STATISTICAL ANALYSIS OF STANDARD PLATE COUNTS OF MILK SAMPLES SPLIT WITH STATE LABORATORIES

C. B. Donnelly, E. K. Harris, L. A. Black
and K. H. Lewis

Robert A. Taft Sanitary Engineering Center,
Public Health Service,

Department of Health, Education, and Welfare,
Cincinnati, Ohio

(Received for publication May 25, 1960)

The split sample procedure is designed to check the performance of laboratories examining milk for interstate shipment by providing actual data from their comparative analyses of milk samples. The procedure generally requires that a sample of fluid milk be divided into portions which are shipped to participating laboratories for examination by agar plate and other methods. The results reported by these laboratories are inspected to determine if any laboratory reports unusually high or low counts. In addition, the counts may be checked to see if they agree within 10 or 20 percent (or some other arbitrary figure) of the counts reported by one or more reference laboratories. The belief was expressed at the 1959 National Conference on Interstate Milk Shipment, that such criteria may not be based on realistic limits of variation and that standards are needed to judge split sample performance (*4*). This paper attempts to develop such standards or limits, based on the statistical analysis of standard plate counts reported by central State laboratories in two split milk sample evaluations.

Materials and Methods

Table 1 shows that in the first evaluation five series of split milk samples were sent out between December 1957 and April 1958 and that the 42 participating states were divided into five groups of seven to nine states each. Each group examined one of the five series of samples. In the second evaluation (Table 1) only two series of samples were shipped out: the first in October, 1958, to a group of 22 states, the second series in November to a different group of 21 states.

The samples for both evaluations were essentially the same, consisting of raw, pasteurized, homogenized and chocolate milk, and of pasteurized cream. Each series included representative low, moderate, and high count samples, and two or more pairs of duplicates. In the first evaluation a set of eight split samples of 8-10 ml. was shipped to each State laboratory. In the second evaluation the set consisted of 10 samples of 20-30 ml. each. Each set was examined by one to four analysts in the State laboratory, and a set of each series was also examined by the senior author.

Table 1 — Plan for Evaluating Central State Milk Sanitation Laboratories by Split Sample Procedure

Evaluation	Group	Number of states	Number of analysts	Series of samples	Number of samples	
					Pairs of duplicates	Unmatched
Dec. 1957-	1st	7	17	1st	3	2
April 1958	2nd	9	14	2nd	3	2
	3rd	8	12	3rd	3	2
	4th	9	16	4th	3	2
	5th	9	17	5th	3	2
Oct.-	1st	22	51	1st	2	6
Nov. 1958	2nd	21	41	2nd	3	4

The preparation and shipment of the samples were essentially as described or suggested by Donnelly *et al.* (*2*) except that each series included certain samples inoculated with pure cultures to provide high plate counts. In most instances, samples were received and examined the day after shipment.

Each analyst was requested to prepare a 1:100 dilution of each sample, to plate duplicate 1.0 ml. (1:100) and 0.1 ml. (1:1000) aliquots from each dilution and to report the number of colonies per plate. From these counts the average number of colonies on duplicate plates and the standard plate count per ml. of milk were computed.

Analysis and Results

The results reported in each evaluation were analyzed primarily to estimate the average variation (*a*) among a large number of analysts examining the same samples of milk and (*b*) between duplicate counts from a single sample of milk prepared and read by the same analyst. This study of variation among analysts and "within" a single analyst was based on the standard plate counts (SPC's) reported

in both evaluations. Each SPC represented an average of one or two pairs of duplicate colony counts, depending on whether one or both dilutions yielded counts within approximately the 30 to 300 colony range. The logarithm of the SPC was used rather than the actual count since estimates of variation obtained under this transform were independent of fluctuating mean counts which ranged from approximately 5,000 to 150,000 per ml.

For the first evaluation, a separate analysis of variance was carried out within each group of states on the SPC's reported for each pair of duplicate samples, i.e., 15 analyses in all. The two unmatched samples were not included in these analyses as they yielded very low counts. In the second evaluation the standard plate counts for duplicate as well as for unmatched samples were analyzed, making a total of seven analyses.

