

AD737567

**AN ALGORITHM FOR CRITICAL VALUES OF THE
TWO-GROUP RANK DISTANCE CLASSIFICATION STATISTIC**

HARRY M. HUGHES, Ph. D.
RICHARD C. McNEE, M.S.

Approved for public release; distribution unlimited.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) USAF School of Aerospace Medicine Aerospace Medical Division (AFSC) Brooks Air Force Base, Texas 78235		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP	
3. REPORT TITLE AN ALGORITHM FOR CRITICAL VALUES OF THE TWO-GROUP RANK DISTANCE CLASSIFICATION STATISTIC			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) October 1971			
5. AUTHOR(S) (First name, middle initial, last name) Harry M. Hughes Richard C. McNee			
6. REPORT DATE December 1971		7a. TOTAL NO. OF PAGES 413	7b. NO. OF REFS 1
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S) SAM-TR-71-50	
b. PROJECT NO. 6319 c. Task No. 631901 d. Work Unit No. 631901012		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
10. DISTRIBUTION STATEMENT Approved for public release; distribution unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY USAF School of Aerospace Medicine Aerospace Medical Division (AFSC) Brooks Air Force Base, Texas 78235	
13. ABSTRACT The rank distance method is restated for discriminating between two groups using a single variable. A formula is derived for the probability of misclassifications in a training set of n observations from each group. Knowledge of this probability permits inference as to the appropriateness of the variable for discrimination purposes. The formula is exact only for values of U below a bound which is proportional to n. A table of numerical values indicates the formula to be useful up to about n = 25 but not beyond n = 40.			

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Classification theory						
Discriminant theory						
Rank distance						
Non-parametric						
Misclassification						
Small samples						
Training set						
Decision rule						
Order statistics						
Measure selection						
Ridit						

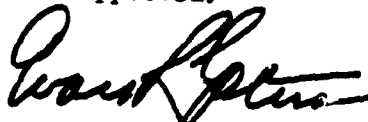
UNCLASSIFIED

Security Classification

FOREWORD

This work was done in the Biometrics Division under task No. 631901 during October 1971. The paper was submitted for publication on 9 November 1971.

This report has been reviewed and is approved.



EVAN R. GOLTRA, Colonel, USAF, MC
Commander

ABSTRACT

The rank distance method is restated for discriminating between two groups using a single variable. A formula is derived for the probability of misclassifications in a training set of n observations from each group. Knowledge of this probability permits inference as to the appropriateness of the variable for discrimination purposes. The formula is exact only for values of U below a bound which is proportional to n . A table of numeric values indicates the formula to be useful up to about $n = 25$ but not beyond $n = 40$.

/

AN ALGORITHM FOR CRITICAL VALUES OF THE TWO-GROUP RANK DISTANCE CLASSIFICATION STATISTIC

I. INTRODUCTION

The rank distance method of discriminating between two groups has been described in a previous report¹ for the case of two groups. It has the advantage of being applicable to a wide variety of measures including those whose scale may not be very well determined but whose ordering is well known. We addressed there the problem of judging how well a particular variable discriminates by noting the number of misclassifications of a training set using that single variable alone. The exact distribution of the number of misclassifications, under the hypothesis of both samples of n being from the same population, was computed and presented for values of n up to 12. Computation beyond that value would require an uneconomic amount of computer time.

We here restate the rank distance method, derive a formula for the lower tail of the misclassification distribution under the null hypothesis, and table some of its numeric values.

II. THE RANK DISTANCE METHOD

Consider a sample of n individuals from each of two defined groups. Let Y_{Ak} be the observation on the k^{th} individual known to be from group A. Let Y_{Bk} be the observation on the k^{th} individual known to be from group B. The Y_{Ak} and Y_{Bk} are ranked together in increasing order from 1 to $2n$. We then divide the rank assigned to Y_{jk} by $2n$ to obtain r_{jk} for $k = 1, 2, \dots, n$, and $j = A, B$. Thus far we have a training set of n fractions (called ridents by Bross) from each group. We next calculate the mean rident for each group, \bar{r}_j , and proceed to classify each observation as belonging to the group to whose mean rident it is closer. Finally, we tend to select those measures which produce a number of misclassifications in the training set whose cumulative probability is small under the null hypothesis of equal distributions within the groups.

