

# The Nature of Repeatability and Reproducibility

JOHN MANDEL and THEODORE W. LASHOF

*National Bureau of Standards, Gaithersburg, MD 20899*

Repeatability and reproducibility are usually defined by the relation  $R = k\sqrt{2}s_r$ , where  $R$  stands for repeatability or reproducibility, and  $s_r$  is the corresponding standard deviation. It is then stated that the probability is  $C$  that a difference between two test results will lie between  $(-R)$  and  $(+R)$ . For  $C = 0.95$ , which is the usual choice, the values that have been proposed for the multiplier  $k$  are 1.96, 2, or Student's  $t$ . However,  $C$  is actually a random variable with a highly skewed distribution. It is shown that regardless of the above choice of  $k$ , the probability that  $C$  will lie in the "vicinity of 0.95," such as for example 0.92 to 0.97, is very small, unless the number of participating laboratories is large (30 or more). Nevertheless, for any given interval defining a "vicinity of 0.95," a value of  $k$  exists that maximizes the probability that  $C$  lies in that interval. For a number of reasonable choices for "vicinity of 0.95" the optimal  $k$  is close to 1.96.

## Introduction

THE concepts of repeatability and reproducibility, while opposed by some, have nevertheless been adopted by so many writers and workers in the field of interlaboratory study of test methods (ASTM [1978-83], BSI [1979], ISO [1979, 1981]) that the time has come for a clarification rather than a confrontation in this area.

The situation in which the need for these concepts arose is easily described. An interlaboratory study of a test method is generally run for the purpose of enabling a committee to formulate a "precision statement" about the test method. There are many ways in which such a statement can be drafted, but one widely used approach is to express it in terms of the "maximum difference" that can "reasonably" be expected between two test results obtained on the same material. The two test results are obtained in accordance with the test method (i.e., by the same measuring technique). If the test results are obtained in the same laboratory one refers to this "maximum difference" as *repeatability*; if they are obtained in different laboratories, the maximum difference in question is referred to as *reproducibility*.

To achieve clarity and avoid the ambiguity inherent in the word "reasonably," the following, more quantitative definitions are often given. *Repeatability*, denoted by  $r$ , is a quantity such that the probability is

95% that two test results obtained in the same laboratory (on the same material) *will not differ by more than  $r$* . A similar definition obtains for *reproducibility*, denoted by  $R$ , by changing the words "in the same laboratory" to "in different laboratories," and the letter  $r$  to  $R$ .

In the following section we examine under what conditions and assumptions such definitions make mathematical sense.

## Conditions of Applicability

First of all, the use of a probability value requires that we specify the statistical distribution of the experimental errors that cause fluctuations in the measurements. Whether explicitly stated or tacitly assumed, the distribution considered is almost always the Gaussian (normal) error distribution. In this paper we will not question this assumption, although the matter is worthy of a closer investigation.

Let us examine what is being required. Let  $\sigma$  represent the (true) standard deviation among different test results. Of course different values of  $\sigma$  will apply depending on how the test results are obtained (i.e., in the same laboratory or in different laboratories). We assume that these populations of test results are properly defined and stated.

The difference, say  $d$ , between two test results chosen at random from the stated population, will have an expected value of zero, if we consider  $d$  as retaining its algebraic sign (say, always, first laboratory chosen minus second laboratory chosen). The standard deviation of  $d$  is  $\sigma\sqrt{2}$ , because each test result contributes its own error of standard deviation  $\sigma$ .

---

John Mandel is a Senior Statistical Consultant at the National Measurement Laboratory. He is a Fellow of ASQC.

Theodore W. Lashof is a Research Associate.

We are interested in the probability that  $d$ , in absolute value, be equal to or less than  $R$ . (To be specific, we are assuming that we are considering *reproducibility*.) By definition of  $R$ , we have

$$\text{Prob}\{|d| \leq R\} = 0.95. \quad (1)$$

We wish to find the value of  $R$  that satisfies the above equation. This is equivalent to

$$\text{Prob}\{-R \leq d \leq R\} = 0.95. \quad (2)$$

Dividing by  $\sigma\sqrt{2}$ , we obtain

$$\text{Prob}\left\{-\frac{R}{\sigma\sqrt{2}} \leq \frac{d}{\sigma\sqrt{2}} \leq \frac{R}{\sigma\sqrt{2}}\right\} = 0.95. \quad (3)$$

