TESTS AND CONFIDENCE INTERVALS BASED ON THE METRIC d_2[1]

By Judah Rosenblatt

University of New Mexico and Sandia Corporation

1. Introduction. In this paper tests and confidence intervals based on a metric similar to that used in the Kolmogorov-Smirnov tests are introduced. While the tests are slightly more difficult computationally, they have somewhat better discriminating power against certain alternatives. The confidence intervals for probabilities of intervals have the advantage over those given by the Kolmogorov-Smirnov statistics, of shorter maximum length for the same sample size. The statistic studied here was investigated by Kuiper in [4].

2. Definition and properties of d_2. Let $\mathfrak{D}$ be the class of one-dimensional distribution functions, and for F, G in $\mathfrak{D}$, let

$$d_2(F, G) = \sup_{x_2 > x_1} |[F(x_2) - F(x_1)] - [G(x_2) - G(x_1)]|.$$

Theorem 2.1. *d_2 is a metric, $d_1 \leqq d_2 \leqq 2d_1$, where d_1 is the uniform metric,*

$$d_2(F, G) = \sup_{I \text{ an interval}} |P_F(I) - P_G(I)|,$$

$$d_2(F, G) = \sup_x [F(x) - G(x)] + \sup_x [G(x) - F(x)].$$

The proof is a straightforward consequence of the definition of d_2, and also appears in Brunk, [1].

Definition. For $F_o \, \varepsilon \, \mathfrak{D}$, e, $k \, \varepsilon \, [0, 1]$, let

$$C_{e,k,F_o} = \{F \, \varepsilon \, \mathfrak{D} \colon F_o(x) + e - k \leqq F(x) \leqq F_o(x) + e, \text{ all } x\}.$$

Theorem 2.2. $\bigcup_{e \varepsilon [0,k]} C_{e,k,F_o} = \{F \, \varepsilon \, \mathfrak{D} \colon d_2(F_o, F) \leqq k\}$.

Proof. Suppose $G \, \varepsilon \, C_{e',k,F_o}$ for some $e' \, \varepsilon \, [0, k]$. For $I = (x_1, x_2]$ the maximum probability that G can assign to I is $[F_o(x_2) + e') - [F_o(x_1) - e' - k] = F_o(x_2) - F_o(x_1) + k$, while similarly the minimum probability G can assign to I is $F_o(x_2) - F_o(x_1) - k$. Hence

$$P_{F_o}(I) - k \leqq P_G(I) \leqq P_{F_o}(I) + k.$$

This holds for each interval I, and since it is true for each $e' \, \varepsilon \, [0, k]$, we have

$$\bigcup_{e \varepsilon [0,k]} C_{e,k,F_o} \subset \{F \, \varepsilon \, \mathfrak{D} \colon d_2(F_o, F) \leqq k\}.$$

Now suppose $G \, \varepsilon \, \{F \, \varepsilon \, \mathfrak{D} \colon d_2(F_o, F) \leqq k\}$. Let $\sup_x [G(x) - F_o(x)] = S \leqq k$, $\sup_x [F_o(x) - G(x)] = t \leqq k - S$. Then clearly

$$G \, \varepsilon \, C_{S,S+t,F_o} \subset C_{S,k,F_o} \subset \bigcup_{e \varepsilon [0,k]} C_{e,k,F_o}.$$

Received February 16, 1960; revised August 20, 1962.

[1] This paper is based on the author's Ph.D. thesis at Columbia University under the guidance of Professor T. W. Anderson. Supported in part by the Office of Naval Research, and in part under the auspices of the United States Atomic Energy Commission.

Hence

$$\{F \varepsilon \mathfrak{D}: d_2(F_o, F) \leqq k\} \subset \bigcup_{e\varepsilon[0,k]} C_{e,k,F_o}.$$

3. Definition of null hypothesis and some properties of d_2 related to this definition. In this section it is assumed that F_o is some given distribution function.

