ON THE ANALYSIS OF PAIRED RANKED OBSERVATIONS

by

Leo Lynch, B.Sc., M.S.

Thesis submitted to the Graduate Faculty of the

Virginia Polytechnic Institute

in candidacy for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

APPROVED:

Chairman, Advisory Committee

August, 1957

Blacksburg, Virginia

TABLE OF CONTENTS

I. INTRODUCTION 4

1.1 A Problem in Nonparametric Statistics 4

II. A BIVARIATE RANK-SUM TEST. 6

2.1 Introduction. 6

2.2 Basic Considerations. 7

2.3 The S_1^2-Statistic 8

2.4 The First Two Moments of S_1^2 9

2.5 An Approximate Test of Significance 12

2.6 An Asymptatic Result. 21

2.7 A Rank Analogue of Wald's Modified T^2-Statistic 23

III. MULTIVARIATE AND MULTI-POPULATION EXTENSIONS 28

3.1 Introduction. 28

3.2 The $S_1^2(k,2)$-Statistic 28

3.3 The $S_1^2(2,p)$-Statistic 32

3.4 The $S_1^2(k,p)$-Statistic 33

IV. DISCRIMINANT ANALYSIS OF RANKED OBSERVATIONS 35

4.1 Introduction. 35

4.2 The Bivariate Case. 36

4.3 The Multivariate Extension. 37

4.4 The Wilcoxon Method 39

V. SUMMARY. 40a

5.1 Summary 40a

BIBLIOGRAPHY 41

APPENDIX A 43

A.1 Derivation of the Expectation of S_1^2 43

A.2 Derivation of the Variance of S_1^2 44

APPENDIX B 51

Table I. Values of a_{oo} 51

Table II. Values of a_{11} 53

Table III. Values of a_{12} 55

Table IV. Values of a_{22} 57

Table V. Values of $a_{11,11}$ 59

ACKNOWLEDGMENTS 61

VITA 62

I. INTRODUCTION

1.1 A Problem in Nonparametric Statistics

In current statistical research considerable attention is devoted to so-called nonparametric methods. Whenever there is a lack of information concerning the form of underlying distributions and there is the indication that assumptions of normality cannot be met, one must employ special kinds of techniques. Such techniques are known as nonparametric methods since they are not concerned with testing or estimating the parameters of distribution functions of given types. These methods are also called distribution-free methods because they do not require a knowledge of how the underlying random variables are distributed. The only assumption needed for most of these methods is that the distribution functions be continuous, a few of them requiring, furthermore, that low-order moments exist.

A statistical test that requires no assumption about the form of underlying distributions can hardly be expected to be as efficient as one requiring such assumption. To compensate for this loss in efficiency, nonparametric methods have the advantage of complete generality in application. Other noteworthy advantages of nonparametric methods of statistics are (i) computational ease and (ii) that they sometimes apply to data available only in ordinal form.

The development of nonparametric methods of statistics has been very rapid during the last decade, touching almost every phase of statistical activity. A common problem in practical statistics which has been attacked by nonparametric methods is that of deciding whether several sam-

ples should be regarded as coming from the same population.

References [1] through [5] of the biblography present some of the more general treatments of this problem. The purpose of this thesis is to consider an extension of the above problem to the case of 2 k-variate populations, particularly, two bivariate populations.

The first approach to the generalized problem is made by means of discriminant analysis and it is discussed in Chapter 4. Here the ranked sample points are projected onto a vector giving maximum discrimination between the two samples and the sample points are then reranked along this vector. The problem is thus reduced to a one-dimensional situation, but this approach is not a fruitful one since, as will be shown, the standard one-dimensional nonparametric tests cannot be used.

An alternative approach, discussed in Chapters 2 and 3, is based on the (Euclidean or more general) distance between the centroids of the ranked samples. Several methods are suggested for constructing approximate tests of significance on the basis of such a distance.

2. A BIVARIATE RANK-SUM TEST

2.1 Introduction

There are many problems of statistical inference in which one is unable to assume the functional form of population distributions. Many of these problems are such that the strongest assumption which can reasonably be made is continuity of the cumulative distribution functions of the populations. Problems of this type, in which the distribution functions are arbitrary within a broad class, come within the framework of nonparametric statistics as defined in Chapter 1.

The following problem belongs to the above class and it was originally suggested to the author by Doctor Frank Wilcoxon. Let π_1 and π_2 be two bivariate populations having unknown cumulative distribution functions $F_1(x_1,x_2)$ and $F_2(x_1,x_2)$, respectively. Assume that F_1 and F_2 are continuous, and identical except possibly in location parameter. It is desired to test the null hypothesis

$$H_o: F_1(x_1,x_2) \equiv F_2(x_1,x_2) \qquad (2.1.1)$$

against the alternative hypothesis that the population distribution functions have different means and it cannot be assumed that the variables x_1 and x_2 are statistically independent.

For example, it may be desired to compare two methods of preparing steel on the basis of compressive strength and elasticity, two teaching methods on the basis of grades obtained in two subjects by two groups of students, or two nationality groups on the basis of two specific skull measurements. If, in situations like these, it is unreasonable to assume

normality for the underlying distributions, one may resort to ranking techniques, ignoring the exact values of the measurements obtained.

In this chapter a bivariate rank-sum test is constructed to test the nonparametric statistical hypothesis (2.1.1) against the specified alternative hypothesis. The main value of such a bivariate rank-sum test is, as is characteristic of all nonparametric tests, that it is free from the assumption that the cumulative distribution functions of the populations have specific functional forms. Another advantage which is often important, is that nonparametric tests, based on ranking techniques, frequently provide computational ease not found in the corresponding parametric methods.

2.2 Basic Considerations

Suppose there are n_1 pairs of observations $(x_{11},x_{21}), \ldots, (x_{1n_1},x_{2n_1})$ from population π_1 and n_2 pairs of observations $(x_{1n_1+1},x_{2n_1+1}), \ldots, (x_{1N},x_{2N})$ from population π_2, where $N = n_1 + n_2$. The $x_{1i}(i=1,\ldots,N)$ are arranged in order of magnitude and ranked, the largest being assigned rank 1 and the smallest assigned rank N. In a similar manner, ranks are assigned to the observations x_{2i} $(i=1,\ldots,N)$. It is assumed that in either case, there are no ties in ranks.

Let u_{1i} and u_{2i} denote the ranks assigned to x_{1i} and x_{2i} if these observations belong to population π_1, and let u'_{1i} and u'_{2i} denote the ranks of these same observations if they belong to population π_2. It follows that

$$\sum_{k=1}^{n_1} u_{ik} + \sum_{k=n_1+1}^{N} u'_{ik} = \frac{N(N+1)}{2} \quad (i=1,2) \qquad (2.2.1)$$

where $\frac{N(N+1)}{2}$ is the sum of the first N integers.

If the $N = n_1 + n_2$ pairs of ranks are plotted on a plane, it is likely that the n_1 points from population π_1 and the n_2 points from population π_2 will be interspersed forming a circular or elliptical pattern under the assumption that $F_1(x_1,x_2)$ and $F_2(x_1,x_2)$ are identical. Under the alternative hypothesis, it is likely that there will be a segregation of the points into two groups. The S_1^2 -statistic that will be proposed in the next section to measure the extent of this segregation, is based on the Euclidean distance between the centroids of the two samples. Under the null hypothesis, S_1^2 can be expected to be smaller than under the alternative hypothesis.

2.3 The S_1^2 -Statistic

Using the Euclidean distance between the centroids of the ranks belonging to π_1 and π_2, the statistic S_1^2 is defined as:

$$S_1^2 = (\bar{u}_1 - \bar{u}_1')^2 + (\bar{u}_2 - \bar{u}_2')^2 \qquad (2.3.1)$$

where,

$$\bar{u}_i = n_1^{-1} \sum_{k=1}^{n_1} u_{ik} \quad \bar{u}_i' = n_2^{-1} \sum_{k=n_1+1}^{N} u_{ik}' \quad (i = 1,2) \qquad (2.3.2)$$

To simplify the notation in the more general case to be considered in Chapter 3, let

$$R_j = \sum_{k=1}^{n_1} u_{jk}$$

$$R_j' = \sum_{k=n_1+1}^{N} u_{jk}' \qquad (j=1,2) \qquad (2.3.3)$$

It follows from equation (2.2.1) that

$$R_j + R_j' = \frac{N(N+1)}{2} \qquad (j=1,2) \tag{2.3.4}$$

The right-hand side of equation (2.3.1) may then be rewritten as:

$$N^2[n_1n_2]^{-2}\left\{[n_1^{-1}R_1-n_2^{-1}(\frac{N(N+1)}{2}-R_1)]^2 + [n_1^{-1}R_2-n_2^{-1}(\frac{N(N+1)}{2}-R_2)]^2\right\} \tag{2.3.5}$$

and after performing some algebraic operations, the formula for the S_1^2-statistic reduces to

$$S_1^2 = N^2\,[n_1n_2]^{-2}\sum_{i=1}^{2}[R_i - \frac{n_1(N+1)}{2}]^2 \tag{2.3.6}$$

2.4 The First Two Moments of S_1^2

To construct a sampling distribution for the S_1^2 statistic, the following conditional randomization procedure will be used. Keeping the ranks paired as given in the sample, n_1 pairs are selected at random (with equal probabilities) from among the $N = n_1 + n_2$ pairs and assigned to population π_1; the remaining n_2 pairs are assigned to population π_2. Since this is a conditional randomization, no attempt will be made to obtain explicit results for the exact sampling distribution of the S_1^2-statistic. The first two moments will be derived and although higher moments could be obtained with identical techniques, their derivation would involve a prohibitive amount of algebraic complications.

To obtain the expectation of S_1^2, the following preliminary results

are required:

$$E(u_{ik}) = \frac{N+1}{2} \qquad (i=1,2) \tag{2.4.1}$$

$$E(R_i) = \frac{n_1(N+1)}{2} \qquad (i=1,2) \tag{2.4.2}$$

$$E(R_i^2) = \frac{n_1(N+1)\,[N(3n_1+1) + 2n_1]}{12} \tag{2.4.3}$$

The derivation of (2.4.3) is given in Appendix A.

