TEST OF SIGNIFICANCE IN A VERIFICATION PROGRAM

By Irving I. Gringorten

Air Force Cambridge Research Center

(Original manuscript received 2 January 1954; revised manuscript received 30 August 1954)

ABSTRACT

The article is limited to consideration of a forecasting program in which a forecaster chooses as his forecast one event out of a finite group of well-defined, mutually exclusive events, for example, "rain" or "no rain." If a table of scores is prepared, from which a forecaster's score can be obtained for each of his forecasts, the sum total of a set of scores will become normally distributed as the number of forecasts is made large. It is possible to set confidence limits on the score total or on the difference of two score totals. But it is necessary to assume that the forecasts by which a forecaster's performance is judged are independent of each other and that the days constitute a fair sample of the climatology of his station.

1. Introduction

The subject "test of level of significance" is not a new one, for it can be found in any recent text on statistical methods (*e.g.*, Mood, 1950). The verifying and the scoring of a set of forecasts, likewise, is not a new subject or a dead issue (Crossley, 1954; Leight, 1953; Vernon, 1953). But there seems to be a dearth of published material on the significance of results in a verification program. How should we determine, *with confidence*, that the forecasters have exhibited significant skills in a series of forecasts? How can we decide conclusively that one system of forecasting is better than another? In addition to these problems, there is the problem of setting forth a procedure to sample a forecast method just often enough to rate the technique as successful or inferior, without necessarily demanding a long-term program for the trial testing of the technique.

In this article, the familiar 95 per cent confidence limit is discussed with respect to any program of verification in which a table of scores is used, and in which a forecaster chooses for each forecast one event out of a classification of well-defined, mutually exclusive events. This article, for example, will apply to the penalty scores of Vernon (1953), but will not apply to Leight's (1953) system of scoring probability statements.

The greatest obstacle in this investigation has been the fact that the distribution of a set of scores is dependent upon the performance or skill of the forecaster, which is the very thing that has to be determined by the verification program. Instead of presenting the 95 per cent confidence limit itself, therefore, the writer will present formulas for values that the 95 per cent confidence limit does not exceed. Even so, several assumptions must be valid for a successful test of the significance of results of verification:

1. If a forecaster predicts an event, there should exist a definite probability that his prediction will be correct, and also a definite probability that each of the alternative possible events will occur; because we shall attempt to find only boundary values for the 95 per cent confidence limit, this assumption is not serious;
2. The forecasts should be independent of each other; that is, the conditions for two consecutive forecasts should not be the same condition extended over two days; to avoid this disturbing factor, the forecasts might be spaced several days apart, so that the forecaster would be faced, on each day, with antecedent conditions that essentially are not related to the conditions of the previous day on which he forecast; this problem, however, may be difficult to avoid completely, because there is a persistence of weather patterns, even from month to month; it is better to live with this difficulty and assume that it will be negligible;
3. When a forecaster's performance is judged by his forecasts on a set of days, it must be assumed that the weather conditions on those days constitute a fair sample of the climatology of the station;
4. A forecaster's score total for one sub-set of his forecasts should be independent of his scores for another sub-set; if he improves his total in one sub-set of his forecasts, it does not necessarily follow that his score total will improve or suffer in another sub-set.

Allowing the above assumptions as tenable, we are able to show that the score total for a large number of forecasts is asymptotically normally distributed. This means that the expected value and the standard deviation of the total of scores define, for practical purposes, the random variations of the total. Since the 95 per cent confidence limit of a normally-distributed function is 1.96 times as great as the standard deviation, we follow convention by rounding this figure to two, and accept the 95 per cent confidence limit as twice as great as the standard deviation.

Since the object of this article is to present several formulae for practical use, the methods and applications of testing the significance of a total or the mean of one or more sets of scores are presented first. Next, the derivation of the formulas is presented.