For $k \varepsilon (0, 1)$ let

$$\mathcal{H}_{o,k} = \{F \varepsilon \mathfrak{D}: d_2(F_o, F) \leqq k\}.$$

It is desired to test the hypothesis that the distribution function of the independent random variables $X_1, \cdots, X_n$ is in $\mathcal{H}_{o,k}$, where k is chosen on the basis of some realistic considerations.

DEFINITION. For $F \varepsilon \mathfrak{D}$, $0 \leqq e \leqq k < 1$, let

$$F_e^*(x) = \begin{cases} F_o(x) - k + e & \text{if } F(x) < F_o(x) - k + e, \\ F(x) & \text{if } F_o(x) - k + e \leqq F(x) \leqq F_o(x) + e, \\ F_o(x) + e & \text{if } F(x) > F_o(x) + e. \end{cases}$$

THEOREM 3.1. *For $F \varepsilon \mathfrak{D}$*

$$\inf_{H\varepsilon\mathcal{H}_{o,k}} d_2(F, H) = \inf_{e\varepsilon[0,k]} d_2(F, F_e^*).$$

PROOF. From Theorem 2.2

$$\inf_{H\varepsilon\mathcal{H}_{o,k}} d_2(F, H) = \inf_{e\varepsilon[0,k]} \inf_{H\varepsilon C_{e,k,F_o}} d_2(F, H).$$

Thus it need only be proved that for each $e' \varepsilon [0, k]$

$$\inf_{H\varepsilon C_{e',k,F_o}} d_2(F, H) = d_2(F, F_{e'}^*). \tag{3.1}$$

To accomplish this it will be shown that for each $G \varepsilon C_{e',k,F_o}$

$$d_2(F, F_{e'}^*) \leqq d_2(F, G), \tag{3.2}$$

which will imply (3.1) because $F_{e'}^* \varepsilon C_{e',k,F_o}$. We can show (3.2) by showing

$$\sup_x [F(x) - F_{e'}^*(x)] \leqq \sup_x [F(x) - G(x)] \tag{3.3}$$

and

$$\sup_x [F_{e'}^*(x) - F(x)] \leqq \sup_x [G(x) - F(x)], \tag{3.4}$$

as is seen from the last part of Theorem 2.1. To prove (3.3) we need only consider those x for which $F(x) > G(x)$, since because F and G are distribution functions, $\sup_x [F(x) - G(x)] \geqq 0$.

Thus suppose $F(x') > G(x')$. Then since $G \varepsilon C_{e',k,F_o}$, either

$$F(x') = F_{e'}^*(x') > g(x'), \qquad \text{or} \qquad F(x') > F_o(x') + e' \geqq G(x'),$$

which proves (3.3). Similarly to prove (3.4) we need only consider those x for which $G(x) > F(x)$.

THEOREM 3.2. *For each distribution function* $F \not\in \mathcal{K}_{o,k}$, *and each* $H^* \in \mathcal{K}_{o,k}$ *for which* $\inf_{H \in \mathcal{K}_{o,k}} d_2(F, H) = d_2(F, H^*)$, *we have*

$$d_2(F, H^*) + d_2(H^*, F_o) = d_2(F, F_o) \qquad \text{and} \qquad d_2(H^*, F_o) = k.$$

PROOF. Since $F \not\in \mathcal{K}_{o,k}$, we may assume

$$d_2(F_o, F) = k + l, \qquad l > 0. \tag{3.5}$$

All we need show is that for some e'

$$d_2(F, F^*_{e'}) = l. \tag{3.6}$$

This follows since by (3.5), the triangle inequality, and the fact that

$$d_2(H^*, F_o) \leqq k, \tag{3.7}$$

we must have $d_2(F, H^*) \geqq l$.