But the ratio  $d/(\sigma\sqrt{2})$  is that of a random variable of zero expectation to its standard deviation. Assuming normality, this ratio is simply the "standard normal deviate" (i.e., a normal random variable of mean zero and standard deviation equal to unity). Denoting it by  $z$ , we obtain the requirement:

$$\text{Prob}\left\{-\frac{R}{\sigma\sqrt{2}} \leq z \leq \frac{R}{\sigma\sqrt{2}}\right\} = 0.95. \quad (4)$$

It is well known that the value of  $R/(\sigma\sqrt{2})$  that will satisfy this equation is given by

$$\frac{R}{\sigma\sqrt{2}} = 1.960$$

which gives

$$R = (1.960)(\sqrt{2})\sigma. \quad (5)$$

There is no doubt that this equation satisfies the conditions stated in the definition of  $R$ , provided that  $\sigma$  is the (true) standard deviation between test results obtained in different laboratories. To make this condition specific, let us denote this standard deviation by  $\sigma_R$ . Thus:

$$R = (1.960)(\sqrt{2})\sigma_R. \quad (5)$$

For test results obtained in the same laboratory a similar reasoning applies, and we obtain:

$$r = (1.960)(\sqrt{2})\sigma_r. \quad (6)$$

Unfortunately, these mathematical relations are applicable only when the values of  $\sigma_r$  and  $\sigma_R$  are known. But these standard deviations are population parameters, and are seldom, if ever, known exactly. Instead, we generally only have *estimates* of these quantities, derived from the data provided by the interlaboratory experiment (the "round robin"). It is customary to denote such estimates by the letter  $s$ . Thus, while  $\sigma$  denotes a *population* standard deviation,  $s$  denotes its *sample estimate*. The difficulty that con-

fronts us now is that if  $s$  is substituted in equations (5) and (6), these equations are no longer valid.

One method that has been proposed to solve this difficulty is to substitute Student's  $t$  for  $z$  (ASTM [1978, 1981, 1983], ISO [1979]). This proposal is based on the reasoning that if one substitutes  $s$  for  $\sigma$  in equation (3), the quantity  $d/(s\sqrt{2})$  thus obtained has a Student's  $t$  distribution. Hence, so goes the argument, it is merely necessary to replace the quantity 1.960 by the critical value of Student's  $t$  for a two-tailed critical region of probability  $1 - 0.95 = 0.05$ , and the number of degrees of freedom involved in the estimate of  $s$ .

For example, if  $s_R$  was estimated with 5 degrees of freedom (round robin with 6 participating laboratories), the appropriate value of  $t$  is 2.571; in this case we would define  $R$  by

$$R = (2.571)(\sqrt{2})s_R.$$

The proposal seems plausible, but a closer analysis of the situation, provided in the next section, shows that it is based on erroneous reasoning.

### The Correct Interpretation of the Probability Statement

Consider the equation

$$\text{Prob}\left\{-\frac{R}{s\sqrt{2}} \leq \frac{d}{s\sqrt{2}} \leq \frac{R}{s\sqrt{2}}\right\} = 0.95. \quad (7)$$

It is correct to write

$$\text{Prob}\left\{-t_c \leq \frac{d}{s\sqrt{2}} \leq t_c\right\} = 0.95 \quad (8)$$

where  $t_c$  is the critical value of Student's  $t$  at the 0.05 level of significance for the degrees of freedom with which  $s$  was estimated. But what is the meaning of equation (8)?

The quantity  $d/(s\sqrt{2})$  is a random variable derived from two other random variables,  $s$  and  $d$ . The former,  $s$ , is obtained from an interlaboratory experiment; the latter,  $d$ , is the difference between two laboratories chosen independently of the interlaboratory experiment. Therefore,  $s$  and  $d$  are statistically independent and in this respect equation (8) is valid. However, the very use of a probability in this equation implies, conceptually, a sequence of trials each of which would yield a new pair of values for  $s$  and  $d$ . Thus, equation (8) has the following meaning. Each conceptual trial is defined as a twofold experiment: 1) conduct an interlaboratory experiment to obtain a value for  $s$ ; 2) choose two laboratories at random (laboratories *not* necessarily included in the interlaboratory experiment), and determine the difference  $d$ .

If such a series of trials were actually carried out, then the set of values

$$d_1/(s_1\sqrt{2}), d_2/(s_2\sqrt{2}), d_3/(s_3\sqrt{2}), \text{ etc.}$$

would be such that, in the long run, 95% of them would fall between the fixed values  $-t_c$  and  $t_c$ . Obviously, it is totally impractical to run a new, complete round robin every time two test results are compared. Therefore, equation (8) is not applicable to real world situations, and consequently the often used equation, derived from a comparison of equations (7) and (8):

$$R = t_c\sqrt{2}s \quad (9)$$

does not achieve the desired probability of 95%. What can be done?