2. Methods and applications

Dichotomy; "black or white" type of forecasts.—The simplest application of the test of significance is to the forecasts of the "black or white" type (Crossley, 1954). It has been frequent practice in U. S. Weather Bureau reports, when forecasting for "rain" or no" rain," to allot one point for each correct forecast. The score total n equals the number of correct forecasts, and the chance score total E_c is the number of correct forecasts that the forecaster is expected to obtain by chance. The difference $(n - E_c)$ is a good criterion of skill, because it will be positive if, and only if, the proportion of rain cases among the forecaster's predictions of rain is increased over the climatic frequency.

If there are N forecasts, of which n are correct, the 95 per cent confidence limit for the difference $(n - E_c)$ is equal to or less than

$$N^{\frac{1}{2}}. \tag{1}$$

For example, in table 1 (after Schmidt, 1951), the official score total was $n = 183$, a number which exceeds the chance score E_c by 22 points. Since the confidence limit

$$N^{\frac{1}{2}} = (271)^{\frac{1}{2}} = 16.5 < 22,$$

the official forecasts are declared skillful forecasts.

Next, consider the *difference of two score totals* $(n_1 - n_2)$ for N forecasts, supposing that n_1 correct forecasts were obtained by one system of forecasting and n_2 correct forecasts were obtained by an independent system. The 95 per cent confidence limit for this difference is equal to or less than

$$(2N)^{\frac{1}{2}}. \tag{2}$$

For example, in table 1, the objective system yielded 232 correct forecasts, whereas the official forecasts obtained 183 correct forecasts. Since

$$(232 - 183) > (2 \times 271)^{\frac{1}{2}},$$

the objective system is credited with a significantly better score, although it must not be concluded from this alone that the objective system is the more skillful. The difference $(n_1 - n_2)$ should not be used to compare one forecaster's skill with another, because the chance score E_c can vary from forecaster to forecaster, except when the climatic frequency of rain is 50 per cent.

In many reports, one forecaster's skill is compared with that of another by means of their ratings, S_1 and S_2, where (Brier and Allen, 1951)

$$S = (n - E_c)/(N - E_c).$$

If E_{c1} and E_{c2} denote the chance score-total for each forecaster, the 95 per cent confidence limit of the difference $(S_1 - S_2)$ is equal to or less than

$$\left[\frac{N}{(N - E_{c1})^2} + \frac{N}{(N - E_{c2})^2}\right]^{\frac{1}{2}}.$$

TABLE 1. Results of official forecasts for Washington, D. C., on 9 winter months in 1946, 1947, 1948 and 1949, and results by objective system for the same days (Schmidt, 1951).

Observed	Official forecasts Rain	Official forecasts No rain	Objective forecasts Rain	Objective forecasts No rain	Total
Rain	31	31	35	27	62
Trace	17	16	5	28	209
No rain	40	136	7	169	
Total	88	183	47	224	271
Correct	183		232		
When trace is counted either way	200		237		
By chance	161		184		

For example, in table 1, the objective system's rating is $S_1 = 0.56$, while the official rating is $S_2 = 0.20$. From the above formula, the 95 per cent confidence limit is not greater than 0.24, which is less than the difference $(S_1 - S_2)$. From this we conclude that the objective forecasts are significantly more skillful. (The determination of skill is discussed further below.)

In some schemes of verification, allowance is made for small errors. If a trace is considered a correct verification to the forecast of either "rain" or "no rain," the number of correct forecasts, either by skill or by chance, will be increased. But the 95 per cent confidence limits for the excess score will remain the same.

In the example of table 1, if "trace" is allowed to be counted either way, the 95 per cent confidence limit is not larger than 16.5 points. The number of correct official forecasts is 200. Since only 170 forecasts are expected correct by chance, the official score still is evidence of skill. For the difference of the score totals between the two systems of forecasting, the 95 per cent confidence limit is not greater than $(2N)^{\frac{1}{2}} = 23.3$. Hence, in the example of table 1, the objective score (237) is still significantly greater than the official score (200).