Now,

$$E(S_1^2) = N^2\,[n_1 n_2]^{-2} \sum_{i=1}^{2} E[R_i - \frac{n_1(N+1)}{2}]^2 \tag{2.4.4}$$

and after substituting the results of (2.4.2) and (2.4.3) this becomes:

$$E(S_1^2) = \frac{N^2(N+1)}{6n_1n_2} \tag{2.4.5}$$

In the special case where $n_1 = n_2 = n$, the expectation of S_1^2 reduces to

$$E(S_1^2) = \frac{2}{3}(2n+1) \tag{2.4.6}$$

To obtain the variance of S_1^2, it will be necessary to evaluate $E(S_1^4)$ and substitute the result together with (2.4.5) into

$$\sigma^2_{S_1^2} = E(S_1^4) - [E(S_1^2)]^2 \tag{2.4.7}$$

Using the expressions obtained in Appendix A, it can be shown that the formula for the variance of S_1^2 can be written in the form

$$\sigma^2_{S_1^2} = a_{oo} + a_{11}A_{11} + a_{12}A_{12} + a_{21}A_{21} + a_{22}A_{22} + a_{11,11}A_{11}^2 \tag{2.4.8}$$

where

$$a_{oo} = \frac{N^4(N+1)}{90n_1^3n_2^4(N-1)\ (N-2)\ (N-3)}\ [N^6(25n_1-29)$$

$$+ N^5(-50n_1^2 + 113n_1 - 65)$$

$$+ N^4(25n_1^3-168n_1^2+ 286n_1-75)$$

$$+ N^3(84n_1^3-442n_1^2+ 369n_1-25)$$

$$+ N^2(221n_1^3-588n_1^2+ 121n_1+ 14)$$

$$+ N(204n_1^3-192n_1^2-14n_1) + 96n_1^3] \qquad (2.4.9)$$

$$a_{11} = \frac{2N^3(N+1)^2\ [\ (n_1-n_2)^2 - n_1^2\ (n_2-1)-n_2^2(n_1-1)]}{n_1^3n_2^3(N-1)\ (N-2)\ (N-3)} \qquad (2.4.10)$$

$$a_{12} = a_{21} = \frac{2N^3(N+1)\ \ [4n_1n_2-n_1(n_1+1)-n_2(n_2+1)]}{n_1^3n_2^3(N-1)\ (N-2)\ (N-3)} \qquad (2.4.11)$$

$$a_{22} = \frac{2n^3\ [(N-2)\ (N-3)-6(n_1-1)\ (n_2-1)]}{n_1^3n_2^3(N-1)\ (N-2)\ (N-3)} \qquad (2.4.12)$$

$$a_{11,11} = \frac{4N^3(n_1-1)\ (n_2-1)}{n_1^3n_2^3(N-1)\ (N-2)\ (N-3)} \qquad (2.4.13)$$

$$A_{11} = \sum_{i=1}^{N} u_{1i}\ u_{2i} \qquad (2.4.14)$$

$$A_{12} = \sum_{i=1}^{N} u_{1i}\ u_{2i}^2 \qquad (2.4.15)$$

$$A_{21} = \sum_{i=1}^{N} u_{1i}^2\ u_{2i} \qquad (2.4.16)$$

$$A_{22} = \sum_{i=1}^{N} u_{1i}^2\ u_{2i}^2 \qquad (2.4.17)$$

It should be noted that the variance of S_1^2 depends on the actual matching of the u_1's and the u_2's only in so far as it depends on the parameters A_{11}, A_{12}, A_{21} and A_{22}. Also it should be recalled that the expression obtained in (2.4.8) is the variance of the theoretical sampling distribution of S_1^2 under the conditional randomization described above.

In the special case where $n_1 = n_2 = n$, the constants a_{ij} in (2.4.8) become

$$a_{oo} = \frac{16(2n+1)\ (50n^3+43n^2+20n+7)}{45n(2n-3)} \tag{2.4.18}$$

$$a_{11} = \frac{-16(2n+1)^2}{n(2n-1)\ (2n-3)} \tag{2.4.19}$$

$$a_{12} = a_{21} = \frac{16(2n+1)}{n^2(2n-1)\ (2n-3)} \tag{2.4.20}$$

$$a_{22} = \frac{-16}{n^2(2n-1)\ (2n-3)} \tag{2.4.21}$$

$$a_{11,11} = \frac{16(n-1)}{n^3(2n-1)\ (2n-3)} \tag{2.4.22}$$

To facilitate the determination of the variance of S_1^2, the constants a_{oo}, a_{11}, a_{12}, a_{22} and $a_{11,11}$ have been calculated for all values of n_1 and n_2 up to n_1 and n_2 equal to 20. These values may be found in Appendix B.

2.5 An approximate Test of Significance

To perform an exact test of hypothesis (2.1.1) against the al-

ternative hypothesis (2.1.2) on the basis of the S_1^2-statistic, it would be necessary to obtain the theoretical sampling distribution of S_1^2. Such a distribution could be derived either by obtaining explicit expressions for the probabilities involved or by enumerating all possible cases. Neither of these two approaches seem to be feasible, the first being complicated by the conditional randomization and the second being impractical since even for n_1 and n_2 as small as 10 it would be necessary to enumerate $\binom{20}{10} = 184,756$ cases.

Workers in the field of nonparametric statistics have encountered considerable difficulties in their attempts of obtaining explicit expressions for sampling distributions of their statistics. Results are scarce and unwieldy even in the case of ordinary unrestricted randomization and no attempt will be made to treat the exact sampling distribution of S_1^2 in a theoretical fashion.

Statisticians frequently approximate sampling distributions of pertinent statistics, at least for large samples, with normal distributions, justifying this either on theoretical grounds or with emperical means. The advantage of this procedure is that knowledge of the mean and variance of the actual sampling distribution of the statistics is sufficient to perform tests of significance.

To see whether a normal curve would provide a satisfactory approximation to the sampling distribution of S_1^2, one could evaluate the third and fourth moments of this statistic (under the previously discussed conditional randomization) and check whether α_3 and α_4 are reasonably close to 0 and 3, respectively. Since the determination of these moments would

involve an enormous amount of algebra, it was decided to use other means, obtaining one sampling distribution by complete enumeration and another by emperical means.

Example 1

In this example the exact sampling distribution of S_1^2 will be obtained by enumeration for the special case where $n_1 = n_2 = 5$ and where the u_1's and u_2's are paired in the following fashion.

u_1	1	2	3	4	5	6	7	8	9	10
u_2	8	3	10	7	6	5	4	9	1	2

In this arrangement the correlation between the u_1's and u_2's is $\rho = -0.50$ and this dependence will be preserved in the conditional randomization in which 5 of the above pairs will be assigned at random (with equal probabilities) to population π_1 and the remaining pairs assigned to population π_2.

Since there are $\binom{10}{5} = 252$ ways in which 5 of the pairs can be assigned to pupulation π_1, the exact sampling distribution of S_1^2 is obtained in this example by actually enumerating these cases and calculating the corresponding values of S_1^2. The result is shown, grouped, in the histogram of Figure 1.

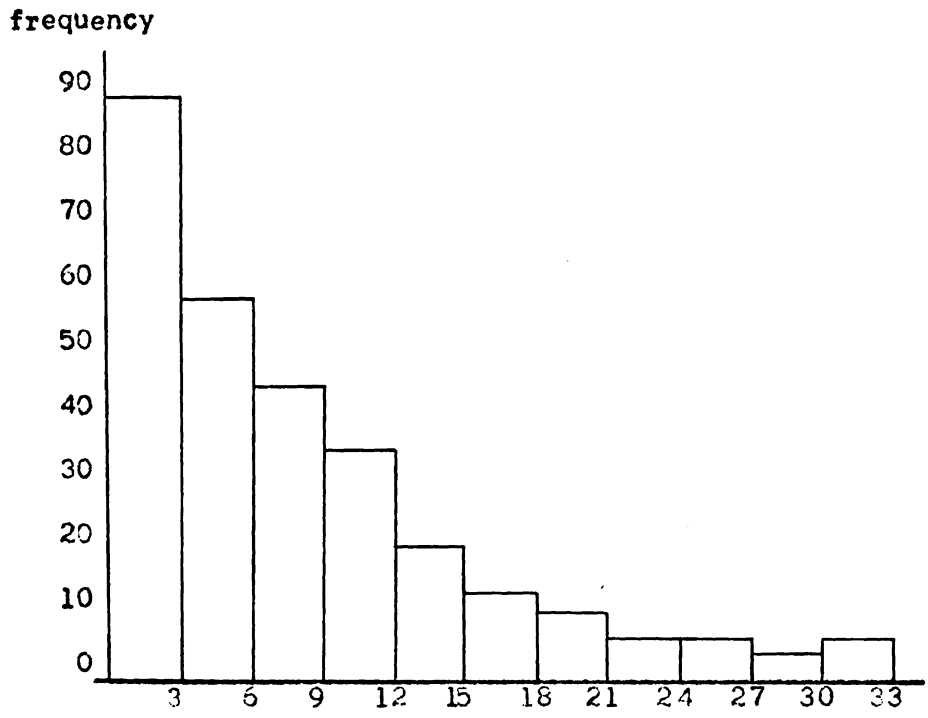

Figure 1. Distribution of S_1^2 for $n_1 = n_2 = 5$.

It is apparent from Figure 1 that, for $n_1 = n_2 = 5$ and the given matching, the sampling distribution of S_1^2 is highly skewed and it would hardly seem reasonable to approximate it with a normal curve. In order to investigate whether the sampling distribution of S_1^2 might be closer to a normal curve when n_1 and n_2 are larger than 5, a second example was worked out.

Example 2

In this example the sampling distribution of S_1^2 will be investigated for $n_1 = n_2 = 10$ with the following matching of the u_1's and u_2's:

u_1	1	2	3	4	5	6	7	8	9	10
u_2	2	11	1	12	4	14	5	7	3	13

u_1	11	12	13	14	15	16	17	18	19	20
u_2	9	8	18	6	19	20	10	16	17	15

In this arrangement the correlation between the u_1's and u_2's is $\rho = 0.63$ and this dependence will be preserved in the conditional randomization in which 10 of the above pairs will be assigned at random (with equal probabilities) to population π_1 and the remaining pairs assigned to population π_2.