In each of these analyses, the milk samples and the analysts were assumed to be random samples from their respective (infinite) populations. Variation in the log SPC was assumed attributable to three components: (*a*) differences in bacterial densities of apparently identical samples, (*b*) variation among analysts, and (*c*) residual variation, i.e., the average variation shown by a single analyst in replicate plate counts from the same milk sample. These variance components are set out in symbolic form in the following analysis of variance table (*1*).

In Table 2, p represents the number of analysts, q the number of replicate samples, σ_a^2 the variance component attributable to different analysts, σ_s^2 the component due to real differences between samples, and σ^2 the average variation among log SPC's reported by a single analyst from aliquots of a given milk sample. This table assumes that "interaction" between sample and analyst may be ignored. Such interaction would arise (and inflate all the expected mean squares in Table 2 by the same amount) if some analysts were to report high counts in one of a pair of samples and lower counts in the second sample while other analysts reported the reverse. A later footnote indicates that this assumption is probably justified.

TABLE 2 — GENERAL ANALYSIS OF VARIANCE ATTRIBUTABLE TO DIFFERENCES BETWEEN ANALYSTS AND SAMPLES

Source of variation	Degrees of freedom	Expected mean square (MS)
Analysts	$p - 1$	$\sigma^2 + q\,\sigma_a^2$
Samples	$q - 1$	$\sigma^2 + p\,\sigma_s^2$
Residual	$N - p - q + 1$	σ^2
Total	$N - 1$	

Estimates of the first two variance components are given by the formulae:

$$\hat{\sigma}_a^2 = \frac{\text{Observed MS analysts} - \text{Observed MS residual}}{q}$$

and

$$\hat{\sigma}_s^2 = \frac{\text{Observed MS samples} - \text{Observed MS residual.}}{p}$$

In general, in both evaluations the mean square between duplicate milk samples was not significantly greater than the residual mean square, indicating no real differences between such samples. However, the mean square among analysts was generally found to be much higher than the residual mean square, indictating significant variability among analysts. Individual estimates of this variance component, for each of the fifteen analyses in the first evaluation and the seven analyses in the second were quite heterogeneous. In the first evaluation, a substantial portion of this heterogeneity was traceable immediately to analysts from two particular laboratories whose discrepant results led to some unusually high later calculations. In the second evaluation, it was values of $\hat{\sigma}_a^2$. These results were omitted from all found necessary to omit data from only one of these laboratories.

The weighted average estimates of $\hat{\sigma}_a^2$ (weighting by the degrees of freedom, p - 1, in each analysis) obtained from the two evaluations were .0069 and .0076, respectively, in terms of log SPC.

Estimates of residual variance σ^2, also weighted averages, were .0028 for the first evaluation and .0160 for the second. The latter figure was too high, due undoubtedly to the presence of substantial interaction between sample and analyst (which would affect the estimate of σ^2 but not of σ_a^2) in those portions of the second evaluation involving four or six unmatched samples. When these results were omitted, the average estimate of residual dropped to .0058, still higher than the estimate from the first evaluation. Nevertheless, striking some averages, results of the entire study point to a variance component between analysts of about .007 (in terms of log SPC) and a residual variance (within sample and analyst) of .004-.005.[1] This study indicates, therefore, that overall

[1]This is about the value one would expect if "interaction" between sample and analyst were absent. Consider, for example, an average colony count of 100, a typical value. The variance of replicate plate counts about this average would approximate the same value, 100. The variance of mean counts based on pairs of such replicates would be 50. Since the variance of the log SPC based on such a mean colony count is equivalent to the variance of the logarithm of the mean colony count itself, and since the latter variance is approximately equal to the variance of the mean divided by the square of the "true" count we obtain,

Variance of log SPC $= \frac{50}{(100)^2} = .005.$

variation among analysts should not greatly exceed .007 + .005, or .012. We may presume that these values hold at least within the plate count range of 5,000 - 150,000 per ml. of milk.

These results may be put in terms of the percentage difference between a pair of plate counts reported by a single analyst, or between SPC's reported by two different analysts. When a larger number of analysts are surveyed, as would be the most common situation, percentage difference generalizes to the ratio of standard deviation to mean. Under the logarithmic transformation, however, the standard deviation or its square, the variance, of the SPC becomes largely independent of the mean, at least over the range of practical interest. Hence an observed variance of log SPC (where each count is based on two plates) may be checked directly against the criterion .012 suggested by the results of this study to determine the acceptability of the observed results.