Multiple-choice type of forecasts.—A relatively simple formula for the test of significance can be used in a program in which positive scores are awarded for correct forecasts only (table 2). Let $\alpha_i(A)$ be the score that the forecaster *intends* to obtain by forecasting the

TABLE 2. Illustrates a table of scores in which forecaster receives positive score α_0, α_1, α_2 or α_3 for correctly forecasting event V_0, V_1, V_2 or V_3, respectively, when antecedent condition is A. Score for any incorrect forecast is zero.

		Forecast V_0	V_1	V_2	V_3
Observed	V_0	$\alpha_0(A)$	0	0	0
	V_1	0	$\alpha_1(A)$	0	0
	V_2	0	0	$\alpha_2(A)$	0
	V_3	0	0	0	$\alpha_3(A)$

event V_i when the existing condition is A. If he obtains n correct forecasts, the forecaster's total will be the sum of the corresponding α's. The 95 per cent confidence limit is equal to or less than

$$[\sum_{N} \alpha_i^2(A)]^{\frac{1}{2}}. \tag{3}$$

In words, the 95 per cent confidence limit is not greater than the square root of the sum of the squares of all of the intended scores.

If two forecasters make independent forecasts for the same days, the difference of their score totals will have a 95 per cent confidence limit equal to or less than

$$[\sum_{N} \alpha_{i1}^2(A) + \sum_{N} \alpha_{i2}^2(A)]^{\frac{1}{2}}, \tag{4}$$

where the subscripts 1 and 2 indicate the intended scores of the first and second forecaster, respectively.

If the problem is to compare the performance of a forecaster on N_1 days with the performance of the same forecaster or another forecaster on N_2 other days, it is better to compare the mean scores. The 95 per cent confidence limit of the difference of the two means is equal to or less than

$$\left[\frac{\sum_{N_1} \alpha_i^2(A)}{N_1^2} + \frac{\sum_{N_2} \alpha_i^2(A)}{N_2^2}\right]^{\frac{1}{2}}. \tag{5}$$

For the purpose of this article, *skill* is defined as *the forecaster's ability to analyze and classify meteorological situations so that, within one class, the probability of one subsequent event is increased more, in proportion to its climatic frequency, than the probability of any other subsequent event.* This definition leads to the result that the skill score for a perfect forecast is inversely proportional to the climatic frequency of the subsequent condition following the initial condition; the score for an incorrect forecast, however close to the observed event, is zero (Gringorten, 1951). If, in a large number of days, there are $N(A)$ days with initial condition A, among which there are $n_i(A)$ days when the event V_i subsequently occurred, the climatic frequency is, closely, $n_i(A)/N(A)$, and the score for the correct forecast of V_i is

$$\alpha_i(A) = N(A)/n_i(A).$$

By this method of scoring, the expected chance score-total becomes independent of the forecaster's method of forecasting. Two forecasters' skills, therefore, can be compared directly by comparing their score totals on the same days. The expected average no-skill score is 1.0 per forecast; random forecasts, persistence forecasts and climatological forecasts each net the forecaster an expected minimal average of 1.0 per forecast. Or, $E_c = N$.

Expressions (3), (4) and (5) apply to this case. If the forecaster makes n correct forecasts, the excess score,

$$\sum_{n} \alpha_i(A) - N,$$

has a 95 per cent confidence limit not greater than (3). The difference in score totals of two forecasters,

$$\sum_{n_1} \alpha_{i1}(A) - \sum_{n_2} \alpha_{i2}(A),$$

has the confidence limit given by (4). The 95 per cent confidence limit of the difference in mean scores is bounded by (5).

For example, table 3 gives scores, measuring skill according to the above definition, obtainable for correct forecasts of weather in four classes at McChord Air Force Base, Washington. Fig. 1 shows the running total of scores between 1 July 1953 and 1 November

TABLE 3. Table of scores for correct forecasts at 1930 PST of ceiling and precipitation at McChord AFB, Washington, verifying at 0430 PST of following morning. (Climatic relative frequencies are bracketed figures.)