### Calculation of the Correct Probabilities

In real life, a single round robin is conducted; a precision statement, using  $s$ , is derived from it; and a possible long series of future comparisons ( $d$ -values) is made in terms of this precision statement. Let us define  $R$  by the equation

$$R = k\sqrt{2}s \quad (10)$$

where  $k$  is a constant to be chosen. The estimate  $s$  is obtained from the single round robin and is therefore temporarily fixed (until a new round robin is run). For the sequence of subsequent  $d$  values, we use the following rule:

if  $|d| \leq R$ , the two test results are compatible;

if  $|d| > R$ , the two test results are not compatible.

Then what is the probability of finding compatibility? It is generally *not* 95%, when  $k$  is equated to Student's  $t$ , and the value of  $s$  obtained from a single round robin is used repeatedly for a series of  $d$  values. This can be seen from the following considerations. If by chance  $s$  derived from the round robin was considerably smaller than  $\sigma$  (which can well happen for relatively small numbers of degrees of freedom), the probability of finding compatible pairs of test results, using the rule above and Student's  $t$  in the calculation of  $R$ , will tend to be much less than 95%. Conversely, a fortuitous large value of  $s$  (much larger than  $\sigma$ ) will tend to give us a much larger number of compatible pairs of test results. It is only *in the long run*, over a long series of round robins, *each* followed by a long series of  $d$  values, that the probability will average out to 95%. If we decide on any given value of  $k$ , then for each value of  $s$ , there is a fixed probability,  $C_s$ , conditional on  $s$ , that any subsequent  $d$  value will satisfy the relation  $|d| < R$ :

$$C_s = \text{Prob}\{-R \leq d \leq R|s\} \quad (11)$$

where  $R$  is defined by

$$R = k\sqrt{2}s.$$

Let us call this value of  $C_s$  the *coverage* corresponding to that value of  $s$ . Thus, the coverage is the probability for a round robin that resulted in  $s$  that a pair of test results will be called compatible.

Now,  $s$  itself is a random variable with a well defined probability distribution. The same is true for  $C_s$ , a function of  $s$ . Taking this into account, we can calculate (see Appendix) the relative frequency with which any given coverage will occur in the long run. From equations (1) and (11), it is seen that the intent of the definition of reproducibility is to make  $C_s$  exactly equal to 0.95; but as we have seen, this is impossible except in the average sense. Instead, we can define as "95%" an *interval around 95%* (the interval, for example, between 90% and 98%) and determine the relative frequency of round robins (i.e., number of values of  $s$ ) for which  $C_s$  will fall within the interval. Let us denote the lower limit of this interval by  $L$  and the upper limit by  $U$ . We are interested in the frequency with which  $C_s$  will lie between  $L$  and  $U$ , and we wish to make this frequency as large as possible.

### Maximization of the Probability of a 95% Coverage

The coverage  $C_s$  depends through  $R$  on both  $s$  and  $k$ . The factor  $k$  may be chosen so as to maximize the probability of  $C_s$  falling within the chosen interval that defines "95%." It is found (see Appendix) that the required value of  $k_{\max}$  is given by the following equation:

$$k_{\max} = \left[ \frac{B^2 - A^2}{\ln B^2 - \ln A^2} \right]^{\frac{1}{2}} \quad (12)$$

where  $A$  and  $B$  are the values of the standard normal deviate corresponding to the cumulative probabilities  $(L + 1)/2$  and  $(U + 1)/2$ , respectively.

Table 1 lists values of  $k_{\max}$  for various chosen intervals. Examination of this table shows that the value  $k = 1.960$ , used by many ASTM committees, is much preferable to  $k = \text{Student's } t$ . Table 2 uses this value,  $k = 1.960$ , and shows several intervals that correspond to it, and that can be used as definitions of "95%."

A remarkable mathematical conclusion relating to equation (12) is that  $k_{\max}$  is independent of the number of laboratories in the round robin (see Appendix for proof). However, the percentage of round robins that lead to  $C_s$  values falling within the chosen interval *does* depend on the number of laboratories, and, as

TABLE 1. Selected "95%" Intervals  
Derived Values of  $k_{\max}$

$L$	$U$	$k_{\max}$
80	98	1.778
85	98	1.865
90	98	1.976
90	97	1.901
92	97	1.957
92	96	1.900
93	96	1.932

shown in Table 3, increases as the number of laboratories increases.