III. DERIVATION OF FORMULA

The probability of obtaining exactly $2k$ misclassifications in the training set under the null hypothesis can be expressed by the formula

$$\text{Prob}(U = 2k) = 2 \left[\frac{n!}{k!(n-k)!} \right]^2 \frac{n!n!}{(2n)!}$$

¹Hughes, H. M., and R. C. McNee. Rank distance to choose discriminators for two groups. SAM-TR-71-40, Oct. 1971.

as long as k is smaller than a bound which we shall derive. To establish this formula, consider first the case in which $\bar{r}_A < \bar{r}_B$. Since the two group mean ridits average to the overall mean $\bar{r} = \frac{1}{2} + \frac{1}{4n}$, an individual ritid in group A will be misclassified if and only if it exceeds \bar{r} . There are n such possible ritid values. (Note that the case of an individual ritid equaling the overall mean cannot occur.) Out of the total

$$\frac{(2n)!}{n!n!}$$

equiprobable ways that n ritids may be selected from $2n$ for group A (with the remaining n obviously falling in group B), there are exactly

$$\frac{n!}{k!(n-k)!}$$

ways of selecting k ritids from the n values exceeding \bar{r} , and exactly

$$\frac{n!}{(n-k)!k!}$$

ways of selecting the $n-k$ ritids from the n values less than \bar{r} which then result in a correct classification. Thus there are

$$\left[\frac{n!}{k!(n-k)!} \right]^2$$

equiprobable ways that will result in exactly k misclassified observations in group A and corresponding k misclassified observations in group B, provided $\bar{r}_A < \bar{r}_B$. When the two group means are equal, half of the classifications are declared incorrect, so that $k = n/2$. As we shall see in the next section, the bound for k is less than $n/2$ so that this case need not be considered. In the remaining case of $\bar{r}_A > \bar{r}_B$, an exactly symmetric argument shows there are

$$\left[\frac{n!}{k!(n-k)!} \right]^2$$

equiprobable ways that will result in k ritids below \bar{r} and $n-k$ ritids above \bar{r} in group A. Putting the two cases together, we have the probability formula: a two for the two symmetric cases, times the number of ways of getting exactly $2k$ misclassifications in each case, times the reciprocal of the total possible ways.

IV. RANGE OF ACCURACY OF FORMULA

The foregoing derivation is valid, provided all of the ways counted actually fall into the case being considered. In particular, for the case $\bar{r}_A < \bar{r}_B$, we must assure that the k ridits larger than \bar{r} do not force \bar{r}_A greater than \bar{r} . The greatest value of \bar{r}_A is achieved when the k chosen ridits are the last k : $1 - \frac{k-1}{2n}$, $1 - \frac{k-2}{2n}$, ..., $1 - \frac{1}{2n}$, 1 and the $n-k$ chosen ridits are the largest ones below \bar{r} : $\frac{k+1}{2n}$, $\frac{k+2}{2n}$, ..., $\frac{1}{2}$. Each of these two sets is an arithmetic sequence; the first totals $k(2 - \frac{k-1}{2n})/2$ and the second totals $[\frac{1}{2} + \frac{k+1}{2n}] (n-k)/2$ so that our restriction is

$$\begin{aligned} \bar{r}_A &= \left[k\left(2 - \frac{k-1}{2n}\right)/2 + \left(\frac{1}{2} + \frac{k+1}{2n}\right) (n-k)/2 \right] /n \\ &= \left[k(4n-k+1) + (n+k+1) (n-k) \right] /4n^2 \\ &= \left[n^2 + n + 4nk - 2k^2 \right] /4n^2 \\ &< \bar{r} = (2n+1)/4n. \end{aligned}$$

Cross multiplying, the condition becomes

$$\begin{aligned} n^2 + n + 4nk - 2k^2 &< 2n^2 + n \\ n^2 &< 2n^2 - 4nk + 2k^2 \\ n^2 &< 2(n-k)^2 \\ n &< \sqrt{2} (n-k) \end{aligned}$$

since k will not exceed n . Thus the condition becomes

$$\begin{aligned} \sqrt{2}k &< (\sqrt{2}-1)n \\ k &< n (2 - \sqrt{2})/2 = .29289n \end{aligned}$$

to insure that our count is exact. The other case, $\bar{r}_A > \bar{r}_B$, reduces to the same condition when we take the k lowest ridits, the $n-k$ ridits that just exceed $1/2$, and require that the mean exceed $(2n+1)/4n$.

V. DISCUSSION

Numeric values of the formula have been computed and accumulated in table I. For any n from 7 to 40, this table presents the cumulative probability for the last three values of U before the bound that insures exact count for that n, and one value beyond. It also lists the bound. For example, the second entry opposite n = 11 was calculated as

$$2 \left[1 + 11^2 + 55^2 \right] \frac{11! 11!}{22!}$$

with at least 10 significant figures, then rounded for entry into the table.