Initial weather at 1930 PST	Correctly forecast: Clear or scattered	Broken or overcast above 5000 ft	Broken or overcast below 5000 ft	Precipitation
Clear or scattered clouds	1.8 (57.1%)	7.5 (13.4%)	4.0 (25.3%)	22.7 (4.4%)
Broken or overcast above 5000 ft	4.2 (23.7%)	3.3 (30.5%)	3.6 (27.5%)	5.4 (18.5%)
Broken or overcast below 5000 ft	5.8 (17.3%)	5.9 (17.0%)	2.2 (45.9%)	5.1 (19.5%)
Precipitation	7.4 (13.6%)	7.8 (12.8%)	3.0 (33.4%)	2.5 (40.2%)

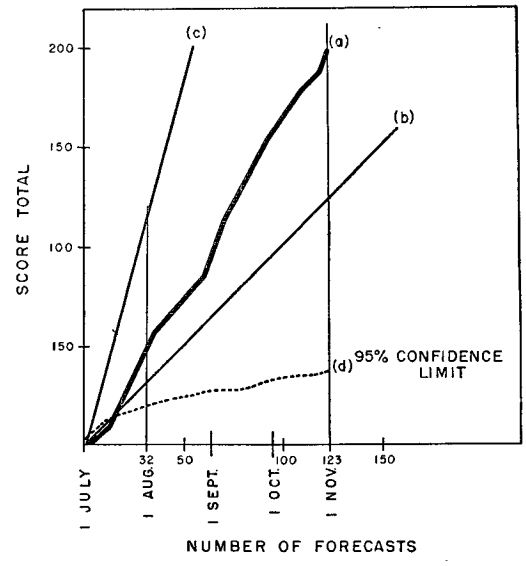

FIG. 1. Showing: (a) running total of scores from 1 July 1953 to 1 November 1953, obtained by forecasters at McChord AFB, Washington, who forecast at 1930 PST for ceiling and precipitation to verify at 0430 PST, (b) expected no-skill running total, (c) expected running total of perfect forecasts, and (d) running value of 95 per cent confidence limit.

1953, achieved by the station forecasters who were forecasting at 1930 PST the sky cover and precipitation in the four classes, to verify at 0430 PST of the next morning. Also shown in fig. 1 is the running top value of the 95 per cent confidence limit of the score total over the expected value. Since the no-skill score should average 1.0 per forecast, the expected no-skill total is given by the straight line with 45-deg. slope. That is, $E_c = N$. Because there are four mutually exclusive events from which to choose a forecast, this system yields an expected average of 4.0 points for each correct forecast. From fig. 1, it is concluded that the forecasters showed significant skill after 32 forecasts. (The test has doubtful value for fewer than 30 forecasts.) After 123 forecasts, it is concluded that the forecasters' average score per forecast is between 1.3 and 1.9.

The running total of the difference between the scores of the forecast team at McChord Air Force Base and the scores of a hypothetical forecaster who always called for persisting weather is shown in fig. 2. The running top value of the 95 per cent confidence limit is plotted on the same figure. When the two lines cross (at 78 forecasts), the decision is reached that the official forecasts are significantly more skillful than forecasts by the other method.

When scores, positive or negative, rewarding or penalizing, are fixed for each combination of initial condition, forecast and observed condition, the formulae are more complex but not prohibitive. For the comparison of the mean scores for N_1 and N_2 forecasts, the 95 per cent confidence limit is equal to or less than

$$\left\{2 \sum_{A,F}\left[\frac{n_1(A,F)}{N_1^2}+\frac{n_2(A,F)}{N_2^2}\right] \times [\alpha(A,F)_{\max}-\alpha(A,F)_{\min}]^2\right\}^{\frac{1}{2}}, \quad (6)$$

where $n_1(A, F)$ denotes the number of days out of the N_1 days on which the initial condition was A and on

