BOUNDED PROBABILITY PROPERTIES OF KOLMOGOROV-SMIRNOV AND SIMILAR STATISTICS FOR DISCRETE DATA

BY JOHN E. WALSH

(Received April 24, 1963; revised Oct. 8, 1963)

Summary

A somewhat general class of situations, that include Kolmogorov-Smirnov type results as special cases, is considered. These situations, which are described in the following sections, are required to have uniquely determined probability properties when the sample values used are from continuous populations of any nature. If the populations sampled are discrete, however, these probability values are not uniquely determined. This paper shows that the values for the continuous case represent bounds for the values that occur in any discrete case. The method used to show that these bound relations hold consists in noting that any discrete data situation can be interpreted as a situation involving the grouping of continuous data. Then bound relationships are established between the values of probabilities for the grouped data situations and the corresponding ungrouped data situations, which are the situations considered for the case of the continuous data. These bounds on probabilities for discrete data cases should be useful for practical applications. In practice, all data are discrete (due to limitations in measurement accuracy).

1. Introduction

The Kolmogorov-Smirnov tests and confidence bands (see, for example, ref. [1]) can be based on statistics each of which can be evaluated as the supremum of a set of values that depend on the observations. These are examples of the many types of results (univariate and multivariate data) that can be based on statistics of this nature. In general, the probabilities considered are for over-all relations that are expressed in terms of some elementary relations. An elementary relation consists of a comparison of a statistic (of the kind considered) with a given number, where this comparison involves a specified one of $<$, $\leqq$, $>$, $\geqq$. Each over-all relation of the type considered can be expressed in terms of elementary relations involving $>$ and $\geqq$, or in terms of ele-

mentary relations involving $<$ and $\leqq$. When $>$ and $\geqq$ are used, the over-all relation asserts that at least a specified number of the elementary relations hold; for $<$ and $\leqq$, the assertion is that at most a specified number of the elementary relations hold.

Some of these situations are such that the joint distribution of the statistics occurring in the elementary relations is the same for all cases where the data are sample values from continuous populations. For a class that seems to include nearly all situations of this nature, the probability of an over-all relation for the case of continuous data is shown to furnish a bound (of a determined kind) for the value of this probability for discrete data cases. In particular, these bounds are shown to be applicable to Kolmogorov-Smirnov results and to some extensions of these results.

The verification of these probability bounds is based on some conceptual ideas that are not of a very complicated nature. Consequently, no attempt is made to present the verification in a formal mathematical way. Instead, a discussion method of presentation is used.

2. Discussion and results

The material of this paper is based on a conceptual use of the grouping of data. Grouping of data is, in reality, a conversion of the data to discrete form. That is, for a given group, a specified representative value is chosen and this value occurs with a probability equal to that for the group; here the representative value could be any of the values in this group. Given a discrete probability distribution, an unlimited number of combinations of a continuous distribution and a grouping of data can result in this discrete distribution.

For a given discrete situation, consider the joint distribution of the statistics occurring in the elementary relations. Suppose that a continuous situation and a grouping of data exist such that the following two conditions are satisfied: First, the joint distribution of the statistics is the same as for the given discrete situation. Second, each statistic is evaluated as the supremum of a set of values that is contained in the corresponding set of values for the ungrouped data case. Then the probability of an over-all relation for the continuous ungrouped case, which has a unique value for the situations considered, is a bound for this probability for the given discrete situation. Thus, if such a combination of a continuous situation and grouping of data exists for every discrete situation, then the unique probability for the continuous case furnishes a bound for all values that can occur for discrete situations.

Let us consider verification of these statements, which represent the principal results of the paper. Since the discrete situation and the

grouped data continuous situation have the same probability properties, it is sufficient to show that the probability for the grouped data continuous situation is bounded by the value for this continuous situation without the data being grouped. By hypothesis, the probability is the same for all continuous situations when the data are not grouped, and therefore equal to the value for this special continuous situation.