For a sample of 11 from each of two populations, either zero or 2 misclassifications in the training set would be highly significant indication that the measure being used is not equally distributed in the two populations and hence is a likely discriminator. Four misclassifications in the training set would be significant at the 0.9% level, which would still appear to indicate a good discriminator, while 6 misclassifications of the 22 observations would be significant only at a level greater than 8.6%. The fact that the bound is 6.4 tells us that the probability listed is exact for 6 misclassifications or less.

The final probability entry on each line is not an exact value, but is listed because it appears to be accurate enough for the purposes of a critical value. For n = 9, the value .3469 approximates the exact value .3455; for n = 10, the approximate is exact to four decimal places. For n = 11, we have shifted to a different set of U values, but the comparison is still .3910 true and .3949 approximate. The approximate and exact probabilities for values of n from 5 through 12 are given in table II. It appears that the differences between the approximate and exact probabilities for values of n above 12 would not be large enough to be of any real importance in the selection of measures. Hence it would appear safe to use the fourth probability column of table I for those larger values of n where we do not know the exact value.

By use of table I we are able to determine 5% significant values up to about n = 20, 1% significant values up to about n = 30, and some idea of the smaller significant values up to n = 40. Because of the inequality restriction, this formula is of no avail for values larger than the ones just summarized.

TABLE I

$P = \text{Prob} \{ \text{number of misclassifications in training set} \leq U \}$ by formula

n	U	P	U	P	U	P	U	P	Bound on U
7	0	.0006	2	.0291	4	.2861	6	1.0000	4.1
8	0	.0002	2	.0101	4	.1319	6	.6193	4.7
9	0	.0000	2	.0034	4	.0567	6	.3469	5.3
10	0	.0000	2	.0011	4	.0230	6	.1789	5.9
11	2	.0003	4	.0089	6	.0861	8	.3949	6.4
12	2	.0001	4	.0033	6	.0391	8	.2203	7.0
13	2	.0000	4	.0012	6	.0169	8	.1152	7.6
14	4	.0004	6	.0070	8	.0570	10	.2568	8.2
15	4	.0001	6	.0028	8	.0268	10	.1431	8.8
16	4	.0000	6	.0011	8	.0121	10	.0756	9.4
17	6	.0004	8	.0053	10	.0381	12	.1694	10.0
18	6	.0002	8	.0022	10	.0184	12	.0943	10.5
19	6	.0001	8	.0009	10	.0086	12	.0502	11.1
20	6	.0000	8	.0004	10	.0038	12	.0256	11.7
21	8	.0001	10	.0017	12	.0126	14	.0629	12.3
22	8	.0001	10	.0007	12	.0060	14	.0336	12.9
23	8	.0000	10	.0003	12	.0028	14	.0174	13.5
24	10	.0001	12	.0012	14	.0087	16	.0422	14.1
25	10	.0000	12	.0005	14	.0042	16	.0227	14.6
26	10	.0000	12	.0002	14	.0020	16	.0118	15.2
27	10	.0000	12	.0001	14	.0009	16	.0060	15.8
28	12	.0000	14	.0004	16	.0029	18	.0154	16.4
29	12	.0000	14	.0002	16	.0014	18	.0081	17.0
30	12	.0000	14	.0001	16	.0007	18	.0041	17.6
31	14	.0000	16	.0003	18	.0021	20	.0105	18.2
32	14	.0000	16	.0001	18	.0010	20	.0055	18.7
33	14	.0000	16	.0001	18	.0005	20	.0028	19.3
34	14	.0000	16	.0000	18	.0002	20	.0014	19.9
35	16	.0000	18	.0001	20	.0007	22	.0038	20.5
36	16	.0000	18	.0000	20	.0003	22	.0020	21.1
37	16	.0000	18	.0000	20	.0002	22	.0010	21.7
38	18	.0000	20	.0001	22	.0005	24	.0026	22.3
39	18	.0000	20	.0000	22	.0002	24	.0014	22.8
40	18	.0000	20	.0000	22	.0001	24	.0007	23.4

TABLE II

Approximate and exact probabilities that the number of misclassifications in training set $\leq U$

n	U	Approx.	Exact	Difference	Bound on U
5	4	1.0000	.8730	.1270	2.9
6	4	.5671	.5498	.0173	3.5
7	6	1.0000	.8596	.1404	4.1
8	6	.6193	.5911	.0282	4.7
9	6	.3469	.3455	.0014	5.3
10	6	.1789	.1789	.0000	5.9
11	8	.3949	.3910	.0039	6.4
12	8	.2203	.2201	.0002	7.0