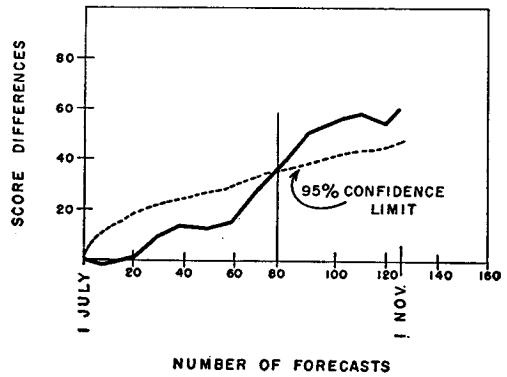

FIG. 2. Running total of differences between scores earned by forecasters at McChord AFB, Washington, in 1953, and scores that would have been earned by persistence.

TABLE 4. Illustration of penalty scores (Vernon, 1953). For example, if V_4 is forecast but V_1 observed, forecaster is penalized 3 points.

	Forecast				
Observed	V_0	V_1	V_2	V_3	V_4
V_0	0	1	2	3	4
V_1	1	0	1	2	3
V_2	2	1	0	1	2
V_3	3	2	1	0	1
V_4	4	3	2	1	0

which the first forecaster predicted the event F. The symbol $n_2(A, F)$ is similarly defined. The symbol $\alpha(A, F)_{\max}$ denotes the maximum score that could be earned (or lost) by the forecaster who, on a day when the initial condition is A, predicts the event F. The symbol $\alpha(A, F)_{\min}$ denotes the minimum score that might be earned (or lost) by the forecaster. The summation is performed over all A's and F's.

Vernon (1953), for instance, has suggested a system that would give a table of penalty scores as in table 4. For the example that Vernon gives in his first table, the average score against forecaster A would be 78/130 and against forecaster B 56/130. The score that would have been expected by chance is the same for the two forecasters. Expression (6) gives the 95 per cent confidence limit as not more than 0.63. (Initial condition is ignored.) But the difference of the two mean scores is 0.17. On the basis of this result, we should require more than 130 forecasts to enable us to declare with confidence that forecaster B is better than A.

Lest the reader think that this article deals only with measures of skill, an example of a measure of the usefulness of a forecast will now be considered. In a given operation, the score awarded to a forecaster on each forecast should be directly proportional to the net profit resulting from the forecast (Gringorten, 1951).

In table 1, let us suppose that the allotted scores, in view of a certain requirement, are as shown in table 5. Then the mean score of the official forecasts would be 0.675, and of the objective forecasts 0.915. From expression (6), the 95 per cent confidence limit is not greater than 0.37. Since the excess score of the objective system is less than 0.37, its operational superiority is not accepted with finality. But we make this statement only with respect to the operation represented by the scores of table 5.

3. Theory

The test of this article is designed for the kind of forecasting in which the forecaster chooses as his

TABLE 5. Scores for hypothetical operational requirement.

	Forecast	
Observed	Rain	No rain
Rain	3	−2
No rain	0	1

forecast one event out of a classification of well-defined, mutually exclusive events, which we might denote as $V_1, V_2, \ldots, V_k$. Further, the score, positive or negative, that is earned by a forecaster for each forecast should be entered in a table of scores such as table 2, 3, 4 or 5.

Let $\alpha(A, F, V)$ be the score that is awarded to a forecaster who forecasts the event F when the initial or antecedent condition is A, but the subsequent observed event is V. For an accurate forecast, F would be the same as V.

Let $n(A, F, V)$ be the number of forecasts that earn for the forecaster the score $\alpha(A, F, V)$. Then the statistic

$$x = \sum_{A,F,V} [\alpha(A, F, V)\, n(A, F, V)] \tag{7}$$

gives the forecaster's score total. When x, or the mean score x/N, is determined from N forecasts, the problem just begins! The experimental average x/N is only an indication of the expected performance of the forecaster. We ask ourseleves: What is the lowest value of x/N below which, we are confident, the value would not lie if we had an indefinitely large sample of forecasts?