Table 3 lists data from the second evaluation, and illustrates the method of computing the variance of the log SPC. The counts of Sample A showed a degree of agreement among analysts typical of the study as a whole. The variance of the log SPC computed for this sample, .012, was, in fact, identical to the average level suggested by this study. On the other hand, analysts examining Sample B reported plate counts whose logarithms showed a variance of .051, approximately four times the average level of .012. Of course, this high value represents an extreme among the many samples distributed in this survey, and naturally we cannot judge the significance of this particular result by the ordinary statistical criteria that we would apply to a sample selected at random. However, if variances of this size were encountered frequently (in say, more than ten percent of samples) in future evaluations, one would certainly suspect that at least some of the analysts were not performing satisfactorily or that certain samples were not being split uniformly.

The problem then arises of determining which analysts need improvement. In general, this question cannot be answered with confidence unless results of all analysts are available on a series of samples examined as a whole. It is conceivable, though unlikely, that counts on each separate sample may show a high variance and yet none of the analysts report consistently high or low counts over all samples. Usually, certain analysts will tend to count consistently higher or lower than their colleagues. The simplest way of determining this is to rank the counts by size within each sample and then study the distribution of ranks for each analyst over all samples. Table 4 presents these ranks for the 21 analysts of Sample B (Table 3) over the entire series of 10 samples examined (Sample B was No. 7).

We see immediately that certain analysts have consistently reported higher or lower counts than most of their colleagues. Analysts 2, 9, 11 and 14 clearly fall into the former category and Analyst 1 into the latter. More on the borderline are 13, 20 and 21, high, and 15 and 18, low. The column of total ranks helps to judge the consistency of each analyst.

Table 4 also discloses peculiar sets of ranks for two Analysts, 12 and 16. Analyst 12 ranked low on Samples 1, 2, 6 and 7 but much more typical or even high, on the remaining samples. Analyst 16, on the other hand, ranked very low on Samples 4, 5, 8, 9 and 10 but quite high on the other five samples. Reviewing the original colony counts, it was apparent that these surprising reversals were not due to under- or over-counting of crowded plates since both .01 and .001 dilutions of each sample showed closely consistent results. Some other defect appeared to have been

TABLE 3 — VARIATION AMONG ANALYSTS IN STATE LABORATORIES AS INDICATED BY THE VARIANCE OF LOGARITHMS OF STANDARD PLATE COUNTS

Sample A			Sample B		
Analyst	SPC/100	(Log SPC)-3	Analyst	SPC/100	(Log SPC)-3
A	130	1.114	1	100	1.000
B	130	1.114	2	320	1.505
C	130	1.114	3	270	1.431
D	130	1.114	4	290	1.462
E	120	1.079	5	200	1.301
F	110	1.041	6	140	1.146
G	130	1.114	7	240	1.380
H	130	1.114	8	200	1.301
I	120	1.079	9	240	1.380
J	93	0.968	10	240	1.380
K	130	1.114	11	320	1.505
L	120	1.079	12	150	1.176
M	120	1.079	13	210	1.322
N	110	1.041	14	270	1.431
O	110	1.041	15	200	1.301
P	40	0.602	16	260	1.415
Q	120	1.079	17	290	1.462
R	100	1.000	18	200	1.301
S	99	0.995	19	210	1.322
T	79	0.897	20	160	1.204
U	120	1.079	21	200	1.301

$$\text{Variance of log SPC} = \frac{\Sigma \log^2 - \frac{(\Sigma \log)^2}{21}}{20}$$

Sample A	*Sample B*
$\Sigma \log = 21.857$	$\Sigma \log = 28.026$
$(\Sigma \log)^2 = 477.72845$	$(\Sigma \log)^2 = 785.45668$
$\Sigma \log^2 = 22.981523$	$\Sigma \log^2 = 38.411906$
Variance of log SPC = 0.0116	Variance of log SPC = 0.0505