Since there are $\binom{20}{10} = 184{,}756$ cases to be enumerated, it was decided to use a Monte Carlo method instead of a complete enumeration. One hundred random samples yielded the values of S_1^2 shown in the distribution of Figure 2.

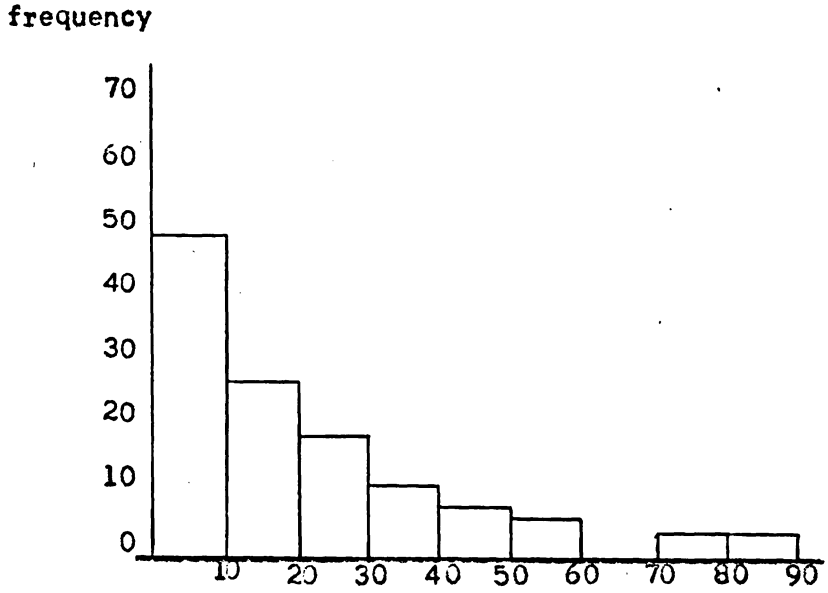

Figure 2. Distribution of S_1^2 for $n_1 = n_2 = 10$

It is clear from Figure 2 that in this second example the sampling distribution of S_1^2 is again highly skewed and that a normal curve approximation would be quite unreasonable.

It was suggested that it might, perhaps, be fruitful to adjust S_1^2 (multiplying by a constant) so that its range of values will be limited to the interval from 0 to 1 and then approximate its sampling distribution with a Beta-distribution. The difficulty posed by this approach is that the maximum value of S_1^2, by which one would have to divide, depends on the matching of the u_1's and u_2's and, hence, cannot be given in a general form. Of course, the maximum value of S_1^2 could be found in any given example, but this would generally entail an enormous amount of work. It is for this reason that the Beta-approximation will not be used.

Since S_1^2 is by definition limited to positive real values, it would perhaps seem reasonable to approximate its sampling distribution with a gamma distribution whose parameters could be obtained by equating suitable expressions involving these parameters with $E(S_1^2)$ and $\sigma^2_{S_1^2}$. In view of the fact that this procedure would entail considerable computations and that, furthermore, tables of gamma distributions are not readily available to most research workers, it will be suggested to use a special kind of gamma distribution, namely, the χ^2 distribution. A heuristic justification for the use of such an approximation lies in the fact that S_1^2 is the sum of two squares which, individually, are linear functions of rank sums.

In order to approximate the sampling distribution of S_1^2 with a

χ^2 distribution, it will be necessary to adjust S_1^2, namely, to multiply it by a constant k, such that the variance of kS_1^2 equals twice the expectation of kS_1^2. Since the variance of kS_1^2 is $k^2\sigma^2_{S_1^2}$ and the expectation of kS_1^2 is $kE(S_1^2)$, the above condition yields

$$\frac{k^2\sigma^2_{S_1^2}}{kE(S_1^2)} = 2 \qquad (2.5.1)$$

or

$$k = \frac{2E(S_1^2)}{\sigma^2_{S_1^2}} \qquad (2.5.2)$$

Since the number of degrees of freedom of a χ^2 distribution equals its expectation, it will be proposed to approximate the sampling distribution of the modified statistic

$$kS_1^2 = \frac{2E(S_1^2)}{\sigma^2_{S_1^2}} S_1^2 \qquad (2.5.3)$$

by means of a χ^2 distribution having

$$kE(S_1^2) = \frac{2\,[E(S_1^2)]^2}{\sigma^2_{S_1^2}} \qquad (2.5.4)$$

degrees of freedom. In order to use a χ^2 table in testing the original hypothesis (2.1.1) against the specified alternative hypothesis, one can either be conservative and use the smallest integer greater than or equal to $\frac{2\,[E(S_1^2)]^2}{\sigma^2_{S_1^2}}$ for the number of degrees of freedom, or interpolate linearly for fractional degrees of freedom [9].

In example 1, $n_1 = n_2 = 5$ and, therefore, from (2.4.6), $E(S_1^2) = \frac{2}{3}(2n+1) = 7\frac{1}{3}$. Using the matching of the u_1's and the u_2's of this example, it can be shown that

$$A_{11} = 261, \quad A_{21} = 1605, \quad A_{12} = 1717, \quad A_{22} = 10033$$

Since the tables of appendix B yield

$$a_{oo} = 830.496 \qquad a_{12} = -.112$$

$$a_{11} = -6.146 \qquad a_{22} = -0.010$$

$$a_{11,11} = 0.008$$

substitution into (2.4.8) gives $\sigma^2_{S_1^2} = 43.1$. According to the approximation discussed above,

$$kS_1^2 = \frac{2E(S_1^2)}{\sigma^2_{S_1^2}} \cdot S_1^2 = 0.34S_1^2$$

is treated as if it had a χ^2 distribution with

$$\frac{2\,[\,E(S_1^2)]^2}{\sigma^2_{S_1^2}} = 2.49 \text{ degrees of freedom.}$$

Using the smallest integer greater than or equal to 2.49, namely 3, the χ^2 table yields, for a level of significance of 0.05, a critical value of $\chi^2_{.05} = 7.81$. Hence, if the approximation were good, 5 per cent of the values of $0.34S_1^2$ should exceed 7.81 or, in other words, 5 per cent of the values of S_1^2 should exceed $\frac{7.81}{0.34} = 22.97$.

In the actual enumeration of the 252 cases it was found that 14 exceed 22.97 or 5.6 per cent. This implies that in Example 1 the χ^2 criterion outlined above provides a very close approximation for a 0.05 test of significance.

In example 2, $n_1 = n_2 = 10$ and, therefore, from (2.4.6) $E(S_1^2) = \frac{2}{3}(2n+1) = 14$. On the basis of the matching of the u_1's and the u_2's of this example, it can be shown that

$$A_{11} = 2626 \qquad A_{12} = 45032$$

$$A_{21} = 40528 \qquad A_{22} = 602{,}874$$

Since the tables of Appendix B yield

$$a_{oo} = 2394.03294 \qquad a_{12} = 45032$$

$$a_{11} = -2.184520 \qquad a_{22} = -0.000495$$

$$a_{11,11} = 0.000446$$

substitution into (2.4.8) gives $\sigma^2_{S_1^2} = 323.2$. Hence, the sampling distribution of

$$kS_1^2 = \frac{2(14)}{323.2} \quad S_1^2 = 0.0866\, S_1^2$$

will be approximated with a χ^2 distribution having $\frac{2[14]^2}{323.2} = 1.21$ degrees of freedom. Using the smallest integer greater than or equal to 1.21, namely 2, the χ^2 tables yield, for a level of significance of 0.05, a critical value of $\chi^2_{.05} = 5.99$. Hence, if the approximation were

good, 5 per cent of the values of $0.0866S_1^2$ should exceed 5.99 or, in other words, 5 per cent of the values of S_1^2 should exceed $\frac{5.99}{0.0866} = 69.17$. In the experimental sampling distribution obtained in this example, 4 of the 100 values of S_1^2 or 4 per cent exceeded 69.17. In view of the fact that this experimental distribution was based only on 100 samples, the results are not conclusive, but they certainly provide further support for the χ^2 approximation of the sampling distribution. (Had the interpolated χ^2-value been used in this example, the percentage of values exceeding the critical value would have been 8 per cent.)

2.6 An Asymptotic Result

The purpose of this section is to evaluate the number of degrees of freedom given by (2.5.4) when $n = n_1 = n_2$ becomes large and when the dependence between the u_1's and u_2's is such that

$$\frac{A_{11}}{n^3} = 2 + O(\frac{1}{n}) \qquad (2.6.1)$$

$$\frac{A_{12}}{n^4} = \frac{8}{3} + O(\frac{1}{n}) \qquad (2.6.2)$$

$$\frac{A_{21}}{n^4} = \frac{8}{3} + O(\frac{1}{n}) \qquad (2.6.3)$$

$$\frac{A_{22}}{n^5} = \frac{32}{9} + O(\frac{1}{n}) \qquad (2.6.4)$$

These four equations correspond to the case when there is a very weak correlation between the u_1's and u_2's, in fact, they were obtained by

substituting $2nE(u_1^r)E(u_2^s)$ for A_{rs} and then letting n become large.