To prove (3.6) we note that by (3.5) there is an interval, $[x_1, x_2]$, such that without loss of generality $F_o(x_1) - F(x_1) = S = \sup_x [F_o(x) - F(x)]$ and $F(x_2) - F_o(x_2) = k + l - S = \sup_x [F(x) - F_o(x)]$. (Here x_i may, for completeness, stand also for $x_i +$. This does not affect the arguments.) If e' is chosen such that $k - S \leqq e' \leqq k + l - S$, then $d_2(F, F^*_{e'}) = l$, since clearly

$$\sup_x [F^*_{e'}(x) - F(x)] = S - (k - e') \qquad \text{and}$$
$$\sup_x [F(x) - F^*_{e'}(x)] = (k + l - S) - e',$$

because both suprema are achieved at the same points as the suprema of

$$[F(x) - F_o(x)] \qquad \text{and} \qquad [F_o(x) - F(x)].$$

Note that from (3.6) and the fact that $F^*_{e'} \in \mathcal{K}^*_{o,k}$ we know that there is an $H^* \in \mathcal{K}_{o,k}$ for which $d_2(F, H^*) = \inf_{H \in \mathcal{K}_{o,k}} d_2(F, H)$, namely $F^*_{e'}$.

4. Formula for the limiting distribution of the statistic on which the tests and estimates are based. For $F \in \mathfrak{D}$ continuous let $U_i = F(X_i)$, $i = 1, \ldots, n$. Then if F is the distribution function of the X_i, the U_i are independent random variables with common uniform distribution on [0, 1]. Let the random process $G_n(u)$, $u \in [0, 1]$, be defined by $G_n(u) =$ proportion of $U_1, \cdots, U_n \leqq u$, and let $F_n(x) =$ proportion of $X_1, \cdots, X_n \leqq x$. ($F_n(\cdot)(\omega)$ is the empirical distribution function based on the observed values $X_1(\omega), \cdots, X_n(\omega)$ of $X_1, \cdots, X_n$.) Then

$$n^{\frac{1}{2}} d_2(F_n, F) = \sup_{0 \leqq u_1 \leqq u_2 \leqq 1} |n^{\frac{1}{2}}[G_n(u_2) - u_2] - n^{\frac{1}{2}}[G_n(u_1) - u_1]|$$
$$= \sup_{u \in [0,1]} n^{\frac{1}{2}}[G_n(u) - u] + \sup_{u \in [0,1]} n^{\frac{1}{2}}[u - G_n(u)] \equiv D_n^+ + D_n^-.$$

In [3], pp. 202–203, the limiting distribution of $\frac{1}{2}(D_n^+ + D_n^-)$ when the U_i are independent with the uniform distribution on [0, 1] was computed. From it we find that for continuous F

$$T_2(z) = \lim_{n \to \infty} P_F\{n^{\frac{1}{2}} d_2(F_n, F) \leqq z\} = 1 + \sum_{m=1}^{\infty} [2 - 8m^2z^2] \exp[-2m^2z^2],$$

the justification for using the limiting stochastic process to find the limiting distribution of $(D_n^+ + D_n^-)$ being due to Donsker [2]. It can be shown that for $z > 1.5$ an excellent approximation to $T_2(z)$ is furnished by using but one term, $1 + [2 - 8z^2] \exp[-2z^2]$, of the above series. For example $T_2(1.65) \doteq .905$, and $T_2(2) \doteq .99$. Tables of T_2 are given in [5], Table 15.7.

5. The proposed tests and their properties. The proposed test is: Reject $\mathcal{H}_{o,k}$ when $X_1(\omega), \cdots, X_n(\omega)$ are observed $\Leftrightarrow n^{\frac{1}{2}} d_2(F_n, F_o)(\omega) > n^{\frac{1}{2}}k + h_{2,\alpha}$, where $T_2(h_{2,\alpha}) = 1 - \alpha$. Since by the triangle inequality for $F \varepsilon \mathcal{H}_{o,k}$ if

$$d_2(F_n, F_o) > k + h_{2,\alpha}/n^{\frac{1}{2}} \qquad \text{then} \qquad d_2(F_n, F) > h_{2,\alpha}/n^{\frac{1}{2}}$$