TABLE 4 — RANKING[1] OF COUNTS REPORTED ON EACH OF A SERIES OF TEN SPLIT SAMPLES (2ND GROUP, OCTOBER - NOVEMBER, 1958)

Analyst	Sample number[2]										Total Rank
	1	2	3	4	5	6	7	8	9	10	
1	5	11	1	2	4	3	1	2	1	2	32
2	15	18	17	18	21	17	21	16	15	19	177
3	10	15	4	12	7	7	12	20	10	9	106
4	17	16	10	10	15	9	19	17	3	14	130
5	14	6	15	17	8	8	6	10	7	6	97
6	4	8	18	6	5	2	2	7	11	10	73
7	7	7	13	7	9	11	14	6	19	11	104
8	8	10	14	8	11	12	8	19	21	20	131
9	20	20	21	20	14	18	15	15	20	17	180
10	9	19	8	4	12	4	13	4	9	4	86
11	21	21	20	16	20	21	20	18	12	21	190
12	1	2	7	14	10	1	3	13	17	7	75
13	16	14	11	19	17	16	10	21	8	13	145
14	12	13	16	21	18	20	17	9	18	18	163
15	13	4	3	11	3	5	7	5	4	5	60
16	18	17	12	1	2	19	16	1	2	3	91
17	11	5	6	15	13	13	18	11	14	12	118
18	3	1	9	3	1	10	5	12	6	1	51
19	6	3	5	9	6	15	11	8	13	8	84
20	2	12	2	5	16	6	4	3	5	15	70
21	19	9	19	13	18	14	9	14	16	16	147

[1]Rank 1 denotes the lowest count. Samples 1 and 6, 2 and 7, 4 and 8 were duplicates.

[2]Since standard plate counts were rounded to the nearest thousand, ties were frequently encountered. These could almost always be eliminated by reference to the original colony counts. In the rare cases where colony counts were also identical, the assignment of separate ranks was decided by tossing a coin.

responsible, particularly in the case of Analyst 16 whose counts of Samples 4, 5, 8, 9, and 10 were all far below average.

Ranking the reported counts within each of a series of samples, as in Table 4, clearly provides useful supplementary information on the relative accuracy and reliability of individual analysts. Statistical significance tests are available as guides in interpreting a table of ranks. For example, the variation in total ranks may be tested to determine whether some analysts consistently under- or over-counted samples. On the assumption that no such consistent differences existed, the expected variance of total ranks, say σ_R^2, is given essentially by the expression, $\sigma_R^2 = \frac{mn(n+1)}{12}$, where m is the number of samples observed, and n the number of analysts (3).

For $n > 7$, and assuming no ties, the ratio of the observed sum of squares of deviations of total ranks from their mean,

$$\text{i.e., } \sum_{i=1}^{n} (R_i - \bar{R})^2 = \sum_{i=1}^{n} R_i^2 - \frac{(\Sigma R_i)^2}{n} = \text{say, } S_R^2$$

to the expected variance given above may be tested by the χ^2 distribution with $n - 1$ degrees of freedom.

In the present example, Table 4, we calculate

$$\sum_{i=1}^{n} R_i^2 = (32)^2 + (106)^2 + \ldots + (70)^2 = 294{,}290 ,$$

$$\Sigma R_i = 2{,}310 , \text{ and}$$

$$S_R^2 = 294{,}290 - \frac{(2{,}310)^2}{21} = 40{,}190 .$$

Also, $\sigma_R^2 = \frac{(10)(21)(22)}{12} = 385$.

Hence, $\chi_R^2 = \frac{40{,}190}{385} = 104$, with 20 degrees of freedom. The probability of this high a value of χ_R^2 under the hypothesis of no consistent bias is extremely small, less than 0.5 percent. We have, therefore, statistical proof of what inspection of Table 4 clearly reveals — a strong, consistent bias in the counts of some analysts.