Using (2.4.6), it follows immediately that when n becomes infinite, the $\frac{E(S_1^2)}{n}$ approaches $\frac{4}{3}$. Also, using (2.4.18), (2.4.19), (2.4.20), (2.4.21) and (2.4.22) together with the above conditions on the A_{rs}, it can be shown that

$$\frac{a_{00}}{n^2} = \frac{160}{9} + O(\frac{1}{n}) \tag{2.6.5}$$

$$\frac{a_{11}A_{11}}{n^2} = -32 + O(\frac{1}{n}) \tag{2.6.6}$$

$$\frac{a_{12}A_{12}}{n^2} = \frac{a_{21}A_{21}}{n^2} = \frac{64}{3n} + O(\frac{1}{n^2}) \tag{2.6.7}$$

$$\frac{a_{22}A_{22}}{n^2} = \frac{128}{9n} + O(\frac{1}{n^2}) \tag{2.6.8}$$

$$\frac{a_{11,11}A_{11}^2}{n^2} = 16 + O(\frac{1}{n}) \tag{2.6.9}$$

Substituting all these values into (2.4.8) yields in the limit

$$\lim_{n\to\infty} \frac{\sigma^2_{S_1^2}}{n^2} = \frac{16}{9} \tag{2.6.10}$$

and the formula for the number of degrees of freedom becomes

$$\lim_{n\to\infty} \frac{2\,[E(S_1^2)]^2}{\sigma^2_{S_1^2}} = \lim_{n\to\infty} \frac{2\,[E(S_1^2)]^2}{\frac{\sigma^2_{S_1^2}}{n^2}} = \frac{2(\frac{4}{3})^2}{\frac{16}{9}} \tag{2.6.11}$$

In the special case discussed above, (2.6.5), (2.6.6) and (2.6.9) equal $\frac{160}{9}$, -32 and 16, respectively, in the limit, while (2.6.7) and (2.6.8) go to zero. This implies here that among the A_{rs}, the variance of S_1^2 will depend mainly on A_{11}. This parameter is functionally related to the correlation coefficient of the u_1's and u_2's and one can write

$$\rho = \frac{6A_{11} - 3n(2n+1)^2}{n(2n+1)(2n-1)} \tag{2.6.12}$$

It follows that (2.6.1) is equivalent to

$$\rho = O(\frac{1}{n}) \tag{2.6.13}$$

or, in other words, that the above argument applies only when the correlation between the u_1's and u_2's is very weak.

2.7 A Rank Analogue of Wald's Modified T^2-Statistic

The type of problem treated in this chapter can, in the parametric case, be handled by the Hotelling T^2-statistic provided, of course, that the assumption of normality and equal variance-covariance matrices can be met.

To construct an analogue to the T^2-statistic one could use ranks instead of the actual observations, obtaining an alternative to the S_1^2-statistic suggested in this chapter. For reasons of simplicity, Wald's [8] modification of T^2 namely,

$$T'^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} q'_{ij}(\bar{x}_i - \bar{x}'_i)(\bar{x}_j - \bar{x}'_j) \tag{2.7.1}$$

will be used, where

$$\left\| q'_{ij} \right\| = \left\| c'_{ij} \right\|^{-1} \tag{2.7.2}$$

$$c'_{ij} = \frac{N}{(N-1)n_1 n_2} \sum_{k=1}^{N} (x_{ik}-\bar{X}_i)(x_{jk}-\bar{X}_j) \quad (i,j=1,2) \tag{2.7.3}$$

$$\bar{X}_i = N^{-1} \sum_{k=1}^{N} x_{ik} \tag{2.7.4}$$

and it has been shown that T'^2 is a monotonic function of T^2. Analogous to (2.7.1), one can now use ranks and write

$$S_2^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} \lambda_{ij}(\bar{u}_i-\bar{u}'_i)(\bar{u}_j-\bar{u}'_j) \tag{2.7.5}$$

where,

$$\left\| \lambda_{ij} \right\| = \left\| c_{ij} \right\|^{-1} \tag{2.7.6}$$

$$c_{11} = \frac{N}{(N-1)n_1 n_2} \sum_{k=1}^{N} (u_{1k}-\bar{U}_1)^2 = \frac{N^2(N+1)}{12n_1 n_2} \tag{2.7.7}$$

$$c_{22} = \frac{N}{(N-1)n_1 n_2} \sum_{k=1}^{N} (u_{2k}-\bar{U}_2)^2 = \frac{N^2(N+1)}{12n_1 n_2} \tag{2.7.8}$$

$$c_{12} = \frac{N}{(N-1)n_1 n_2} \sum_{k=1}^{N} (u_{1k}-\bar{U}_1)(u_{2k}-\bar{U}_2)$$

$$= \frac{N}{(N-1)n_1 n_2} \left[A_{11} - \frac{N(N+1)^2}{2}\right] \tag{2.7.9}$$

Putting,

$$B_{11} = A_{11} - \frac{N(N+1)^2}{2} \tag{2.7.10}$$

equation (2.7.9) can be written

$$c_{12} = \frac{N}{(N-1)n_1 n_2} B_{11} \tag{2.7.11}$$

Now,

$$\Lambda = C^{-1} = \frac{N^2(N^2-1)^2}{N^2(N^2-1)^2 - 144B_{11}^2} \begin{Vmatrix} 1 & \frac{-12B_{11}}{N(N^2-1)} \\ \frac{-12B_{11}}{N(N^2-1)} & 1 \end{Vmatrix} \tag{2.7.12}$$

and omitting the multiplicative constant, let

$$\Lambda^* = \begin{Vmatrix} 1 & \frac{-12B_{11}}{N(N^2-1)} \\ \frac{-12B_{11}}{N(N^2-1)} & 1 \end{Vmatrix} \tag{2.7.13}$$

Using this matrix instead of Λ , one can now define a third statistic

$$S_3^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} \lambda^*_{ij} (\bar{u}_i - \bar{u}'_i)(\bar{u}_j - \bar{u}'_j) \tag{2.7.14}$$

so that

$$S_3^2 = \frac{N^2(N^2-1)^2 - 144B_{11}^2}{N^2(N^2-1)^2} S_2^2 \tag{2.7.15}$$

It should be noted that if in (2.7.14) the cross-product term is omitted, the formula reduces to that of S_1^2 .

The mean and variance of S_3^2 may be found by the use of methods like those employed in Appendix A. The results are

$$E(S_3^2) = \frac{N^2(N+1)}{6n_1 n_2} - \frac{243B_{11}^2}{(N+1)(N-1)^2 n_1 n_2} \tag{2.7.16}$$

and

$$\sigma^2_{S_1^2} = b_{oo} + b_{11}A_{11} + b_{12}A_{12} + b_{21}A_{21} + b_{22}A_{22} + b_{11,11}A_{11}^2 \qquad (2.7.17)$$

where A_{11}, A_{12}, A_{21} and A_{22} are the parameters defined in (2.4.14), (2.4.15), (2.4.16) and (2.4.17), B_{11} is defined in (2.7.10), and the coefficients b_{oo}, b_{11}, b_{12}, b_{22}, and $b_{11,11}$ are functions of n_1, n_2 and B_{11}. Explicitly,

$$\begin{aligned}
b_{oo} = a_{oo} &- \frac{576B_{11}^4}{(N+1)(N-1)^4n_1^2n_2^2} + \frac{8N^2B_{11}^2}{(N^2-1)^2n_1^2n_2^2} \\
&+ \frac{576B_{11}^2}{N^2(N^2-1)^2}\left\{\frac{n_1^3(N+1)^3[2N-n_1(3N+5)]}{48}\right. \\
&+ \frac{n_1(n_1-1)N(N+1)^2[9n_1^2(N+1)^2-6n_1(N+1)(2N+1)+(2N+1)^2]}{36(N-1)} \\
&+ \frac{n_1(n_1-1)(n_1-2)N(N+1)^2(3N+2)[(2N+1)-3n_1(N+1)]}{36(N-2)} \\
&\left.+ \frac{n_1(n_1-1)(n_1-2)(n_1-3)N(N+1)^2[9N^2(N+1)^2-4(2N+1)(3N^2+N-1)]}{144(N-1)(N-2)(N-3)}\right\} \\
&- \frac{48B_{11}}{N(N^2-1)}\left\{\frac{n_1(n_1-1)N(N+1)^2[N(3n_1^2-4n_1+2)+n_1(3n_1-2)]}{8(N-1)}\right. \\
&+ n_1(n_1-1)(n_1-2)N(N+1)^3[4-n_1(3N+2)] \\
&+ \frac{n_1(n_1-1)(n_1-2)(n_1-3)N^2(N+1)^2}{24(N-3)} \\
&\left.+ \frac{n_1^2(N+1)^3[N(13n_1^2-18n_1+6)+(5n_1^2-2n_1-4)]}{16}\right\} \qquad (2.7.18)
\end{aligned}$$

and

$$b_{11} = ta_{11} \tag{2.7.19}$$

$$b_{12} = b_{21} = ta_{12} \tag{2.7.20}$$

$$b_{22} = ta_{22} \tag{2.7.21}$$

$$b_{11,11} = ta_{11,11} \tag{2.7.22}$$

where,

$$t = 1 + \frac{288B_{11}^2}{N^2(N^2-1)^2} \tag{2.7.23}$$

Although the S_3^2-statistic may provide a slight advantage over the S_1^2-statistic from the viewpoint of power, the S_1^2-statistic has the very definite advantage that its first two moments are much easier to find. It is for this reason that the S_1^2-statistic and the χ^2 approximation are suggested as a criterion for testing hypothesis (2.1.1) against the alternative (2.1.2). In view of the results of Section 2.6, it is of interest to note that it has been shown that the distribution of T'^2, for which S_2^2 and S_3^2 are rank analogues, is asymptotically χ^2 with 2 degrees of freedom.

3. MULTIVARIATE AND MULTI-POPULATION EXTENSIONS

3.1 Introduction

In this chapter the method of Section 2.3, the S_1^2-statistic, will be generalized first to the multivariate case for two populations. This problem is considered in Section 3.2, where a statistic, $S_1^2(k,2)$ is defined analogous to S_1^2, as the square of the Euclidean distance in the k-dimensional rank space. (Note that $S_1^2(2,2) = S_1^2$).

In Section 3.3, the method of Section 2.3 is extended to the case where there are p bivariate populations. The test statistic proposed here, $S_1^2(2,p)$, is the sum of the squares of the Euclidean distances between all pairs of centroids in the bivariate rank space. The extension to p k-variate populations is discussed briefly in Section 3.4.

3.2 The $S_1^2(k,2)$ - Statistic

Suppose there are n_1 k-tuplets of observations $(x_{1i},x_{2i},..,x_{ki})$, $i=1,2,...,$ and n_1, from population π_1, and n_2 k-tuplets of observations $(x_{1i},x_{2i},...,x_{ki})$, $i=n_1+1,...,$ $n_1+n_2=N$, from population π_2. Suppose, furthermore, that the N values x_{1i} are ranked jointly, receiving ranks that are written symbolically as u and u' depending on whether they belong to population π_1 or π_2. The observations x_{2i}, x_{3i},..., and x_{ki} are ranked in a similar fashion and, in general, u_{ji} stands for the rank of an observation on the j^{th} variable if it belongs to population π_1, while u'_{ji} stands for the rank of an observation on the j^{th} variable if it belongs to population π_2.

