THE RELIABILITY OF LABORATORY RESULTS

A PROCTER & GAMBLE CONTRIBUTION

When the laboratory reports that the titer of a stock is 40.3° C., that the moisture in a sample of soap flakes is 16.1%, that an I. V. is 44.6, or that any analysis gives a certain result, it is understood that there is a

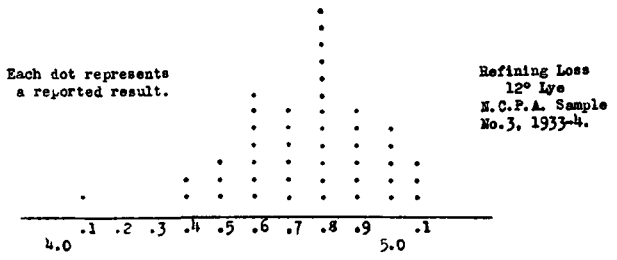


FIG. 1.

certain tolerance or "error" in the figure reported. That is, if we asked the laboratory to repeat the determination on another portion of the same sample, a result differing from the first by a small amount would be obtained; for instance, if the laboratory had reported a 7.9% refining loss and, upon repeating, the loss was 7.6%, we would say, "That's a good check." But suppose the second value were 7.3. Is this value sufficiently different from the first result, 7.9, to warrant suspicion of the validity of either figure? To settle such a question, we need some appreciation of and a quantitative measure for the variations not caused by mistakes or poor work on the part of the analyst.

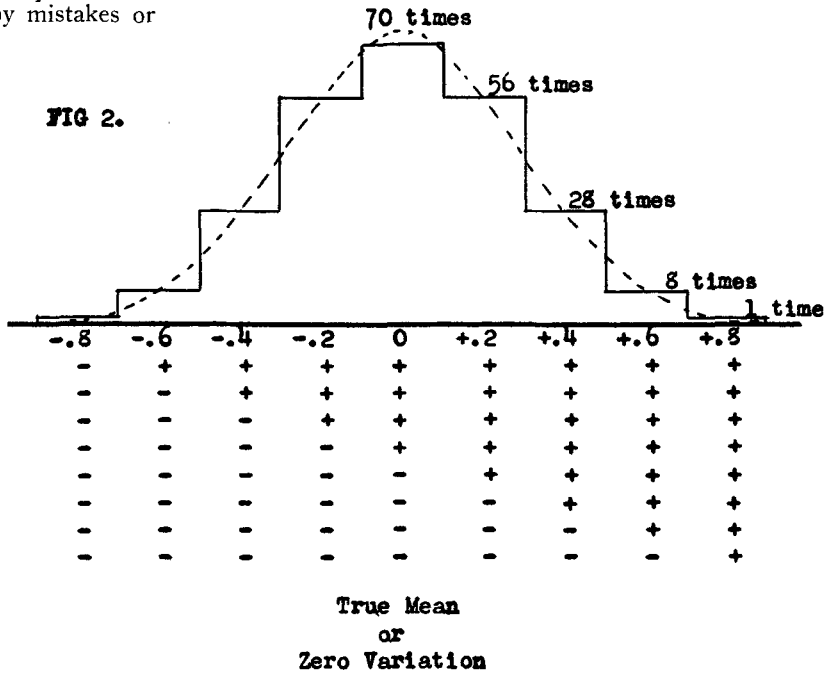
Figure 1 presents a picture of the scattering of results when a large number of analyses are run.

It is not surprising that results do not check exactly. The final figure that is reported as the result of an analysis is usually made up of a number of measurements, each of which involves an estimate of the relation between a variable line and a graduated scale. The variable line may be a meniscus, the shadow on a refractometer, the pointer on a balance, or the mercury in a thermometer. There is always some "estimating" involved in deciding upon the location of the moving line. Moreover, the nature of the operations carried out in the laboratory is such that they do not go to precisely the same degree of completion under the conditions that we are able to maintain. A screen test for fines is an extreme example of this variation in completeness. In making a test for fines we shake the screen in a manner that causes the fine particles to work their way down thru the coarse particles and out thru the screen. Obviously, in any test, a few fine particles may remain wedged between the coarse particles. If we attempt to carry the test to such a length that all fine particles are shaken thru the screen, we may break up or wear out the coarse particles and get erroneous results. Consequently, we have to set up a standard method and a standard time of shaking which attempts to hit the line between fine particles left behind and fine particles broken off or the coarse ones. It is easy to see that in such an empirical test two portions from the same sample jar may not behave precisely the same. In somewhat the same manner certain crystallizations and reactions, such as titers and iodine values, may fail to repeat themselves exactly and precisely when carried out on duplicate samples in what we consider to be the standard procedures.