Crude statistical limits for isolating these analysts may be obtained by adding and subtracting twice the standard error, σ_R, from the average of the total ranks, i.e. $\frac{m(n+1)}{2} \pm 2\sigma_R$. All analysts whose total ranks lie outside these limits may be suspected of consistent bias, although under the hypothesis of no bias in any analyst, we would expect five percent of the total ranks to fall above or below these limits. In the present example, $\sigma_R = \sqrt{385} = 19.6$. Hence, the limits are 110 ± 39.2 or 70.8 and 149.2. Analysts 1, 2, 9, 11, 14, 15 and 18 fall well outside these

limits, while Analysts 13, 20 and 21 are on the borderline as indicated earlier by inspection of Table 4.

SUMMARY

Standard plate counts (SPC's) reported by 41 State laboratories on five to eight split milk samples were analyzed statistically to estimate the average variation among groups of analysts examining the same samples of milk, as well as between duplicate counts from a single sample obtained by the same analyst and, further, to develop criteria for deciding which analysts need to improve their performance. Analysis of variance of the log SPC showed that overall variation among analysts should not greatly exceed .012. Two samples were selected to illustrate the calculation of variance and, in one case, was found to be .051, or about four times the typical value of .012. Variances of this size in more than ten percent of the samples provide a valid basis for suspecting inadequate performance by the analysts or nonuniformity in splitting the samples in question.

Further information about the performance of individual analysts may be obtained by ranking the counts according to size within each sample and computing the total rank of each analyst for the entire series of samples. All analysts whose total ranks fell above or below twice the standard error of the average total rank for the group of analysts may be regarded (tentatively) as showing consistent bias which should be corrected. In the example shown, seven of twenty-one analysts fell in this category and three others were on the borderline.

REFERENCES

1. Bennett, C. A. and Franklin, N. L. *Statistical Analysis in Chemistry and the Chemical Industry*, pg. 377. John Wiley and Sons, New York, 1954.
2. Donnelly, C. B., Black, L. A., and Lewis, K. H. Containers, Refrigerants and Insulation for Split Milk Samples. J. Milk and Food Technol., **21**: 5, 131-137. 1958.
3. Kendall, M. G. *Rank Correlation Methods*, pg. 98. Hefner Publishing Company, New York. 1955.
4. Minutes of Seventh Meeting, National Conference on Interstate Milk Shipments. St. Louis, Missouri, April 1959.

NEWS AND EVENTS

QUESTIONS AND ANSWERS

Note: Questions of technical nature may be submitted to the Editorial Office of the Journal. A question in your mind may be in the minds of many others. Send in your questions and we will attempt to answer them.

QUESTION:

What can a small dairy plant with some laboratory facilities do to ensure that milk they ship in interstate commerce will be free of pesticide residues?

ANSWER:

The problems of pesticide residues in milk are much more complex than antibiotics. Your laboratory will not be of much use to you in checking your milk supply as the test procedures are complex and beyond the resources of most dairy laboratories. However, your laboratory or some other member of your organization can prepare information on the problem and see that each individual patron is fully informed on the sources of the residues and what can be done to keep them out of the milk. You should work closely with your local regulatory agency in preparing a list of approved pesticides and herbicides with proper directions for their use. Your field force should be alerted to see that your patrons follow the directions you have outlined. In short, an intensive campaign of education and supervision is the only answer to the pesticide problem for the small plant.

QUESTION:

We hear about rancid milk resulting from improperly installed pipe line milkers on the farm. Can improper plant practices also result in rancidity problems?

ANSWER:

Yes. The same factors can cause rancidity in the plant up to the time milk has been heated to 135°F. Air leaks should be avoided. Elevations of temperature should be avoided prior to heating for pasteurization. Addition of homogenized milk to raw milk must be avoided. The milk should not be homogenized until the temperature reaches at least 135°F.

QUESTION:

What is the Astell Roll tube method for bacteria counts?

ANSWER:

This is a procedure to measure the bacterial content of milk by adding the sample directly to a tube of melted agar and spinning it to form a film of hardened agar around the inside of the tube. Tubes are incubated and counts are made of colonies growing in the agar layer. This method is reputed to be less costly and time consuming than conventional plate counting. The initial investment is less than for a plate count procedure and should be adaptable to small plant laboratories. The method is to be included in the new 11th Edition of *Standard Methods for the Examination of Dairy Products*. Additional information can be obtained from the APV Company, Inc., 137 Arthur Street, Buffalo 7, New York.