In the following we show that, for large N, x/N is, for practical purposes, normally distributed. We find an expression for the expected value of the score total and its variance in terms of the number of forecasts, the scores, and the true probabilities that the forecaster will be right (or wrong). Since these probabilities are not known beforehand, we show that the variance has a maximum value that is determinable for a set of forecasts regardless of the subsequents. From the variance, or its maximum value, we go to the 95 per cent confidence limit which, for a normal distribution, is approximately twice the square root of the variance. If the observed value of the score total exceeds a hypothetical value by more than the 95 per cent confidence limit, we decide that the expected value is significantly greater than the hypothetical value.

Likewise, we find the 95 per cent confidence limit of the difference between two mean scores. If, in two sets of forecasts, the mean score of one set exceeds the mean score of the other set by an amount greater than the 95 per cent confidence limit of this difference, we accept the superiority of the first set of forecasts over the second set.

Let $n(A, F)$ represent the number of times that the forecaster predicts F when the initial condition is A. The ratio $n(A, F, V)/n(A, F)$ is the observed frequency with which the event V occurs among all the days characterized by A and F. When F is the same as V, this ratio is the observed frequency with which the $n(A, F)$ forecasts were accurate. As the n's are made larger, all such ratios approach true probabilities:

$$\frac{n(A, F, V)}{n(A, F)} \underset{n\to\infty}{\longrightarrow} p(V|A, F), \tag{8}$$

where $p(V|A, F)$ represents the conditional probability that the event V will occur when the initial condition is A and the forecaster predicts F.

The following are simple identities:

$$\sum_{i=1}^{k} p(V_i|A, F) = 1, \tag{9}$$

and

$$\sum_{i=1}^{k} n(A, F, V_i) = n(A, F). \tag{10}$$

If, as was assumed in the introduction, the $n(A, F)$ forecasts are independent of each other, the discrete variables $n(A, F, V)$ have the well-known multinomial distribution (Rao, 1952). The sub-total of scores $x(A, F)$ is simply the sum of the $n(A, F, V)$'s weighted by the scores $\alpha(A, F, V)$. As in Rao (1952), the expected sub-total is

$$E[x(A, F)] = n(A, F) \sum_{i=1}^{k} [\alpha(A, F, V_i)\, p(V_i|A, F)]. \tag{11}$$

For the variance, Rao gives

$$\operatorname{var} x(A, F) = n(A, F) \left\{ \sum_{i=1}^{k} \alpha^2(A, F, V_i)\, p(V_i|A, F) - \left[\sum_{i=1}^{k} \alpha(A, F, V_i)\, p(V_i|A, F) \right]^2 \right\}.$$

With a regrouping of terms, this becomes

$$\operatorname{var} x(A, F) = n(A, F) \sum_{j\neq i} \{[\alpha(A, F, V_i) - \alpha(A, F, V_j)]^2 \times p(V_i|A, F)\, p(V_j|A, F)\}. \tag{12}$$

By making $n(A, F) = 1$ in (11) and (12), we conclude that, among all days characterized by A and F, the expected value and the variance of a single forecast are both finite. By the central-limit theorem of statistics (Mood, 1950), therefore, the distribution of the mean score approaches a normal distribution as $n(A, F)$ approaches infinity.

In the work that follows, we suppose that $n(A, F)$ is large enough to make the mean score normally distributed for practical puposes. The sub-total $x(A, F)$, being $n(A, F)$ times as large as the mean score, is also presumed normally distributed with expected value and variance given by (11) and (12).