First, consider an elementary relation. Evidently, the probability that the supremum of a set of random values exceeds (or is $\geqq$) a given number is at least equal to the probability that the supremum of a subset of this set exceeds (or is $\geqq$) this number. Similarly, the probability that the supremum of a set is less than (or $\leqq$) a given number is at most equal to the probability that the supremum of a subset is less than (or $\leqq$) this given number.

Now consider an over-all relation of the $>$ and $\geqq$ kind. Again, the reasoning is based on the suprema of sets as compared to the suprema of subsets of these sets, combined with the fact that the common elements of each set and subset have identical probability properties. The result is easily seen to be that the probability of the over-all relation for the grouped data case is at most equal to the value for the ungrouped data case. Likewise, for an over-all relation of the $<$ and $\leqq$ kind, the probability for the grouped data situation is at least equal to the probability for ungrouped data.

Thus, the stated bounds hold when a suitable combination of a continuous situation and data grouping exist for every discrete situation. The large amount of freedom available in choosing a grouping (including the selection of representative values), and the ability to use any continuous situation, indicate that the bounds are nearly always applicable for all discrete situations. In fact, these bounds would seem to be applicable for all situations where the statistics that appear in the elementary relations are of a reasonable nature and unique probabilities occur for the continuous case. However, no verification is presented for these conjectures.

Justification for the applicability of the bounds is presented for a class of situations that includes the usual Kolmogorov-Smirnov situations as special cases. This material is contained in the next section.

3. Kolmogorov-Smirnov type cases

Here the data are univariate and two types of cases occur. For one case, the data are a sample from a population with cumulative distribution function $F(x)$. Here the interest is obtaining confidence bands for $F(x)$; these bands could be used to test the null hypothesis that

$F(x)=F_0(x)$ over the interval of values considered for x. In the other case, the data are two independent samples and tests are considered. Here, the null hypothesis asserts that these samples are from populations that have the same distribution over the interval of values considered for x and the interest is in significance levels.

For the case involving only one sample, the confidence bands can be based on one or both of statistics of the forms

$$d_1=\sup_{x\in I}[F_n(x)-aF(x)], \qquad d_2=\sup_{x\in I}[F(x)-a'F_n(x)]$$

where I, perhaps random depending on the sample, is the specified interval over which the suprema are taken and $F_n(x)$ is the empirical cumulative distribution function for the sample (of size n). The Kolmogorov-Smirnov situations occur for I equal to the entire x-axis and $a=a'=1$. A number of other situations have been considered in the statistical literature (see, for example, refs. [2], [3], [4], [5]). Each of the relations $d_1\leqq d^{(1)}$, $d_1<d^{(1)}$, $d_2\leqq d^{(2)}$, $d_2<d^{(2)}$ furnishes a one-sided confidence band for $F(x)$ over I. Each of $d_1\leqq d^{(2)}$ and/or $d_2\leqq d^{(2)}$, $d_1\leqq d^{(1)}$ and/or $d_2<d^{(2)}$, $d_1<d^{(1)}$ and/or $d_2\leqq d^{(2)}$, $d_1<d^{(1)}$ and/or $d_2<d^{(2)}$ furnishes a two-sided confidence band for $F(x)$ over I. In particular, for $a=a'=1$ and I equal to the entire x-axis, tests based on the Kolmogorov statistic

$$\sup_{-\infty<x<\infty}|F_n(x)-F(x)|$$

can be obtained.

For many choices of I, the joint distribution of d_1 and d_2 is the same for all continuous populations. Consideration is limited to these choices.

Let the intervals for grouping the data be of the form $y_1<x\leqq y_2$ (all data are grouped). The effect of this grouping is to limit the values of x for which $F_n(x)$ is known to the boundary points of these intervals (which are taken to be the representative values for the groups). For grouped data situations, d_1 and d_2 are evaluated as the suprema taken over the values of x (in I) for which $F_n(x)$ is known. Thus, the possible values of $F_n(x)-aF(x)$ used in determining d_1 for the grouped data case are a subset of these values for the case of ungrouped data. Likewise with respect to $F(x)-a'F_n(x)$ and d_2. This implies that the confidence coefficient for a band based on ungrouped data is at most as large as the confidence coefficient for the corresponding band based on any grouping of these data.