Using the means $\bar{u}_j$ and $\bar{u}'_j$ as defined in (2.3.2), the Euclidean distance between $(\bar{u}_1,\bar{u}_2,\ldots,\bar{u}_k)$ and $(\bar{u}'_1,\bar{u}'_2,\ldots,\bar{u}'_j)$ is

$$S_1^2(k,2) = \sum_{j=1}^{k} (\bar{u}_j - \bar{u}'_j)^2 \qquad (3.2.1)$$

Using (2.3.3), this squared distance can be rewritten in the form

$$S_1^2(k,2) = \frac{N^2}{n_1^2 n_2^2} \sum_{j=1}^{k} \left[R_j - \frac{n_1(N+1)}{2}\right]^2 \qquad (3.2.2)$$

analogous to (2.3.6).

The derivation of the first two moments of $S_1^2(k,2)$ is essentially the same, though somewhat more tedious, as the derivation of these moments for S_1^2.

It can easily be shown that

$$E\left[S_1^2(k,2)\right] = \frac{kN^2(N+1)}{12n_1n_2} \qquad (3.2.3)$$

and that for the special case where $n_1=n_2=n$, this formula reduces to

$$E\left[S_1^2(k,2)\right] = \frac{k}{3}(2n+1) \qquad (3.2.4)$$

The variance of $S_1^2(k,2)$ is of the form

$$\begin{aligned}\sigma^2_{S_1^2}(k,2) = {} & c_{oo} + c_{11}[A_{11,12} + A_{11,13} + \cdots + A_{11,(k-1)k}] \\ & + c_{12}[A_{12,12} + A_{12,13} + \cdots + A_{12,(k-1)k}] \\ & + c_{21}[A_{21,12} + A_{21,13} + \cdots + A_{21,(k-1)k}] \\ & + c_{11,11}[A^2_{11,12} + A^2_{11,13} + \cdots + A^2_{11,(k-1)k}] \\ & + c_{22}[A_{22,12} + A_{22,13} + \cdots + A_{22,(k-1)k}]\end{aligned} \qquad (3.2.5)$$

where,

$$c_{oo} = \frac{kN^4(N+1)}{240n_1^3n_2^4}\ [N^3(5n_1-2) + N^2(-10n_1^2 + 9n_1-2)$$

$$+ N(5n_1^3 - 14n_1^2 + 2n_1) + 7n_1^3]$$

$$+ \frac{k(k-1)N^4(N+1)^2}{144(N-1)\ (N-2)\ (N-3)n_1^3n_2^4}\ [N^5(19n_1-22)$$

$$+ N^4(-38n_1^2 + 74n_1-36) + N^3(19n_1^3-104n_1^2+ 163n_1-18)$$

$$+ N^2(52n_1^3 - 254n_1^2 + 108n_1 + 4) + 4N(127n_1^3-180n_1^2-4n_1)+90n^3]$$

$$- \frac{k^2N^4(n+1)^2}{144n_1^2n_2^2} \qquad (3.2.6)$$

and c_{11}, c_{12}, c_{22} and $c_{11,11}$ are, respectively, the came as a_{11}, a_{12}, a_{22} and $a_{11,11}$ as defined in (2.4.10), (2.4.11), (2.4.12) and (2.4.13). The parameters $A_{11,ij}$, $A_{12,ij}$, $A_{21,ij}$ are as defined in (2.4.14) through (2.4.17) with the second pair of subscripts referring to the variables. For example,

$$A_{11,12} = \sum_{i=1}^{N} u_{1i}u_{2i} \qquad (3.2.7)$$

$$A_{11,13} = \sum_{i=1}^{N} u_{1i}u_{3i} \qquad (3.2.8)$$

$$A_{12,23} = \sum_{i=1}^{N} u_{2i}u_{3i}^2 \qquad (3.2.9)$$

In the special case where $n_1 = n_2 = n$, the coefficients c_{11}, c_{12}, c_{22} and $c_{11,11}$ of (3.2.6) are the same as in (2.4.19) through (2.4.22) and c_{oo} reduces to the following:

$$c_{oo} = \frac{k(2n+1)\,(10n^2-n-4)}{15n} - \frac{k^2(2n+1)^2}{9}$$

$$+ \frac{k(k-1)\,(2n+1)^2(38n^2+21n+4)}{9n(2n-3)} \qquad (3.2.10)$$

Similar to Section 2.1, the hypothesis to be tested is

$$F_1(x_1,x_2,\ldots,x_k) \equiv F_2(x_1,x_2,\ldots,x_k)$$

and the alternative hypothesis is

$$F_1(x_1,x_2,\ldots,x_k) \not\equiv F_2(x_1,x_2,\ldots,x_k)$$

where F_1 and F_2 are continuous distribution functions, identical except possibly in location parameters. It is suggested that an approximate test of the hypothesis be performed by approximating the sampling distribution of $S_1^2(k,2)$ with a χ^2 distribution with the number of degrees of freedom being the smallest integer greater than $\frac{2[ES_1^2(k,2)]^2}{\sigma^2_{S_1^2}(k,2)}$.

When the pairwise dependence between the u_i's and u_j's is weak and the parameters are as defined in (2.6.1) through (2.6.4), it follows that when $n = n_1 = n_2$ becomes large, $\frac{E[S_1^2(k,2)]}{n}$ approaches $\frac{2k}{3}$.

Also, the coefficients c_{11}, c_{12}, c_{22} and $c_{11,11}$ become equal to the expressions given in (2.6.6) through (2.6.9) and (3.2.10) becomes

$$\frac{c_{oo}}{n^2} = \frac{72k^2-64k}{9} + O\left(\frac{1}{n}\right) \qquad (3.2.11)$$

Substituting these values into (3.2.5) yields in the limit

$$\lim_{n\to\infty} \frac{\sigma^2_{S_1^2}(k,2)}{n^2} = \frac{8k}{9} \qquad (3.2.12)$$

and the formula for the number of degrees of freedom becomes

$$\lim_{n\to\infty} \frac{2[ES_1^2(k,2)]^2}{\sigma^2_{S_1^2}(k,2)} = \lim_{n\to\infty} \frac{2\left[\frac{E[S_1^2(k,2)]}{n}\right]^2}{\frac{\sigma^2_{S_1^2}(k,2)}{n^2}}$$

$$= \frac{2\left(\frac{2k}{3}\right)^2}{\frac{8k}{9}} = k \qquad (3.2.13)$$

3.3 The $S_1^2(2,p)$-Statistic

Consider p bivariate populations $\pi_1, \pi_2, \ldots \pi_p$ having unknown distribution functions that are assumed to be identical except possibly in location parameters. Suppose there are n_1 pairs of observations $(x_{11},x_{21}),\ldots,(x_{1n_1},x_{2n_1})$ from population π_1, n_2 pairs of observations $(x_{1n_1+1},x_{2n_1+1}),\ldots,(x_{1n_1+n_2},x_{2n_1+n_2})$ from populations π_2, etc., and n_p pairs of observations $(x_{1n_1+n_2+\ldots+n_{p-1}+1},x_{1n_1+n_2+\ldots+n_{p-1}+1}),\ldots,$ (x_{1N},x_{2N}) from population π_p, where $N=n_1+n_2+\ldots+n_p$. Suppose, furthermore, that the N values x_{1i} are ranked jointly, receiving ranks that are written symbolically as $u^{(1)}_{1i}, u^{(2)}_{1i},\ldots, u^{(p)}_{1i}$ depending on whether they belong to population $\pi_1, \pi_2,\ldots,$ or π_p. Similarly, the ranks for the observations x_{2i} are $u^{(1)}_{2i}, u^{(2)}_{2i},\ldots,u^{(p)}_{2i}$ depending on whether they belong to population $\pi_1,\pi_2,\ldots,$ or π_p.

The test statistic proposed here, $S_1^2(2,p)$, is the sum of squares of the Euclidean distances between all pairs of centroids in the

bivariate rank space. Using the means $\bar{u}_1^{(r)}$ and $\bar{u}_2^{(r)}$ as defined in (2.3.2), the Euclidean distances between the pair $(\bar{u}_1^{(1)}, \bar{u}_2^{(1)})$, $(\bar{u}_1^{(2)}, \bar{u}_2^{(2)}),\ldots, (\bar{u}_1^{(p)}, \bar{u}_2^{(p)})$ is

$$S_1^2(2,p) = \sum_{s=2}^{p} \sum_{r=1}^{s-1} \sum_{i=1}^{2} (\bar{u}^{(r)} - \bar{u}^{(s)})^2 \qquad (3.3.1)$$

and by using (2.3.3), it follows that

$$S_1^2(2,p) = \sum_{s=2}^{p} \sum_{r=1}^{s-1} \sum_{i=1}^{2} \frac{1}{n_r^2 n_s^2} [n_r R_i^{(r)} - n_r R_i^{(s)}]^2 \qquad (3.3.2)$$

The derivation of the first two moments of $S_1^2(2,p)$ is essentially the same, though more complicated, as the derivation of these moments for S_1^2.

Using the results of Appendix A, it can be shown that

$$E[S_1^2(2,p)] = \frac{N(N+1)}{6} \sum_{s=2}^{p} \sum_{r=1}^{s-1} \frac{n_r + n_s}{n_r n_s} \qquad (3.3.3)$$

and in the special case where $n=n_1=n_2=\ldots=n_p$, this formula reduces to

$$E[S_1^2(2,p)] = \frac{p^2(p-1)(np+1)}{6} \qquad (3.3.4)$$

The derivation of the variance of $S_1^2(2,p)$ is similar to that of S_1^2 given in (2.4.8), but it will not be given here.