### An Example

Table 4 is a data set taken from Johnson (1978) and represents measurements of phosphorus pentoxide ( $P_2O_5$ ) in fertilizers. Table 5 lists the averages, the standard deviations, and the 95% reproducibility limits at all levels. (Only single test results were obtained at each level by each laboratory.) For this example the value  $k = 1.96$  was used.

If we wish to interpret the reproducibility values in the light of our findings, we would proceed as follows:

1. We replace "95 percent" by the interval "90 to 98 percent"; thus  $L = 0.90$  and  $U = 0.98$ .

2. We calculate  $A$  and  $B$ : they are the values of the standard normal deviate for which the cumulative probabilities are  $(L + 1)/2$  and  $(U + 1)/2$ , respectively. Since  $(L + 1)/2 = 0.95$  and  $(U + 1)/2 = 0.99$ , we find in the appropriate table (Pearson and Hartley [1972]):

$$A = 1.6449 \quad \text{and} \quad B = 2.3263.$$

3. Applying equation (12), we obtain:

$$k_{\max} = \left[ \frac{B^2 - A^2}{\ln B^2 - \ln A^2} \right]^{1/2} = 1.9757$$

which is not far from  $k = 1.96$ .

4. Using the equations

$$G_A = \nu A^2 / k^2 \quad \text{and} \quad G_B = \nu B^2 / k^2$$

where  $\nu$ , for our data, is  $8 - 1 = 7$ , we find:  $G_A = 4.8516$  and  $G_B = 9.7047$ .

5. Equation (8) of the Appendix provides the integral that gives us the probability that the coverage is between 0.90 and 0.98. The result of the integration is

$$\text{Prob}[L \leq C_s \leq U] = 0.4721.$$

If 1.96 had been taken as the multiplier, the probability would be 0.4719. The practical meaning of this

result is that even if we use a liberal interval of 0.90 to 0.98 to replace the strict 0.95 specified in the usual definition of reproducibility, we still have a less than 50% chance that the coverage  $C_s$ , will be in the interval. This, however, is the best that can be achieved with eight participating laboratories.

This example, while apparently discouraging, should not lead one to conclude that interlaboratory studies of test methods are uninformative and perhaps unnecessary. While the usual definitions of repeatability and reproducibility are seen to be untenable, the interlaboratory study nevertheless provides essential and valuable information about the relation of within to between-laboratory variability, between both of these and the level of the measurement, and about the order of magnitude of these precision parameters.

### Summary and Conclusion

We have shown that the usual definitions of repeatability ( $r$ ) and reproducibility ( $R$ ) cannot be implemented because they are based on standard deviations that are not known exactly but can only be *estimated* from a round robin. Substitution of the critical value of Student's  $t$  for the usual multiplier 1.960 is also incorrect, because it is based on a probability model that does not fit the situation prevailing in real life. The realistic situation is one in which a single round robin is used for the estimation of the within- and between-laboratory standard deviations, and these estimates are subsequently used for many comparisons of test results.

To obtain an optimum value for the multiplier  $k$  used in the definitions of  $r$  and  $R$ , namely,

$$r = k\sqrt{2}s_r \quad \text{and} \quad R = k\sqrt{2}s_R$$

we first replace the 95% probability requirement in the definitions of  $r$  and  $R$  by an interval  $[L, U]$ , such as, for example, [0.92 to 0.97]. The definitions of  $r$  and  $R$  would now require that the probability that two test results agree to within  $r$ , or  $R$ , be at least  $L$  and at most  $U$  (rather than exactly 0.95).

TABLE 2.  $k_{\max} = 1.960$  Given  $L$ , Derived Upper Limit  $U$

$k_{\max}$	$L$	$U$
1.960	80	99.37
1.960	85	98.87
1.960	88	98.34
1.960	90	97.82
1.960	91	97.47
1.960	92	97.05
1.960	93	96.53

TABLE 3. Maximum Percentages of Round Robins That Will Satisfy Different Definitions of "About 95%",  $k_{max} = 1.960$