Finally, consider the case where $F(x)$ is discrete. Let the boundary points for the grouping be the points of the discrete population that have nonzero probability. There evidently exist an unlimited number of continuous populations such that the cumulative distribution function

for the discrete distribution which results from the grouping is the same as the discrete $F(x)$ that is considered. Since this holds for any discrete $F(x)$, the confidence coefficient value for the continuous case is a lower bound for any value that occurs for a discrete case.

Next, consider the two-sample problem case. Let $F_m(x)$ be the empirical cumulative distribution function for the first sample while $G_n(x)$ is this function for the second sample. The tests can be based on one or both of statistics

$$D_1=\sup_{x\in I'}[F_m(x)-AG_n(x)], \qquad D_2=\sup_{x\in I'}[G_n(x)-A'F_m(x)],$$

where I', perhaps random depending on the samples, is the specified interval over which the suprema are taken. The Smirnov situations occur for I' equal to the entire x-axis and $A=A'=1$. A number of other situations have been considered in the literature (see, for example, refs. [6], [7], [8], [9], [10]).

Here, consideration is limited to the many situations where the joint distribution of D_1 and D_2 is the same for all continuous populations under the null hypothesis. The same type of grouping is used and the necessary set-subset relations are found to hold when the boundaries are used as the representative values for the groups.

Under the null hypothesis, both populations have the same cumulative distribution function, say $G(x)$. Consider a case where $G(x)$ is discrete. The grouping is chosen in the same manner as for the one-sample confidence band case and it is easily seen that the discrete distribution resulting from the grouping is, for every discrete population, the same as the discrete $G(x)$ considered. Thus, the significance level for the continuous case furnishes an upper bound for the significance level that occurs for any discrete situation.

REFERENCES

[1] D. A. Darling, "The Kolmogorov-Smirnov, Cramér-von Mises tests," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 823-838.

[2] L.-C. Chang, "On the ratio of an empirical distribution function to the theoretical distribution function," *Acta Math. Sinica*, Vol. 5 (1955), pp. 347-368 (in Chinese, English summary).

[3] S. Malmquist, "On certain contours for distribution functions," *Ann. Math. Statist.*, Vol. 25 (1954), pp. 523-533.

[4] G. M. Maniya, "Generalization of the criterion of A. N. Kolmogorov for an estimate for the law of distribution for empirical data," *Doklady Akademii Nauk SSSR (N.S.)*, Vol. 69 (1949), pp. 495-497 (in Russian).

[5] A. Rényi, "On the theory of order statistics," *Acta Mathematica Academiae Scientiarum Hungaricae*, Vol. 4 (1953), pp. 191-232.

[6] I. D. Kvit, "On N. V. Smirnov's theorem concerning the comparison of two samples," *Doklady Akademii Nauk SSSR (N.S.)*, Vol. 71 (1950), pp. 229-231 (in Russian).

[7] H. L. Berlyand and I. D. Kvit, "On a problem of comparison of two samples,"

Dopovidi Akad. Nauk Ukrain. RSR 1952, (1952), pp. 13-15 (in Ukrainian, Russian summary).

[8] E. F. Drion, "Some distribution-free tests for the difference between two empirical cumulative distribution functions," *Ann. Math. Statist.*, Vol. 23 (1952), pp. 563-574.

[9] C. K. Tsao, "An extension of Massey's distribution of the maximum deviation between two sample cumulative step functions," *Ann. Math. Statist.*, Vol. 25 (1954), pp. 587-592.

[10] S.-J. Wang, "On the limiting distribution of the ratio of two empirical distributions," *Acta Math. Sinica*, Vol. 5 (1955), pp. 253-267 (in Chinese, English summary).