed to correct the situation.

THE IMPROVEMENT OF TEST RESULTS

The preceding section pointed out two ways in which progress may be made in reducing the scatter of the results reported by laboratories testing the same material. A relatively small amount of data, plotted as a joint scatter diagram for two materials, permitted the consideration of all three elements that have to be considered in seeking the causes for an undesirably large dispersion among the reported results. First, some assurance is needed that lack of uniformity in the material distributed is not mainly responsible for the scatter. This is a problem of some difficulty where the samples are perforce large and the laboratories numerous. Second, the identification of systematic errors as the cause for the more divergent results is easily made and the singling out of the responsible laboratories is a simple matter. Finally, faulty test procedures are likely to be the trouble when most of the points determined by the test results are strung along the 45-deg line in the graph. Improvement can only come by obtaining satisfactory evidence of the presence and causes of errors in test results and then taking whatever remedial steps are needed.

SCORING SYSTEM FOR TEST RESULTS

At the outset of the work a computation procedure was instituted to calculate the standard deviation for each test for each cement. The 103 laboratories available provided an unusually large basis for calculating the standard deviation of the hundred or so results as reported. Only the most obvious blunders were deleted before making the computation. It was realized that a more careful screening of the data was desirable, but there was considerable hesitation about deleting results and thereby obtaining an estimate of the standard deviation that might be too small. In the scoring scheme about to be described the above standard deviation was used with the idea of determining whether the distribution of the test results could be satisfactorily described using this estimate of the standard deviation. The scoring scheme also greatly reduced the amount of arithmetic required in dealing with over 24,000 test results. The scores made for simplicity in comparing tests because the individual units used in reporting results were replaced by multiples of the appropriate standard deviation.

TABLE II.—SHOWS HOW (1) SCORE DEPENDS ON DEVIATION FROM THE AVERAGE AND (2) THE PROBABILITY OF ACHIEVING THE SCORE.

Difference Between Result and Average	Score	Probability
0–1.0 σ	4	0.69
1.0 σ–1.5 σ	3	0.18
1.5 σ–2.0 σ	2	0.09
2.0 σ–2.5 σ	1	0.03
over 2.5 σ	0	0.01

A test result within one standard deviation of the average was given a score of 4—plus if high, minus if low. Results more than one standard deviation away but within one and a half

standard deviations were given a score of 3. The complete scoring system is shown in Table II.

The appended column of probabilities gives the chances that a laboratory will receive any one of these five scores if the laboratory has in fact a standard deviation equal to that computed from all the data. One laboratory in a hundred may deviate by 2.5 σ, and receive a score of zero, even if its work is up to par. A single set of scores is not very informative. If as many as ten scores have been awarded, then the average score becomes meaningful. A laboratory might have an average of 4.0 for the ten physical (or chemical) tests on one cement but only if it scored 4 on every test. The chance of this happening is the tenth power of 0.69 or 0.0245. One in forty laboratories will amass such an impressive average, even though its work is no better than the standard deviation used in setting up the scoring. The numerical work is tedious but the chances of getting various average scores based on ten tests can be computed. These chances are displayed in Table III, column 2, which gives the expected number of laboratories per hundred making each average score.

TABLE III.—EXPECTED NUMBER OF LABORATORIES PER HUNDRED MAKING INDICATED AVERAGE SCORES BASED ON TEN TEST RESULTS.

Average Score	Number of Laboratories		
	Computed σ	80 per cent of Computed σ	Adjusted to 94 Laboratories
4.0	2.45	9.35	8.78
3.9	6.38	17.77	16.00
3.8	10.68	20.89	19.64
3.7	13.77	18.74	17.62
3.6	14.92	14.00	13.16
3.5	14.10	9.05	8.52
3.4	11.93	5.22	4.91
3.3	9.18	2.72	2.56
3.2	6.51	1.30	1.22
3.1	4.28	0.58	0.55
3.0	2.55	0.23	0.22
2.9	1.52	...	...
2.8	0.82	...	...
2.7	0.43	...	...
2.6	0.21	...	...
Total	99.73	99.85	93.88

The computed frequencies, listed in column 2, Table III are represented graphically by the outline bars in Fig. 3. When this theoretical curve was compared with the observed distribution of the scores made by the 100 laboratories, pronounced discrepancies were found. Invariably the theoretical curve underestimated the number of high scores. Also, in theory, in studies of this size, no scores were expected below 2.6. In fact, several such scores were usually found. Computations showed that the deletion of six low-scoring laboratories would reduce the standard deviation by at least 20 per cent. If the smaller standard deviation is used, the chances of achieving various scores on a single test changes from those given in the last column of Table II to the following:

Score	Probability
4	0.789
3	0.150
2	0.048
1	0.011
0	0.002

The chance of a laboratory achieving an average of 4.0 is now the tenth power of 0.789 or 0.0935. The number of laboratories out of 94 that can be expected to score 4.0 is 94 × 0.0935 or 8.78 laboratories. The expected number of laboratories obtaining various scores using this smaller standard deviation are shown in the last column of Table III. The expected number of laboratories making these scores are shown by the solid bars in Fig. 3. The contrast in the two distributions is pronounced especially in the sharp increase in the ex-

pected number of laboratories making high scores. The solid bar chart is a much better approximation to the observed distribution of laboratory average scores and is confirmed by the high proportion in the scoring. All in all the evidence is unmistakable that about 90 per cent of the laboratories are performing the test in such a way that the over all scatter of their results is associated with a standard deviation about 20 per cent below that calculated from all the laboratories. (All the laboratories here means after discarding two or three patently absurd results arising from what appear to be typing or computing slips.)