In a determination like the bichromate glycerine, we have a fairly large number of points at which measurement is made, the final result having an "error" which is the algebraic sum of all of the "errors" made. We have the standardizing of the bichromate, the weighing of the sample of glycerine, the absorption of water by the glycerine, the transference of the glycerine to the

(*Continued on page 220*)

FIG 2.



keeping properties of various shortenings and the crackers in which they are used. Evidence is presented to show that some anti-oxidants which are effective in prolonging the life of shortenings largely disappear during the mixing, fermentation and baking of crackers. A theory is proposed to explain the contradictory results on the relative stability of the various shortenings and crackers containing them.

Table 7

Stable	Fairly stable	Unstable	Very unstable
Saturated: stearic, palmitic, myrstic, etc.	Iso-oleic	Oleic	Linolic, linolenic

2. Metals, particularly iron, are shown to have a powerful effect in promoting the rancidity of crackers. For comparative stability work on crackers, it is necessary to remove all sources of metallic contamination.

3. Determination of the stability of shortenings does not necessarily indicate the stability the shortening will show in crackers. To determine the latter, it is necessary to use the shortenings in crackers and test the stability of the crackers.

Table 8

Shortening	Proportion of stable fatty acid glycerides	Proportion of unstable fatty acid glycerides	Stability in crackers
Oils	low	very high	very poor
Lard	medium	comparatively high	fair
Oleo oil	high	comparatively low	good
Regular all-hyd.	high	comparatively low	good
B. & C. all-hyd.	very high	almost none	excellent

LABORATORY RESULTS

(*Continued from page 211*)

flask, the making up to volume, the uniformity of solution by shaking, the variations in pipetting, the variations in the completeness of the oxidation, the error in back titrating as to (a) burette reading, (b) reading the end point.

Suppose in a determination like this, we have eight causes of analytical deviation, each equally as likely to be in one direction as another; suppose, for the purpose of discussion, the deviations are all of the same size and equal to 0.1; if we ran the determination 256 times we would have the deviations piling up as shown in Figure 2.*

In an actual system the variations are not of equal size, nor do they change so as to be either +0.1 or —0.1 with no steps between.

Therefore, the shape of the distribution smooths out and becomes somewhat like the dotted line in Figure 2. Standing alone, and portrayed more accurately, it would look like Figure 3.

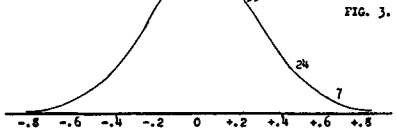


FIG. 3.

This bell-shaped curve, called the normal distribution curve, is smoothly symmetrical as shown, only when a great many results are plotted.

The results obtained when the value is ±0 8 from the average are not caused by any poorer work than those obtained when the value hits the average "on the nose." The large deviations are not the result of "error" in the sense of mistake, but are merely the accidental piling up of all of the small deviations in an all plus or all minus direction.

Methods of Expressing Variation

The degree of scattering in a set of results like that shown for the refining test can be expressed in a number of ways. The simplest method is to add up the deviation from the mean, regardless of sign, divide by the number of determinations and call the result the average deviation.

A more useful measure of the scatter of a set of results is obtained by squaring each deviation from the mean, adding the squares, dividing the sum by the number of deviations, (or the number minus one, for relatively few results) and extracting the square root of the quotient. This rather round-a-bout procedure gives us the root-mean-square or standard deviation. The term "standard" is misused somewhat; the standard deviation is not a standard for deviating, it is just a name for a mathematical value. It is usually called sigma (σ), and can be stated mathematically as

$$\sigma = \sqrt{\frac{\epsilon\, d^2}{n-1}}$$

where d is a deviation from the mean, and n is the number of results.

The utility of the standard deviation lies in the fact that it represents variation more "efficiently" than the average deviation. A standard deviation based on 100 results measures variation as dependably as an average deviation based on 114 results. For this reason, most statistical tables are computed with σ as the base.

In a set of results that are considered in making up a distribution curve, no results due to mistakes or blunders should occur. Any one result is, as far as known, obtained by as trustworthy laboratory manipulation as any other result. Only first-class work was considered in securing the results which were used to compute the standard deviations to be presented.

Assuming that we are dealing with an approximately normal distribution, we may use tables showing the relation between the standard deviation and the area under the normal curve. For example:

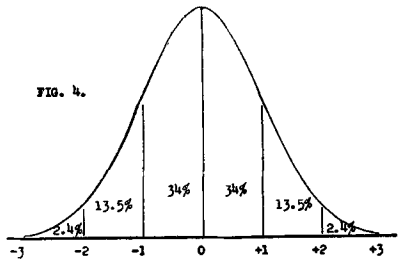


FIG. 4.

± Distance, in Sigmas, from the True Value	Area Under the Curve, Expressed in % of Results
1 σ	68 %
2 σ	95 %
3 σ	99.7%

Figure 4 shows these relationships.