No. Labs	80-99.37	85-98.87	88-98.34	90-97.82	91-97.47	92-97.05	93-96.53
3	50.84	39.46	30.85	24.01	20.14	15.91	11.23
4	61.26	48.39	38.21	29.91	25.17	19.93	14.09
5	68.81	55.27	44.06	34.70	29.27	23.23	16.46
6	74.56	60.83	48.94	38.76	32.79	26.08	18.52
8	82.65	69.37	56.80	45.50	38.68	30.91	22.03
10	87.92	75.64	62.96	50.97	43.55	34.97	25.02
12	91.47	80.41	67.95	55.59	47.73	38.49	27.64
15	94.85	85.67	73.92	61.35	53.06	43.07	31.11
20	97.71	91.28	81.12	68.85	60.21	49.40	36.01
30	99.52	96.60	89.65	79.02	70.52	59.02	43.79
60	99.99	99.76	98.02	92.72	86.60	76.16	59.31
100	100.00	99.99	99.75	98.00	94.80	87.37	71.78
500	100.00	100.00	100.00	100.00	100.00	99.94	98.44

The fulfillment of this requirement will itself depend on the values of  $s$  obtained in the round robin. We wish to maximize the number of instances in which the obtained values of  $s$  allow fulfillment of this requirement.

We show that for any given interval  $[L, U]$ , a single multiplier,  $k_{max}$ , exists, which achieves this maximization requirement;  $k_{max}$  is a simple function of  $L$  and  $U$ . It turns out, surprisingly, that for many reasonable choices for  $L$  and  $U$ ,  $k_{max}$  is fairly close to 1.960, the value corresponding to the case where  $\sigma$  is known; it is certainly closer to 1.960 than to Student's  $t$  for the proper degrees of freedom.

**Appendix: Theoretical Development**

**Probability Distribution of the Coverage**

We consider an estimate  $s$  of a standard deviation  $\sigma$ , and assume that  $s$  has been estimated with  $\nu$  degrees of freedom. (For reproducibility in a round robin involving  $N$  laboratories,  $\nu = N - 1$ , when within-laboratory repeatability is numerically small compared with reproducibility.)

The probability of concern to us is

$$C = \text{Prob}\{-R \leq d \leq R\} \tag{A.1}$$

where  $d$ , the difference between the two test results, is a random variable of zero mean and standard deviation equal to  $\sigma\sqrt{2}$ . The letter  $C$  stands for "coverage."

Equation (A.1) when written in the form

$$C = \text{Prob}\left\{-\frac{R}{\sigma\sqrt{2}} \leq \frac{d}{\sigma\sqrt{2}} \leq \frac{R}{\sigma\sqrt{2}}\right\}$$

is therefore equivalent to

$$C = \text{Prob}\left\{-\frac{R}{\sigma\sqrt{2}} \leq z \leq \frac{R}{\sigma\sqrt{2}}\right\}$$

where  $z$  is a standard normal deviate.

By definition we have

$$R = k\sqrt{2}s$$

where the multiplier  $k$  is, at this point, unspecified. If we replace  $R$  by  $k\sqrt{2}s$ , the probability becomes conditional on  $s$ , and we obtain

TABLE 4. Interlaboratory Results for P<sub>2</sub>O<sub>5</sub> in Fertilizers

Lab	Material									
	A	B	C	D	E	F	G	H	I	J
1	7.70	8.68	12.65	13.60	18.70	20.20	30.20	31.40	45.88	46.75
2	7.63	8.64	12.73	14.16	18.95	19.92	30.09	30.42	45.48	47.14
3	8.04	8.45	13.17	13.71	19.52	20.91	29.10	30.18	45.51	48.00
4	7.74	8.66	12.98	13.68	19.00	20.65	29.85	31.34	44.82	46.37
5	7.83	8.73	12.88	13.66	19.08	19.94	30.29	31.11	44.63	46.63
6	7.70	8.59	12.60	13.08	18.85	20.30	29.88	31.00	45.13	46.75
7	7.69	8.54	12.25	12.75	18.83	19.43	29.80	29.50	43.50	44.91
8	7.85	8.75	12.99	13.26	19.20	19.97	29.40	30.25	45.18	46.78