we have

THEOREM 5.1. *For* $F \varepsilon \mathcal{H}_{o,k}$, $\limsup_{n\to\infty} P_F\{\text{rej } \mathcal{H}_{o,k}\} \leqq \alpha$. It should be noted for computational purposes that

$$d_2(F_n, F_o) = \max_{i=1,\cdots,n} [F_o(X_{[i]}-) - (i-1)/n] + \max_{i=1,\cdots,n} [i/n - F_o(X_{[i]})],$$

where $X_{[j]}(\omega)$ is that $X_i(\omega)$ which is jth in order of magnitude.

Let $[[r]]$ denote the greatest integer less than or equal to r. Then from considerations similar to those in [6], we can easily prove

THEOREM 5.2. *If n is any integer for which*

$$\inf_{p \varepsilon [0,1-l]} \sum_{\nu=0}^{[[n(p+l)-n^{\frac{1}{2}}h_{2,\alpha}]]} \binom{n}{\nu} p^\nu (1-p)^{n-\nu} \geqq 1 - \beta,$$

then

$$P_F\{\text{rej } \mathcal{H}_{o,k}\} \geqq 1 - \beta \quad \textit{for all} \quad F \quad \textit{such that} \quad \inf_{H \varepsilon \mathcal{H}_{o,k}} d_2(F, H) \geqq l.$$

The proof is the same as that of Theorem 3.1 in [6], when one notices the facts that $d_2(F, F_o) \geqq k + l$ (from the triangle inequality), and the number of observations falling in $(y_1, y_2]$ has the binomial distribution with parameter $n[F(y_2) - F(y_1)]$.

Clearly n may be computed as in [6]. As to whether the n thus found is too conservative, for α sufficiently small the results are similar to those in [6], since after applying Theorem 3.2 the reasoning is the same as that of [6].

It should be mentioned that d_1 (the uniform metric) and d_2 are basically different in the sense that there are some alternatives against which only one will discriminate. To see this let F_o be the uniform distribution function on [0, 1], in which case $\mathcal{H}_{o,0}$ will consist only of F_o. In order to keep the tests of $\mathcal{H}_{o,0}$ based on d_1 and d_2 comparable, we let $\alpha = .05$ and $n = 1600$ for both. Then $h_{1,\alpha} = 1.36$ and $h_{2,\alpha} = 1.75$. Let

$$F^*(x) = \begin{cases} 0 & \text{if } x < 0, \\ .025 & \text{if } 0 \leqq x \leqq .05, \\ x & \text{if } .05 < x \leqq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

It is easily seen that the test based on d_2 rejects $\mathcal{H}_{o,0}$ with probability one when $F = F^*$. However with the aid of Uspensky's result as given in [6], Equation (3.28), it is easy to show that the probability of rejection when $F = F^*$ using the test based on d_1 is close to .05.

Similarly it can be shown that the test based on d_1 has better discriminating power against alternatives of the form

$$F^{**}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leqq x \leqq .5 - \delta, \\ .5 - \delta & \text{if } .5 - \delta < x \leqq .5 + \delta, \\ x & \text{if } .5 + \delta < x \leqq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The suggested tests are certainly not the only desirable ones. For example it can easily be shown that if we define $\mathcal{H}_o^*$ as in [6], Definition 2.1, F_n^* as in [5] Definition 2.3 with $G = F_n$, then

$$\inf_{H \epsilon \mathcal{H}_o^*} d_2(F_n, H) = d_2(F_n, F_n^*).$$

In this case the test given by

Reject $\mathcal{H}_o^*$ when $X_1(\omega), \cdots, X_n(\omega)$ are observed $\Leftrightarrow n^{\frac{1}{2}} d_2(F_n, F_n^*)(\omega) > h_{2,\alpha}$

satisfies Theorems 5.1 and 5.2, and n can be found in the same way as before.

In this case the alternatives against which there is good power are different from those treated in [6], though the null hypothesis is the same.