So far, we have dealt with only a part of the score total, $x(A, F)$. The over-all total, x, is the sum of all

the $x(A, F)$'s:

$$x = \sum_{A,F} x(A, F). \tag{13}$$

The expected value of x is the sum of all the expected values:

$$E(x) = \sum_{A,F,V} [n(A, F)\, \alpha(A, F, V)\, p(V|A, F)]. \tag{14}$$

The variance of x, however, is not so simple without the assumptions in the introduction of this article. The variance of a sum of variables is itself the sum of the individual variances plus twice the sum of the covariances of all pairs of variables. In our case, the covariances should all vanish. True, a forecaster might display a high degree of accuracy, say, among his "no rain" forecasts and a low degree of accuracy among his "rain" forecasts. But, from one day to the next, as long as the forecaster's style of forecasting does not change, the score total for one set of his forecasts should not necessarily improve with, or at the expense of, the score total for another set of his forecasts. Again, it is conceivable that, following one antecedent, the score total will suffer whereas, following another antecedent, the score total will flourish. However, from one month to the next, if each month has a suitable representation of all antecedents, the score total for forecasts following one antecedent should not increase with, or at the expense of, the score total following another antecedent. This argument leaves us with the desirable result, that the variance of the score total x is simply the sum of the variances of the individual totals $x(A, F)$:

$$\text{var } x = \sum_{A,F} \text{var } x(A, F),$$

or

$$\text{var } x = \sum_{A,F} n(A, F) \times \{\sum_{j \neq i} [\alpha(A, F, V_i) - \alpha(A, F, V_j)]^2 \times p(V_i|A, F)\, p(V_j|A, F)\}. \tag{15}$$

Regrettably, the variance of x is dependent upon the performance of the forecaster as represented by the probabilities $p(V|A, F)$. A maximum value of this variance, however, can be determined for a given set of N forecasts. Since, for any $p(V|A, F)$,

$$p(1 - p) \leqslant \tfrac{1}{4},$$

(15) gives

$$\text{var } x \leqslant \tfrac{1}{4} \sum_{A,F} n(A, F) \times \sum_{i=1}^{k-1} [\alpha(A, F, V_i) - \alpha(A, F, V_j)]_{\max}^2. \tag{16}$$

More crudely, since

$$\sum_{j \neq i} p_i p_j < \tfrac{1}{2},$$

$$\text{var } x < \tfrac{1}{2} \sum_{A,F} n(A, F) \times [\alpha(A, F, V_i) - \alpha(A, F, V_j)]_{\max}^2. \tag{17}$$

For the mean score x/N,

$$\text{var } x/N = N^{-2} \text{var } x. \tag{18}$$

If two forecasters meet the same requirements for N forecasts, and obtain the totals x_1 and x_2, respectively,

$$\text{var } (x_1 - x_2) = \text{var } x_1 + \text{var } x_2 - 2 \text{ cov } (x_1, x_2).$$

If the forecasters use similar methods of forecasting, they would tend to make the same forecasts and the same errors. The covariance should be positive, tending to reduce the size of $\text{var}(x_1 - x_2)$. Therefore,

$$\text{var } (x_1 - x_2) \leqslant \text{var } x_1 + \text{var } x_2. \tag{19}$$

If the mean scores of two forecasters are compared when their series of forecasts are not the same, or if a forecaster's mean score for one set of N_1 forecasts is compared with an earlier performance, the two averages should be independently distributed, and therefore

$$\text{var } \left(\frac{x_1}{N_1} - \frac{x_2}{N_2}\right) = \frac{1}{N_1^2} \text{var } x_1 + \frac{1}{N_2^2} \text{var } x_2. \tag{20}$$

Putting the results (16) and (20) together, we can say it is generally true that

$$\text{var } (x_1/N_1 - x_2/N_2) \leqslant \frac{1}{4} \left\{ \sum_{A,F} \left[\frac{n_1(A, F)}{N_1^2} + \frac{n_2(A, F)}{N_2^2}\right] \times \sum_{i=1}^{k-1} [\alpha(A, F, V_i) - \alpha(A, F, V_j)]_{\max}^2 \right\}, \tag{21}$$

or,

$$\text{var } (x_1/N_1 - x_2/N_2) < \frac{1}{2} \left\{ \sum_{A,F} \left[\frac{n_1(A, F)}{N_1^2} + \frac{n_2(A, F)}{N_2^2}\right] \times [\alpha(A, F, V_i) - \alpha(A, F, V_j)]_{\max}^2 \right\}. \tag{22}$$