TABLE 5. P<sub>2</sub>O<sub>5</sub> in Fertilizer Precision Parameters

Material	Average	S <sub>s</sub> (σ)	Reproducibility limit
A	7.772	0.131	0.363
B	8.630	0.100	0.277
C	12.781	0.288	0.798
D	13.488	0.438	1.214
E	18.954	0.351	0.973
F	20.165	0.463	1.283
G	29.826	0.403	1.117
H	30.650	0.669	1.854
I	45.016	0.730	2.023
J	46.666	0.862	2.389

$$C_s = \text{Prob}\left\{-k\frac{s}{\sigma} \leq z \leq k\frac{s}{\sigma} \mid s\right\} \quad (\text{A.2})$$

where we have replaced the symbol *C* by *C<sub>s</sub>*, to indicate the conditional nature of the probability and its dependence on the value *s*. We will call *C<sub>s</sub>* the “coverage corresponding to *s*.” Figure 1 shows that, for any fixed value of *k*, *C<sub>s</sub>* is a random variable between 0 and 1, and that it is a monotonic function of *s*: for any given value of *s*, say *s\**, there corresponds a value of *C<sub>s</sub>*, say *C<sub>s\*</sub>*. Thus, the distribution of the random variable *C<sub>s</sub>* is determined by that of *s*, which is related to the chi-square distribution, by

$$\frac{vs^2}{\sigma^2} = \chi_v^2$$

where  $\chi_v^2$  is the central chi-square distribution with *v* degrees of freedom. We now have (see Figure 1)

$$\begin{aligned} \text{Prob}\{C_s \leq C_{s^*}\} &= \text{Prob}\{s \leq s^*\} = \text{Prob}\{s^2 \leq s^{*2}\} \\ &= \text{Prob}\left\{\frac{vs^2}{\sigma^2} \leq \frac{vs^{*2}}{\sigma^2}\right\} = \text{Prob}\left\{\chi_v^2 \leq \frac{vs^{*2}}{\sigma^2}\right\}. \end{aligned}$$

Hence

$$\text{Prob}\{C_s \leq C_{s^*}\} = \text{Prob}\left\{\chi_v^2 \leq \frac{vs^{*2}}{\sigma^2}\right\}. \quad (\text{A.3})$$

Referring again to Figure 1, let *B* represent the abscissa corresponding to the upper limit of the interval defining *C<sub>s</sub>*, and let *s\** be the corresponding value of *s*. Then  $B = ks^*/\sigma$  and

$$\frac{vs^{*2}}{\sigma^2} = v\frac{B^2}{k^2} \quad (\text{A.4})$$

and the right side of Equation (A.3) becomes

$$\text{Prob}\left\{\chi_v^2 \leq v\frac{B^2}{k^2}\right\}.$$

For the chi-square distribution

$$\text{Prob}\{\chi_v^2 \leq G\} = \int_0^G \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2} dx. \quad (\text{A.5})$$

For *G* we substitute the value given by equation (A.4)

$$G = vB^2/k^2. \quad (\text{A.6})$$

To calculate the integral in equation (A.5), we need a value for *B*, which we obtain as follows.

From Figure 1 we see that the *total* area under the curve, to the left of *B*, is

$$P_B = C_s + \frac{1 - C_s}{2} = \frac{C_s + 1}{2}. \quad (\text{A.7})$$

*P<sub>B</sub>* is the cumulative probability corresponding to the value *B* of a standard normal variate. Given *P<sub>B</sub>*, we obtain *B* from a table of the normal distribution. For example, for *P<sub>B</sub>* = 0.8, we find *B* = 0.84162 (Pearson and Hartley [1972]).

Thus given a coverage-value *C<sub>s\*</sub>*, we calculate *P<sub>B</sub>* from equation (A.7), then find *B* from a table of the normal distribution. We then calculate *G* from equation (A.6), and the probability integral given by equation (A.5). By virtue of equation (A.3), this represents the probability that the coverage will be less than the value *C<sub>s\*</sub>*.

The intent of the definition of reproducibility is to make *C* exactly equal to 0.95, but this is impossible except in an average sense, since *C* depends on *s*, a random variable changing from one round robin to another (even for the same test method). Instead, we can define an *interval around 95%* (e.g., the interval between 90 and 98%) and require that “for as many

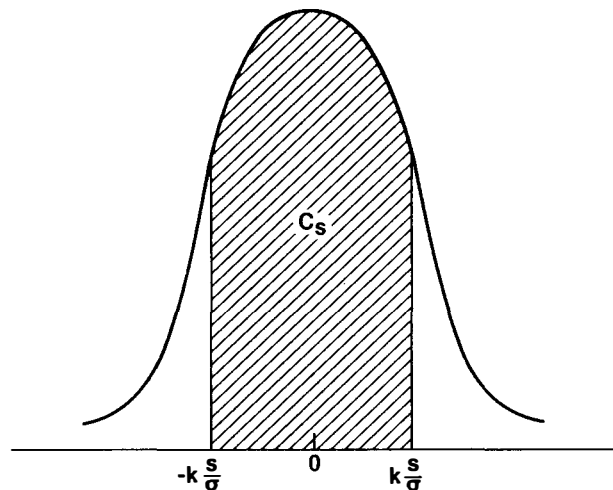


FIGURE 1. Probability Distribution of  $d/(\sigma\sqrt{2})$ .

round robins as possible" (i.e., for as many  $s$  values as possible)  $C_s$  will indeed fall in this interval.