3.4 The $S_1^2(k,p)$-Statistic

The test statistic, $S_1^2(k,p)$, proposed here is a direct generalization of the statistics, S_1^2, $S_1^2(k,2)$ and $S_1^2(2,p)$ discussed in previous sections. It is the sum of squares of the Euclidean distances between

all pairs of centroids in the k-dimensional rank space. It can be written as

$$S_1^2(k,p) = \sum_{s=2}^{p} \sum_{r=1}^{s-1} \sum_{i=1}^{k} (\bar{u}_i^{(r)} - \bar{u}_i^{(s)})^2 \qquad (3.4.1)$$

and using (2.3.3), it follows that

$$S_1^2(k,p) = \sum_{p=2}^{p} \sum_{r=1}^{s-1} \sum_{i=1}^{k} \frac{1}{n_r^2 n_s^2} [n_s R_i^{(r)} - n_r R_i^{(s)}]^2 \qquad (3.4.2)$$

The first two moments of this statistic could be derived by employing the methods of Appendix A. However, due to extensive algebraic complications, the variance of $S_1^2(k,p)$ will not be given here. It can be shown that

$$E[S_1^2(k,p)] = \frac{kN(N+1)}{12} \sum_{s=2}^{p} \sum_{r=1}^{s-1} \frac{n_r+n_s}{n_r n_s} \qquad (3.4.3)$$

and in the special case where $n=n_1=n_2=\ldots=n_p$, this formula reduces to

$$E[S_1^2(k,p)] = \frac{kp^2(p-1)(np+1)}{12} \qquad (3.4.4)$$

The general case has not been investigated in any extensive detail.

4. DISCRIMINANT ANALYSIS OF RANKED OBSERVATIONS

4.1 Introduction

When multivariate measurements have been obtained on two or more populations, it is often of interest to consider certain linear functions of these measurements in order to discriminate between the populations. The main objective of linear discriminant analysis is to find a particular linear function

$$z_{ij} = \lambda_1 x_{1ij} + \lambda_2 x_{2ij} + \cdots + \lambda_k x_{kij} \qquad (4.1.1)$$

which provides optimum discrimination in the sense that the quantity

$$G = (\bar{z}_1 - \bar{z}_2)^2 / \sum_{i=1}^{2} \sum_{j=1}^{n_i} (z_{ij} - z_i)^2 \qquad (4.1.2)$$

is maximized with respect to the λ's. Here x_{pij} is the p^{th} observation on the j^{th} variable in the i^{th} population, and

$$\bar{z}_i = \lambda_1 \bar{x}_{1i} + \lambda_2 \bar{x}_{2i} + \cdots + \lambda_k \bar{x}_{ki} \qquad (4.1.3)$$

In the treatment of this theory it is generally assumed that the variates have multivariate normal distributions with equal dispersion matrices. Some general discussions of discriminant analysis may be found in references [10] through [18].

The purpose of this chapter is to present an alternative approach to the problem considered in Chapters 2 and 3. It is proposed that the original measurements are first ranked as in Chapter 3, namely, that the values obtained for the different populations are ranked jointly for each individual variable. The method of discriminant analysis is then applied to these ranks and the resulting z's are reranked.

In Sections 4.2 and 4.3 methods are derived for simplifying the calculations of the λ's when dealing with ranked data. In Section 4.4 it will be investigated whether it is reasonable to apply a Mann-Whitney U-test (or a linear function thereof) to the rankings obtained for the Z's.

4.2 The Bivariate Case

Using the same notation as before, let u_{1i} and u_{2i} denote the ranks assigned to x_{1i} and x_{2i} if these observations belong to population π_1, and let u'_{1i} and u'_{2i} denote the ranks of these same observations if they belong to population π_2.

If the sample points in the two-dimensional rank space are projected onto a vector giving maximum discrimination in the sense of maximizing (4.1.1) and reranking along this vector, the problem is reduced to the analysis of one-dimensional ranks. The purpose of this section is to obtain simplified formulas for the components of the vector providing optimum discrimination.

As shown in [18], the maximization of (4.1.1) gives rise to the following equations:

$$\lambda_1 s_{11} + \lambda_2 s_{12} = cd_1 \qquad (4.2.1)$$

$$\lambda_1 s_{21} + \lambda_2 s_{22} = cd_2$$

where

$$s_{ii} = \sum_{k=1}^{n} (u_{ik}-\bar{u}_i)^2 + \sum_{k=1}^{n} (u'_{ik}-\bar{u}'_i)^2 \qquad (4.2.2)$$

$$s_{ij} = \sum_{k=1}^{n} (u_{ik}-\bar{u}_i)(u_{jk}-\bar{u}_j) + \sum_{k=1}^{n} (u'_{ik}-\bar{u}'_i)(u'_j-\bar{u}'_j) \qquad (4.2.3)$$

$$d_i = \bar{u}_i - \bar{u}'_i \tag{4.2.4}$$

and

$$c = \frac{\sum_{i=1}^{2} \lambda_i d_i}{\dfrac{\sum_{i=1}^{2}\sum_{j=1}^{2} \lambda_i \lambda_j d_i d_j}{\sum_{i=1}^{2}\sum_{j=1}^{2} \lambda_i \lambda_j s_{ij}}} \tag{4.2.5}$$

Recalling that the sum and sum of squares of the first N integers are $\frac{N(N+1)}{2}$ and $\frac{N(N+1)(2N+1)}{6}$, respectively, the above expressions can be rewritten as

$$s_{ii} = \frac{N(N+1)(2N+1)}{6} - \frac{N}{2}(\bar{u}_i^2 + \bar{u}_i'^2) \tag{4.2.6}$$

$$s_{ij} = \sum_{k=1}^{N} u_{ik} u_{jk} - \frac{N}{2}(\bar{u}_1\bar{u}'_1 + \bar{u}_2\bar{u}'_2) \tag{4.2.7}$$

and

$$d_i = 2\bar{u}_i - \frac{N+1}{2} \tag{4.2.8}$$

Substituting these values into equations (4.2.1) and solving simultaneously for λ_1 and λ_2, one obtains, after simplification,

$$\lambda_1 \propto d_1\left[2a - \frac{N(N+1)^2}{2}\right] - d_2\left[2\sum_{k=1}^{N} u_{1k}u_{2k} - \frac{N(N+1)^2}{2}\right] \tag{4.2.9}$$

$$\lambda_2 \propto d_2\left[2a - \frac{N(N+1)^2}{2}\right] - d_1\left[2\sum_{k=1}^{N} u_{1k}u_{2k} - \frac{N(N+1)^2}{2}\right] \tag{4.2.10}$$

where a denotes the sum of squares of the first N integers.

The desired vector thus has direction numbers whose ratio is

$$\frac{\lambda_2}{\lambda_1} = \frac{d_2 - d_1\left[\dfrac{4\sum_{k=1}^{n} u_{1k}u_{2k} - N(N+1)^2}{4a - N(N+1)^2}\right]}{d_1 - d_2\left[\dfrac{4\sum_{k=1}^{n} u_{1k}u_{2k} - N(N+1)^2}{4a - N(N+1)^2}\right]} \qquad (4.2.11)$$

To simplify this further, it can be shown that the expression within the brackets of (4.2.11) is the rank corrrlation coefficient, r'. Therefore, the ratio of the direction numbers of the vector is

$$\frac{\lambda_2}{\lambda_1} = \frac{d_2 - r'd_1}{d_1 - r'd_2} \qquad (4.2.12)$$

This result was previously obtained by Dr. Frank Wilcoxon as communicated to the author in personal correspondence.

4.3 The Multivariate Extension

In this section the method of Section 4.2 is generalized to the k-variate case for two populations. Simplified formulas are obtained for the components of a vector in the k-dimensional rank space, which provides optimum discrimination between the two populations.

The equations to be considered in the multivariate case are

$$\begin{aligned}
\lambda_1 s_{11} + \lambda_2 s_{12} + \cdots + \lambda_k s_{1k} &= cd_1 \\
\lambda_1 s_{21} + \lambda_2 s_{22} + \cdots + \lambda_k s_{2k} &= cd_2 \\
&\\
\lambda_1 s_{k1} + \lambda_2 s_{k2} + \cdots + \lambda_k s_{kk} &= cd_k
\end{aligned} \qquad (4.3.1)$$

Solving these equations for the λ_i and substituting the results

shown in (4.2.6), (4.2.7) and (4.2.8), it can be shown that

$$\lambda_i \propto \sum_{j=1}^{k} C_{ij} d_{ij} \qquad (4.3.2)$$

where $\|C_{ij}\|$ is the matrix of cofactors of $\|r'_{ij}\|$, the matrix of rank correlation coefficients

$$r'_{ij} = 1 - \frac{6 \sum_{k=1}^{N} (u_{ik}-u_{jk})^2}{N(N^2-1)} \qquad (4.3.3)$$

4.4 The Wilcoxon Method

The original purpose for reranking the data after they have been projected on a vector providing maximum discrimination was to perform a test of the null hypothesis that the samples came from populations with identical distribution functions against the alternative that there may be differences in means.

In the bivariate two population case, Dr. Frank Wilcoxon suggests that this test be based on the rank sum obtained along the vector for either population. He proposes to calculate the expectation and variance of this rank sum by means of the formulas $\frac{n_1(N+1)}{2}$ and $\frac{n_1 n_2(N+1)}{12}$, the usual expressions for the mean and variance of rank sums. Then he proposes to use a normal curve approximation to the distribution of the rank sums or a χ^2 approximation to the square of the difference between the observed and the expected rank sum divided by the variance.

It is felt that this test of significance might be reasonable if the direction of the vector were chosen at random, but not if the

vector provides optimum discrimination. Clearly, the vector is chosen to maximize, among other things, the square of the distance between the means and hence it would hardly seem reasonable to use the expectation and variance for ordinary rank sums.

To investigate the reasonableness of the above method suggested by Wilcoxon, the exact distribution of the rank sum was obtained by complete enumeration in the special bivariate two population case, where $n_1=n_2= 2$.

Table VI.

Probabilities of the Rank Sum for Population π_1

Rank Sum for Population π_1	Probability
3	.347
4	.097
5	.111
6	.097
7	.347

It is of interest to note the U-shape of this distribution caused by the fact that the ranked data are projected on a vector providing optimum discrimination, thus giving ranks 3 and 7 very high probabilities. The variance of the above distribution is 2.96 which is much larger than 1.67, the value obtained by substituting $n_1 = n_2 = 2$ into the variance formula given above.