If we have a number of results on the same sample, say Iodine Value, and we know that the standard deviation on Iodine Values is 0.14, we are justified in rejecting a value more than three times 0.14 units from the mean because the probability of a result falling more than three standard deviations away from the mean is less than 3 in a 1000.

What is a Good Check

The procedure of rejecting results more than 3 σ from the mean is useful only when we have enough results to establish the true mean value and the true standard deviation with some certainty. Ordinarily we get a single value —let us say 7.9% loss—that does not look just right; we send the sample back to the laboratory for a repeat and get another value—say 7.3%. This difference raises two questions in our minds; first, "Are these good checks?"; and second, "What's the true loss?" To answer the second and easier question first, the true value can only be determined by running a number of good analyses and taking the average. The first question, "Are two determinations —so and so apart—good checks?" requires a more involved explanation.

The result of the first test may be anywhere within ± 3 σ of the average, (neglecting the 3 results out of 1000 likely to fall beyond ± 3 σ), but we do not know what the average is. The result of the second test will likewise fall within ± 3 σ of the unknown average. How then can we set up a criterion for "good checks"? Suppose we look at a normal distribution curve (Figure 5); this curve can be assumed to represent the way the "errors" distribute in many analyses. Therefore any conclusions we may draw concerning the normal distribution curve can be applied to all analyses for which we have a standard deviation (σ).

The result from the first test, "A," can fall almost anywhere within the

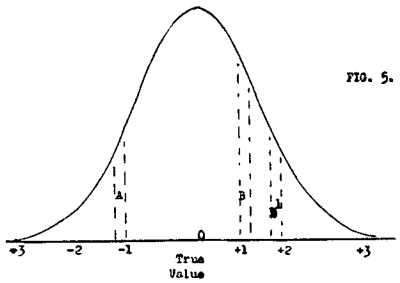


FIG. 5.

curve, the first result having fallen at "A" the second result can fall anywhere within the curve; at "A" again, or at another point "B"; or "B'," for examples. We can take the result "A," and tabulate the differences between it and all of the other results in a set of results possessing normal distribution. We can then take another result, and tabulate the differences between it and all of the other results in the distribution. By continuing this process for all of the results, that is, getting the differences between each point and all of the other points, we can build up a table of differences. This table will show the size of the differences and the frequency with which they occur. As a statistical mathematician says, "The differences in a normal distribution are themselves normally distributed with a standard deviation 1.41 times the original standard deviation." We can then say that:

It is a good rough rule to say that duplicate results which fail to check within 2σ should be looked into, for differences greater than 2σ occur, with good analytical practice, only about one out of seven times.

All of the above σ's except those for iodine value were determined on co-operative work. The standard deviation for a set of co-operative results is greater than that for a set of repeated determinations made in a single laboratory. The increase in deviation of co-operative results over single laboratory results depends upon the number of calibrated instruments involved in the determination (assuming uniform methods of analysis). With titers, for instance, the standard deviation for co-operative results is only slightly greater than that for single laboratory results, because only thermometers are involved and thermometer calibrations are fairly accurate. With iodine values, in which balances, burettes, temperature expansion of solutions, etc., are involved, the standard deviation of co-operative results will be about twice that of repeated results in a single laboratory.

Standard deviations for refining tests on crudes of higher fatty acid than those available this year will be compiled and presented as soon as co-operative work on such samples is carried out.

The value of σ for good work following N. C. P. A. and A. O. C. S. Methods is as follows on the various determinations:

Determination	σ	Reference
Glycerine		Compiled from accepted results A. O. C. S. Glycerine Analysis Com. for years 1931-2-3.
Acetin	0.4	
Bichromate	0.55**	
Iodine Value		
Range 40-70	0.14 (Single Lab.)	Barbour, A. D. "A Comparson of Various Methods of Running I. V.'s," "Oil & Soap," Jan., 1934, p. 9.
Refining		
C. S. O., 1—2 FFA.		Compiled from results of collaborators on N. C. P. A. samples for 1933-4.
F.F.A.	0.08	
Loss	0.22	
Color	0.25	
Bleach Test	0.15	
	Each σ from the mean for each lye.	

Total Fatty Acid in Soap Stock (Samples 1 and 2, 1933-34) .27 and .39.

The average difference is	1.13σ
The median difference is	$.95\sigma$
(This difference is the middle difference—there are as many larger than it, as there are smaller.)	
75% of the results are closer together than	1.62σ
85% " " " " " " "	2.02σ
90% " " " " " " "	2.32σ
95% " " " " " " "	2.76σ

*These values are based on probability mathematics as applied to "heads or tails" for 8 coins. The +0.1 or —0.1 value for each cause of variation is considered to occur just like heads or tails.

**Several non-standard methods were used by the committee labs.