Equation (A.5) gives us the probability that  $C_s$  will be less than the chosen value  $C_s$ . Applying this equation to a lower and an upper bound,  $C_s = L$  and  $C_s = U$  (e.g.,  $L = 0.90$  to  $U = 0.98$ ), we obtain from equation (A.4)

$$\text{Prob}\{L \leq C_s \leq U\} = \int_{G_A}^{G_B} \frac{1}{2^{v/2} \Gamma(v/2)} x^{v/2-1} e^{-x/2} dx \quad (\text{A.8})$$

where

$$G_A = vA^2/k^2 \quad \text{and} \quad G_B = vB^2/k^2$$

and  $A$  and  $B$  are the abscissa values of the normal distribution corresponding to the cumulative probabilities  $(L + 1)/2$  and  $(U + 1)/2$ ,  $L$  being the lower limit and  $U$  being the upper limit of the interval that defines "95%."

**Maximization of the Probability of Coverage**

The value of  $k$  has so far not been specified. We now seek to specify it in such a way that the probability expressed by equation (A.8) be as large as possible. This is reasonable since we wish the coverage to lie in the preassigned interval,  $L$  to  $U$ , as often as possible.

To obtain this value of  $k$ , which we denote by  $k_{\max}$ , we differentiate equation (A.8) with respect to  $k$  and set the derivative equal to zero. The parameter  $k$  appears in the two limits of integration,  $G_A$  and  $G_B$ , or  $vA^2/k^2$  and  $vB^2/k^2$ . Denoting the integrand by  $F(x)dx$ , and the integral by  $I$ , we have

$$\begin{aligned} \frac{dI}{dk} &= \frac{d \int_{G_A}^{G_B} F(x) dx}{dk} = -\frac{\partial I}{\partial G_A} \frac{dG_A}{dk} + \frac{\partial I}{\partial G_B} \frac{dG_B}{dk} \\ &= -F(G_A) \frac{dG_A}{dk} + F(G_B) \frac{dG_B}{dk} \\ &= -F(vA^2/k^2) vA^2(-2)k^{-3} + F(vB^2/k^2) vB^2(-2)k^{-3}. \end{aligned}$$

Setting the derivative equal to zero, we obtain (since  $k$  cannot equal infinity):

$$B^2 F(vB^2/k^2) - A^2 F(vA^2/k^2) = 0. \quad (\text{A.9})$$

The constant multiplier,  $(2^{v/2} \Gamma(v/2))^{-1}$ , in the function  $F$  can be omitted in this equation, which becomes

$$(vB^2/k^2)^{v/2-1} e^{-vB^2/2k^2} B^2 = (vA^2/k^2)^{v/2-1} e^{-vA^2/2k^2} A^2.$$

After algebraic simplification this equation becomes

$$(B^2/A^2)^{v/2} = e^{(B^2-A^2)v/2k^2}$$

or:

$$\frac{v}{2} \ln \frac{B^2}{A^2} = \frac{v}{2k^2} (B^2 - A^2)$$

and finally

$$k^2 = \frac{B^2 - A^2}{\ln B^2 - \ln A^2}. \quad (\text{A.10})$$

This then gives us the value of  $k_{\max}$  as

$$k_{\max} = \left[ \frac{B^2 - A^2}{\ln B^2 - \ln A^2} \right]^{1/2}. \quad (\text{A.11})$$

That this value is indeed a maximum, and not a minimum or a stationary value, can be shown.

A remarkable conclusion from equation (A.11) is that  $k_{\max}$  is independent of  $v$ , and therefore of the sample size (number of laboratories)  $N$ . However, the maximum itself (i.e., the value of the integral  $I$  given by equation (A.8)) will depend on  $v$ . Since  $k$  is now known, this integral can be calculated for every value of  $v$ .