This illustration supports the contention that it is quite unreasonable to apply the test suggested by Wilcoxon to data that have been reranked along a vector providing optimum discrimination.

V. SUMMARY

5.1 Summary

The problem considered in this dissertation deals with two bivariate populations π_1 and π_2 having unknown distribution functions $F_1(x_1,x_2)$ and $F_2(x_1,x_2)$ that are continuous and identical except possibly in location parameters. It is desired to test the null hypothesis

$$H_o: \quad F_1(x_1,x_2) = F_2(x_1,x_2)$$

against the alternative hypothesis that the population distribution functions have different means and it cannot be assumed that the variables x_1 and x_2 are statistically independent.

A test statistic, S_1^2, based on the Euclidean distance between the centroids of the ranks belonging to bivariate samples from π_1 and π_2 is proposed to test the above hypothesis. The first two moments of S_1^2 are derived under a conditional randomization procedure which retains the rank pairs as given in the sample.

The exact sampling distribution of S_1^2 is unknown. However, it is shown in examples that the distribution of a constant multiple of $E(S_1^2)$ divided by the variance of S_1^2 can, at least in these instances, be approximated with a ^{2} distribution with the number of degrees of freedom equal to $2[E(S_1^2)]^2$ divided by $\sigma^2_{S_1^2}$. Tables which facilitate the calculation of $\sigma^2_{S_1^2}$ are given in Appendix B.

As an extension, a statistic, $S_1^2(k,2)$, is proposed for the multivariate two-population case, and its first two moments are derived. Further, statistics are proposed for the bivariate p-population case and for the multivariate p-population case. The first moments are given in each case.

An alternative approach to the solution of the above problem is considered in Chapter 4. In this chapter discriminant analysis is employed to obtain a vector which provides optimum discrimination. It is shown that this method is not a fruitful one for the construction of tests of significance pertaining to the original null hypothesis.

BIBLIOGRAPHY

1. Wilcoxon, F., "Individual Comparisons by Ranking Methods," Biometrics Bulletin, 1(1045) 80-3.

2. _____, "Probability Tables for Individual Comparisons by Ranking Methods," Biometrics, 3(1047) 119-22.

3. Mann, H. B. and Whitney, D. R., "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," Annals of Mathematical Statistics, 18(1947) 50-60.

4. Kruskal, W. H. and Wallis, W. A., "Use of Ranks in One-Criterion Variance Analysis," Journal of the American Statistical Association, 47(1952) 583-621.

5. Kruskal, W. H., "A Non-Parametric Test for the Several Sample Problem," Annals of Mathematical Statistics, 23(1952) 525-40.

6. Hotelling, H., "The Generalization of Student's Ratio," Annals of Mathematical Statistics, 2(1931) 360-78).

7. Bose, R. C., "On the Exact Distribution and Moment-Coefficients of the D^2-Statistic," Sankhya, 2(1936) 143-54.

8. Wald, A., and Wolfowitz, J., "Statistical Tests Based on Permutations of the Observations," Annals of Mathematical Statistics, 15(1944) 367-71.

9. Patnaik, P. B., "The Non-Central χ^2 - and F-Distributions and Their Applications," Biometrika, 36(1949) 202-32.

10. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7(1936) 179-88.

11. _____, "The Statistical Utilization of Multiple Measurements," Annals of Eugenics, 8(1938) 376-86.

12. _____, "The Precision of Discriminant Functions," Annals of Eugenics, 10(1940) 422-9.

13. Bartlett, M. S., "The Goodness of Fit of a Single Hypothetical Discriminant Function in the Case of Several Groups," Annals of Eugenics, 16(1951) 199-214.

14. Rao, C. R., "Tests with Discriminant Functions in Multivariate Analysis," Sankhya, 7(1946) 407-14.

15. _____, "On Some Problems Arising Out of Discriminants with Multiple Characters," Sankhya, 9(1949) 343-66.

16. Williams, E. J., "Some Exact Tests in Multivariate Analysis," Biometrika, 39(1952) 17-31.

17. _____, "Significance Tests for Discriminant Functions and Linear Functional Relationships," Biometrika, 42(1955) 360-81.

18. Hoel, P. G., Introduction to Mathematical Statistics, First Edition, John Wiley and Sons, Inc., New York, 1947.

APPENDIX A

A.1 Derivation of the Expectation of S_1^2

The expectation of S_1^2 under the conditional randomization of Section 2.4 can be written:

$$E(S_1^2) = \frac{N^2}{n_1^2 n_2^2} E[(\sum_{k=1}^{n_1} u_{1k} - \frac{n_1(N+1)}{2})^2 + (\sum_{k=1}^{n_1} u_{2k} - \frac{n_1(N+1)}{2})^2]$$

$$= \frac{2N^2}{n_1^2 n_2^2} E[(\sum_{k=1}^{n_1} u_{1k} - \frac{n_1(N+1)}{2})^2]$$

$$= \frac{2N^2}{n_1^2 n_2^2} [E(\sum_{k=1}^{n_1} u_{1k})^2 - n_1(N+1)E(\sum_{k=1}^{n_1} u_{1k}) + \frac{n_1^2(N+1)^2}{4}]$$

$$= \frac{2N^2}{n_1^2 n_2^2} [n_1 E(u^2) + n_1(n_1-1)E(uu') - n_1(N+1)E(u) + \frac{n_1^2(N+1)^2}{4}] \qquad (A.1.1)$$

where $E(u)$ is the average of the first N integers, $E(u^2)$ is the average of the squares of the first N integers, and $E(uu')$ is the average of the products formed by all possible pairs of integers selected (without replacement) from the first N integers.

Using the well-known fact that the sum and the sum of squares of the first N integers are $\frac{N(N+1)}{2}$ and $\frac{N(N+1)(2N+1)}{6}$, respectively, it follows that

$$E(u) = \frac{N+1}{2} \qquad (A.1.2)$$

and

$$E(u^2) = \frac{(N+1)(2N+1)}{6} \qquad (A.1.3)$$

To evaluate E(uu'), note that

$$(1+2+..+N)^2 = \frac{N(N+1)(2N+1)}{6} + N(N-1)E(uu')$$

and it follows that

$$E(uu') = \frac{(N+1)(3N+2)}{12} \qquad (A.1.4)$$

Substituting (A.1.2), (A.1.3) and (A.1.4) into (A.1.1) gives, after simplification,

$$E(S_1^2) = \frac{N^2(N+1)}{6n_1n_2} \qquad (A.1.5)$$

A.2 Derivation of the Variance of S_1^2

To obtain the variance of S_1^2 it will be necessary to evaluate $E(S_1^4)$ under the conditional randomization of Section 2.4 and substitute the result together with (A.1.5) into (2.4.7). Using (2.3.6), the expectation of S_1^4 can be written:

$$E(S_1^4) = E\frac{N^4}{n_1^4n_2^4}\left[\left(\sum_{k=1}^{n_1}u_{1k} - \frac{n_1(N+1)}{2}\right)^2 + \left(\sum_{k=1}^{n_1}u_{2k} - \frac{n_1(N+1)}{2}\right)^2\right]^2$$

$$= \frac{2N^4}{n_1^4n_2^4}E\left[\left(\sum_{k=1}^{n_1}u_{1k} - \frac{n_1(N+1)}{2}\right)^2\right]^2$$

$$+ E\left[\left(\sum_{k=1}^{n_1}u_{1k} - \frac{n_1(N+1)}{2}\right)^2\left(\sum_{k=1}^{n_1}u_{2k} - \frac{n_1(N+1)}{2}\right)^2\right] \qquad (A.2.1)$$

and after simplification this reduced to:

$$E(s_1^2) = \frac{2N^4}{n_1^4 n_2^4} \quad n_1E(u^4)+4n_1(n_1-1)E(u^3u')$$

$$+ 3n_1(n_1-1)E(u^2(u')^2) + 6n_1(n_1-1)(n_1-2)E(u^2u'u'')$$

$$+ n_1(n_1-1)(n_1-2)(n_1-3)E(uu'u''u''')-2n_1^2(N+1)E(u^3)$$

$$-6n_1^2(n_1-1)(N+1)E(u^2u')+n_1(n_1-1)Eu^2(v')^2)$$

$$-2n_1^2(n_1-1)(n_1-2)(N+1)E(uu'u'')+2n_1^3(N+1)^2E(u^2)$$

$$+ 2n_1^3(n_1-1)(N+1)^2E(uu') - n_1^4(N+1)^3E(u)$$

$$+ \frac{n_1^4(N+1)^4}{8} + n_1^3(N+1)^2E(uv)+n_1^3(n_1-1)(N+1)^2E(uv')$$

$$-n_1^2(N+1)E(u^2v)-n_1^2(N+1)E(uv^2)-n_1^2(n_1-1)(N+1)E(u^2v')$$

$$-n_1^2(n_1-1)(N+1)E(u'v^2)-2n_1^2(n_1-1)(N+1)E(uu'v)$$

$$-2n_1^2(n_1-1)(N+1)E(uvv')-n_1^2(n_1-1)(n_1-2)(N+1)E(uu'v'')$$

$$-n_1^2(n_1-1)(n_1-2)(N+1)E(uv'v'')+n_1E(u^2v^2)$$

$$+ n_1(n_1-1)(n_1-2)E(uu'(v'')^2)+2n_1(n_1-1)E(u^2vv')$$

$$+ n_1(n_1-1)(n_1-2)E(u^2v'v'')+2n_1(n_1-1)E(uvu'v')$$

$$+ 4n_1(n_1-1)(n_1-2)E(uvu'v'') + 2n_1(n_1-1)E(uu'v^2)$$

$$+ n_1(n_1-1)(n_1-2)(n_1-3)E(uu'v''v''') \qquad (A.2.2)$$

where $E(u)$, $E(u^2)$, and $E(uu')$ are as defined in Section A.1. Also $E(u^i)$ is the average of the i^{th} power of the first N integers, $E(u^2u')$ is the average of the products u^2u' formed by all possible pairs of integers, u and u', selected (without replacement) from the first N integers, and $E(uu'u'')$ is the average of the products $uu'u''$ formed

by all possible triplets of integers u, u' and u'', selected (without replacement) from the first N integers. Quantities such as $E(u^2(u')^2)$, $E(u^3u')$, $E(uu'u''u''')$, etc., are defined in a similar fashion.