The distribution of the averages of *ten* scores was chosen because initially averages were taken of the ten physical (or ten chemical) scores by a laboratory on one cement. There were twelve cement samples but sometimes a pair was skipped. The theoretical distribution for averages based on twelve results differs very slightly from that given for averages of ten.

The standard deviation is a property of the test procedure, provided the test is performed with due care with the specified equipment. A few laboratories by lack of care or by unwitting departures from the specified method of performing the test can obtain results suffi-

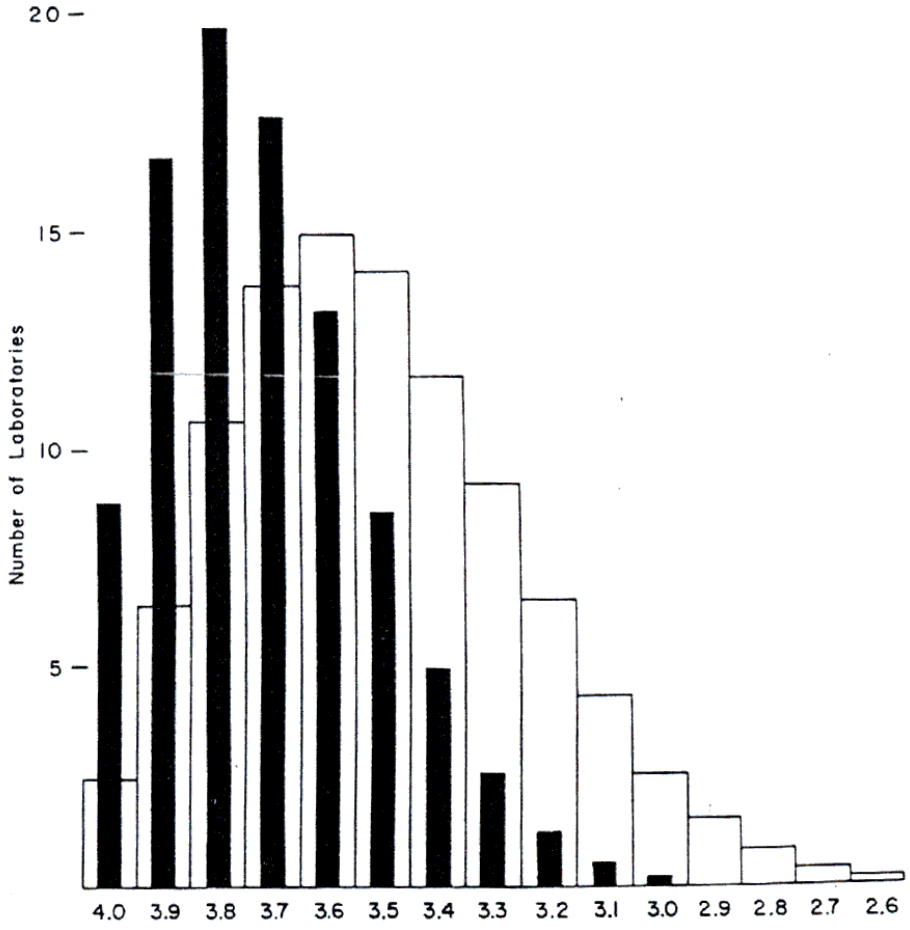

FIG. 3.—Average Score for Ten Results.

ciently out of line to inflate substantially the estimated standard deviation. The present study has established beyond any possibility of controversy that a handful of laboratories is responsible for making the test procedures appear substantially less satisfactory than they really are.

SUMMARY

A graphic procedure using paired samples has been developed to demonstrate to what extent systematic errors are present in interlaboratory tests. This procedure also indicates the extent of random errors and the ultimate precision of a test procedure.

A scoring system for evaluating the test results of a group of laboratories has been developed. It has been shown that the results of a few laboratories doing poor work can substantially inflate the apparent standard deviation obtained for any test.

REFERENCES

(1) J. R. Crandall and R. L. Blaine, "Statistical Evaluation of Interlaboratory Cement Tests," see p. 1129, this publication.

(2) W. J. Youden, *Industrial and Engineering Chem.*, Vol. 50, p. 63a, Aug.; p. 91a, Oct.; p. 77a, Dec. 1958.