**Relation to Tolerance Intervals**

Tolerance intervals are intervals dealing with coverages. They are expressed in the scale of measurement, say  $y$ , and can be classified into two groups (Proschan [1969]):

a) Intervals for which the *expected value* of  $C$  is a specified number, where  $C$  is the proportion of  $y$  values in the specified interval; for example, an interval  $(y_1, y_2)$  constructed in such a way that  $C$  is, on the average, equal to 0.90.

b) Intervals for which  $C$  is at least equal to a specified number, say  $C_0$ , with a specified probability  $P$ ; for example, an interval  $(y_1, y_2)$  such that

$$C \geq 0.95$$

with probability  $P = 0.80$ .

It can be shown that with a multiplier  $k$  equal to Student's  $t$ ,  $r$  and  $R$  become tolerance intervals of the type in (a) above. The reason for not using them in this sense is that, while  $C$ , *on the average* (that is, over a very large number of round robins), will indeed be equal to 0.95, the distribution of  $C$  is highly skewed, so that a very large proportion of round robins will result in a  $C$  value well above 0.95, compensated by a small proportion for which  $C$  will be far less than 0.95.

Tolerance intervals of the type in (b) are also inappropriate for our objectives. In repeatability and reproducibility intervals, we specify *an upper and a lower bound* for  $C$ , say  $C_1$  and  $C_2$ , and we construct the interval  $(y_1, y_2)$  in such a way that  $P$  is *maximum*, whereas in type (b) tolerance intervals, we specify a lower bound  $C_0$  for  $C$ , and require  $P$  to be a specified number. Had we tried to specify  $P$  at, say, 90% in our problem, we would have required a very large number

of laboratories in each round robin to satisfy our requirements. Realistically, it simply is not feasible to meet such requirements. Furthermore, even if it were possible to fulfill these requirements, they would relate to a coverage *at least equal to*  $C_0$ , whereas we wish to achieve a coverage *close to* a value such as 0.95, say 0.92 to 0.98 ( $C_1$  to  $C_2$ ).

In conclusion, the rationale underlying the conventional tolerance intervals is not applicable to our problem. It was therefore necessary to reformulate it completely. The reformulation presented in this paper is one that addresses the problem of repeatability and reproducibility in realistic terms.

### Acknowledgment

The authors are indebted to Professor Ingram Olkin for very valuable suggestions relating to the derivation of the frequency distribution of the 'coverages' as defined in this paper.

### References

- ASTM (1978). "Standard Recommended Practice E 180-78 for Developing Precision Data on ASTM Methods for Analysis and Testing of Industrial Chemicals." American Society for Testing and Materials, Philadelphia, PA.
- ASTM (1979). "Standard Practice E 691-79 for Conducting an Interlaboratory Test Program to Determine the Precision of Test Methods." ASTM, Philadelphia, PA.
- ASTM (1980). "Standard Practice E 173-80 for Conducting Interlaboratory Studies of Methods of Chemical Analysis of Metals." ASTM, Philadelphia, PA.
- ASTM (1981). "Standard Practice F 465-76(81) for Developing Precision and Accuracy Data on ASTM Methods for the Analysis of Meat and Meat Products." ASTM, Philadelphia, PA.
- ASTM (1983). "Standard Practice D 3980-83 for Interlaboratory Testing of Paint and Related Materials." ASTM, Philadelphia, PA.
- BSI (1979). "Precision of Test Methods. Part 1: Guide for the Determination of Repeatability and Reproducibility for a Standard Test Method." British Standard 5497:1979, London, England.
- ISO (1979). "Petroleum Products—Determination and Application of Precision Data in Relation to Methods of Test." International Standard 4259-1979, Geneva, Switzerland.
- ISO (1981). "Precision of Test Methods—Determination of Repeatability and Reproducibility by Interlaboratory Tests." International Standard 5725-1981, Geneva, Switzerland.
- JOHNSON, F. J. (1978). "Automated Determination of Phosphorus in Fertilizers; Collaborative Study," *Journal of the Association of Official Analytical Chemists* 61, pp. 533-536.
- PEARSON, E. S. and HARTLEY, H. O. (1972). *Biometrika Tables for Statisticians* 2, Table 1, Cambridge University Press, England.
- PROSCHAN, F. (1969). "Confidence and Tolerance Intervals for the Normal Distribution." *Precision Measurement and Calibration. Statistical Concepts and Procedures*. Special Publication 300, Vol. 1, National Bureau of Standards, Department of Commerce, Washington, D.C., pp. 373-387.

Key Words: *Interlaboratory Testing, Repeatability, Reproducibility, Round Robin Tests, Tolerance Intervals.*