For the "black or white" type of forecast, when one point is awarded for a correct forecast, α equals 0 or 1. From (16),

$$\text{var } x = \text{var } n \leqslant \tfrac{1}{4} \sum_{A,F} n(A, F) = \tfrac{1}{4}N.$$

From (16) and (19),

$$\text{var } (x_1 - x_2) = \text{var } (n_1 - n_2) \leqslant \tfrac{1}{4}(2N).$$

When scores are awarded for correct forecasts only, (16) gives

$$\text{var } x \leqslant \tfrac{1}{4} \sum_{A,\,F} [n(A, F)\, \alpha^2(A, F)] = \tfrac{1}{4} \sum_{i=1}^{N} \alpha_i^2,$$

where $\alpha(A, F)$ is the only score obtainable for a forecast of F following the antecedent A, and α_i denotes the forecaster's intended score on the i-th forecast.

Using the approximation that the 95 per cent confidence limit is equal to twice the square root of the variance, the results of section 2 follow easily. The resemblance of (6) to (22) is readily apparent.

4. Concluding remarks

The above treatment of the subject has limited application, limited to the kind of forecasting program in which a forecaster selects for each forecast one out of a set of well-defined, mutually exclusive events. A table of scores, such as table 2, 3, 4 or 5 must be used, constructed to meet the specific requirements of a verification program. There is no attempt in this article to test the significance of indices of usefulness of skill, except the sum total of scores, the mean score, and the frequently used "skill score" (Brier and Allen, 1951).

The excess of the total number of correct forecasts of a "black or white" type over the number that are expected to be correct by chance is a measure of skill. In a multiple-choice type of program, the total of the scores can also be a measure of skill, if the definition of skill given above is accepted. Thus, the 95 per cent confidence limit of a score total enables us to determine, with confidence, that the forecasters have exhibited significant skills.

Through expressions (2), (4) and (5), we are able to decide conclusively that one system of forecasting is better than another. By "better" we might mean better in skill or better in operational value, depending upon the selection of scores in the scoring table.

Lastly, there has been included in this article one example of a procedure for sampling a forecast system just often enough to declare the forecasts as successful, or to declare one system superior to another. Further sampling of the forecasts enables us to narrow the limits of the forecaster's expected performance.

Acknowledgments.—Mr. R. C. Schmidt, Meteorologist-in-Charge, U. S. Weather Bureau, Washington National Airport, Washington, D. C., kindly furnished the extra information on the number of cases of trace in table 1. Professor Max A. Woodbury, University of Pennsylvania, while examining the manuscript of this article suggested the formula for the 95 per cent confidence of the "skill score" (Brier and Allen, 1951) which was used in this article.

REFERENCES

Brier, G. W., and Allen, R. A., 1951: Verification of weather forecasts. *Compendium Meteor.*, Boston, Amer. meteor. Soc., 841–848.

Crossley, A. F., 1954: Measures of success in forecasting. *Meteor. Mag.*, **83**, 66–73.

Gringorten, I. I., 1951: The verification and scoring of weather forecasts. *J. Amer. stat. Assoc.*, **46**, 279–296.

Leight, W. G., 1953: The use of probability statements in extended forecasting. *Mon. Wea. Rev.*, **81**, 349–356.

Mood, A. M., 1950: *Introduction to the theory of statistics.* New York, McGraw-Hill Book Co., 433 pp.

Rao, C. R., 1952: *Advanced statistical methods in biometric research.* New York, John Wiley and Sons, 390 pp.

Schmidt, R. C., 1951: A method of forecasting occurrence of winter precipitation two days in advance. *Mon. Wea. Rev.*, **79**, 81–95.

Vernon, E. M., 1953: A new concept of skill score for rating quantitative forecasts. *Mon. Wea. Rev.*, **81**, 326–329.