6. Construction of simultaneous confidence intervals. It can be seen from Theorem 2.1 that $d_2(F_n, F) \leqq h_{2,\alpha}/n^{\frac{1}{2}}$ if and only if for all intervals I

$$-h_{2,\alpha}/n^{\frac{1}{2}} \leqq P_{F_n}(I) - P_F(I) \leqq h_{2,\alpha}/n^{\frac{1}{2}}.$$

Thus simultaneous asymptotic $1 - \alpha$ level confidence intervals for probabilities of intervals I are given by

$$[P_{F_n}(I) - h_{2,\alpha}/n^{\frac{1}{2}}, P_{F_n}(I) + h_{2,\alpha}/n^{\frac{1}{2}}].$$

Here if $I = (a, b]$ then $P_{F_n}(I) = F_n(b) - F_n(a)$, with the usual modifications if I is for example closed.

Such intervals could be constructed using d_1, but the intervals given here have the advantage of smaller maximum length for given α than those from d_1.

It seems worthwhile to examine the advantage of the d_2 confidence intervals over those from d_1. The asymptotic $1 - \alpha$ confidence intervals from d_1 are

$$[P_{F_n}(I) - 2h_{1,\alpha}/n^{\frac{1}{2}}, P_{F_n}(I) + 2h_{1,\alpha}/n^{\frac{1}{2}}].$$

Hence for given α the ratio

$$\frac{\text{maximum length of } d_2\ 1 - \alpha \text{ asymptotic confidence interval}}{\text{maximum length of } d_1\ 1 - \alpha \text{ asymptotic confidence interval}} = \frac{h_2, \alpha}{2h_1, \alpha}.$$

For $\alpha = .05$ this ratio is $1.75/2.72 \doteq .64$, (i.e. for large n the d_2 interval is approximately 64 per cent as long as the longest d_1 interval), while for $\alpha = .01$ this ratio is $2/3.3 \doteq .61$.

In particular the above can be used to obtain simultaneous asymptotic $1 - \alpha$ level confidence intervals for the parameters $p_1, \cdots, p_k$ of a multinomial distribution by letting X_m be a random variable which has value j if the mth observation is in the jth class. Then

$$F_{X_m}(\lambda) = \sum_{j \leqq \lambda} p_j .$$

Letting $J_n(j)$ be the jump of the empirical distribution function at j, we have

$$p_j \, \varepsilon \, [J_n(j) - h_{2,\alpha}/n^{\frac{1}{2}}, J_n(j) + h_{2,\alpha}/n^{\frac{1}{2}}]$$

with asymptotic probability at least $1 - \alpha$, simultaneously for all $j = 1, 2, \cdots, k$.

This method can be made exact for finite samples by replacing $h_{2,\alpha}$ by $h_{2,\alpha,n}$, where for F continuous $P_F\{n^{\frac{1}{2}} d_2(F_n, F) > h_{2,\alpha,n}\} = \alpha$.

REFERENCES

[1] Brunk, H. D. (1962). On the range of the difference between hypothetical distribution function and Pyke's modified empirical distribution function. *Ann. Math. Statist.* **33** 525–532.

[2] Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov limit theorems. *Ann. Math. Statist.* **23** 277–281.

[3] Kac, M., Kiefer, J. and Wolfowitz, J. (1955). On tests of normality and other tests of goodness-of-fit based on distance methods. *Ann. Math. Statist.* **26** 189–211.

[4] Kuiper, N. H. (1960). Tests concerning random points on a circle. *Proc. Kon. Nederl. Akad. van Wet., Amsterdam Ser. A*, **63**; *Indag. Math.* **22** 38–47.

[5] Owen, D. B. (1962). *Handbook of Statistical Tables*. Addison-Wesley, Cambridge.

[6] Rosenblatt, Judah (1962). Some modified Kolmogorov-Smirnov tests of approximate hypotheses and their properties. *Ann. Math. Statist.* **33** 513–524.