Expectations involving both u's and v's are defined as follows: E(uv) is the average of all products uv preserving the matching of the conditional randomization and selecting u from the first N integers. Expectations involving primed variables, for example, E(uv') stand for the average of all products uv', where u and v' are selected individually from among the first N integers, but v' is not the particular v that is matched with u. Similarly, $E(u^2v'v'')$ is the average of all products $u^2v'v''$, where u is selected from the first N integers, v' and v'' are a pair of distinct integers selected without replacement from the first N integers, and neither v' nor v'' is the particular v that is paired with the given u in the matching of the conditional randomization. All other expectations are defined in an identical manner.

Using the results given in (A.1.2), (A.1.3) and (A.1.4) along with the fact that the sums of the cubes and the fourth power of the first N integers are $\frac{N^2(N+1)^2}{4}$ and $\frac{N(N+1)(2N+1)(3N^2+3N-1)}{30}$, respectively, it follows that

$$E(u^3) = \frac{N(N+1)^2}{4} \qquad \text{(A.2.3)}$$

and

$$E(u^4) = \frac{(N+1)(2N+1)(3N^2+3N-1)}{30} \qquad \text{(A.2.4)}$$

In order to evaluate $E(u^2u')$ and $E(uu'u'')$ consider the follow-

ing equations:

$$(1^2+2^2+..+N^2)(1+2+..+N)=NE(u^3)+N(N-1)E(u^2u')$$

$$(1+2+..+N)^3=NE(u^3)+3N(N-1)E(u^2u')+N(N-1)(N-2)E(uu'u'')$$

These can be rewritten in the following form:

$$\frac{N^2(N+1)^2(2N+1)}{6} \equiv NE(u^3) + N(N-1)E(u^2u') \qquad (A.2.5)$$

and

$$\frac{N^3(N+1)^3}{8} = NE(u^3) + 3N(N-1)E(u^2u') + N(N-1)(N-2)E(uu'u'') \qquad (A.2.6)$$

Substituting (A.2.3) and solving simultaneously, these equations yield

$$E(u^2u') = \frac{N(N+1)^2}{6} \qquad (A.2.7)$$

and

$$E(uu'u'') = \frac{N(N+1)^2}{8} \qquad (A.2.8)$$

Similarly, the equations

$$(1^3+2^3+..+N^3)(1+2+..+N) = NE(u^4)+N(N-1)E(u^3u')$$

and

$$(1\cdot2+1\cdot3+..+(N-1)\cdot N)(1^2+..+N^2) = N(N-1)E(u^3u') + \frac{N(N-1)(N-2)}{2}E(u^2u'u'')$$

after substituting (A.2.4), can be solved simultaneously to yield

$$E(u^3u') = \frac{(N+1)(15N^3+21N^2-4)}{120} \qquad (A.2.9)$$

and

$$E(u^2u'u'') = \frac{(N+1)(30N^3+35N^2-11N-12)}{360} \qquad (A.2.10)$$

Substituting (A.2.4), (A.2.9) and (A.2.10) into

$$(1^2+2^2+..+N^2) = NE(u^4)+N(N-1)E(u^2(u')^2)$$

and

$$(1+2+..+N)^4 = NE(u^4)+4N(N-1)E(u^3u')$$
$$+ 3N(N-1)E(u^2(u')^2)+6N(N-1)(N-2)E(u^2u'u'')$$
$$+ N(N-1)(N-2)(N-3)E(uu'u''u''')$$

and solving simultaneously gives

$$E(u^2(u')^2) = \frac{(N+1)(2N+1)(2N-1)(5N+6)}{180} \qquad (A.2.11)$$

and

$$E(uu'u''u''') = \frac{(N+1)(15N^3+15N^2-10N-8)}{240} \qquad (A.2.12)$$

Expectations involving both u's and v's are expressed in terms of the parameters A_{11}, A_{12}, A_{21} and A_{22} as defined by (2.4.14), (2.4.15), (2.4.16) and (2.4.17). For example:

$$E(uv) = \frac{A_{11}}{N} \qquad (A.2.13)$$

$$E(u^2v) = \frac{A_{21}}{N} \qquad (A.2.14)$$

$$E(uv^2) = \frac{A_{12}}{N} \qquad (A.2.15)$$

$$E(u^2v^2) = \frac{A_{22}}{N} \qquad (A.2.16)$$

To evaluate E(uv'), note that

$$(1+2+..+N)(1+2+..+N) = A_{11}+N(N-1)E(uv')$$

and it follows that

$$E(uv') = \frac{N(N+1)^2}{4(N-1)} - \frac{A_{11}}{N(N-1)} \qquad (A.2.17)$$

To evaluate E(uvv'), the following equation is employed:

$$(1+2+..+N)A_{11} = A_{12} + N(N-1)E(uvv')$$

It follows that

$$E(uvv') = \frac{(N+1)A_{11}}{2(N-1)} - \frac{A_{12}}{N(N-1)} \qquad (A.2.18)$$

Similarly, the equation

$$(1^2+2^2+..+N^2)(1+2+..+N) = A_{12}+N(N-1)E(u'v^2)$$

gives

$$E(u'v^2) = \frac{N(N+1)^2(2N+1)}{12(N-1)} - \frac{A_{12}}{N(N-1)} \qquad (A.2.19)$$

and the following expectations are obtained in an identical manner:

$$E(u^2v') = \frac{N(N+1)^2(2N+1)}{12(N-1)} - \frac{A_{21}}{N(N-1)} \qquad (A.2.20)$$

$$E(uu'v) = \frac{(N+1)A_{11}}{2(N-1)} - \frac{A_{21}}{N(N-1)} \qquad (A.2.21)$$

$$E(uu'v'') = \frac{N(N+1)^2(3N+2)}{24(N-2)} - \frac{(N+1)A_{11}}{(N-1)(N-2)} + \frac{2A_{21}}{N(N-1)(N-2)} \qquad (A.2.22)$$

$$E(uv'v'') = \frac{N(N+1)^2(3N+2)}{24(N-2)} - \frac{(N+1)A_{11}}{(N-1)(N-2)} + \frac{2A_{12}}{N(N-1)(N-2)} \qquad (A.2.23)$$

$$E(u^2(v')^2) = \frac{N(N+1)^2(2N+1)^2}{36(N-1)} - \frac{A_{22}}{N(N-1)} \tag{A.2.24}$$

$$E(uu'v^2) = \frac{(N+1)A_{12}}{2(N-1)} - \frac{A_{22}}{N(N-1)} \tag{A.2.25}$$

$$E(uu'(v'')^2) = \frac{N(N+1)^2(2N+1)(3N+2)}{72(N-2)} - \frac{(N+1)A_{12}}{(N-1)(N-2)}$$

$$+ \frac{2A_{22}}{N(N-1)(N-2)} \tag{A.2.26}$$

$$E(u^2vv') = \frac{(N+1)A_{21}}{2(N-1)} - \frac{A_{22}}{N(N-1)} \tag{A.2.27}$$

$$E(u^2v'v'') = \frac{N(N+1)^2(2N+1)(3N+2)}{72(N-2)} - \frac{(N+1)A_{21}}{(N-1)(N-2)}$$

$$+ \frac{2A_{22}}{N(N-1)(N-2)} \tag{A.2.28}$$

$$E(uvu'v') = \frac{A_{11}^2 - A_{22}}{N(N-1)} \tag{A.2.29}$$

$$E(uvu'v'') = \frac{N(N+1)^2A_{11}}{4(N-1)(N-2)} - \frac{A_{11}^2}{N(N-1)(N-2)} \tag{A.2.30}$$

$$- \frac{(N+1)A_{12}}{(N-1)(N-2)} + \frac{2A_{22}}{N(N-1)(N-2)}$$

$$E(uu'v''v''') = \frac{1}{N(N-1)(N-2)(N-3)} \quad \frac{N^4(N+1)^4}{16} \; N^2(N+1)^2A$$

$$- \frac{N2(N+1)^2(2N+1)(3N^2+N-1)}{36} + 2A_{11}^2 - 6A_{22}$$

$$+ 2N(N+1)(A_{12}+A_{21}) \tag{A.2.31}$$

Substituting the above results into (2.4.7) and simplifying, it follows that the variance of S_1^2 equals the expression given in (2.4.8).

APPENDIX B

Table I. Values of a_{oo}

n_1 \ n_2	3	4	5	6	7	8	9	10	11
3	498.8839	567.6438	678.2456	829.9152	1025.032	1258.825	1560.260	1920.040	2342.660
4		639.0484	730.7307	875.6935	1049.391	1262.237	1517.934	1820.835	2175.799
5			830.4965	955.1957	1112.661	1304.280	1532.370	1799.893	2110.281
6				1067.819	1212.079	1387.980	1733.558	1921.823	2120.368
7					1343.689	1506.868	1700.636	1926.232	2185.189
8						1657.515	1838.585	2050.027	2292.590
9							2007.651	2206.729	2435.773
10								2394.033	2611.016
11									2816.442
12									
13									
14									
15									
16									
17									
18									
19									
20									

Table I (continued)

n_1 \ n_2	12	13	14	15	16	17	18	19	20
3	2841.363	3419.065	4089.653	4860.509	5741.249	6741.954	7873.169	9145.912	10571.626
4	2588.112	3063.449	3608.733	4227.574	5443.714	5720.226	6607.393	7598.486	8701.401
5	2467.161	2876.333	3338.210	3861.041	4448.561	5105.873	5838.306	6651.403	7550.920
6	2440.563	2803.668	3212.958	3671.967	4184.412	4754.191	5350.210	6080.145	6849.077
7	2479.477	2811.383	3182.129	3598.359	4057.694	4568.680	5130.267	5747.405	6423.392
8	2567.555	2876.566	3221.543	3604.615	4027.214	4494.435	5006.223	5566.172	6177.107
9	2695.398	2986.681	3311.031	3670.073	4065.638	4499.691	4956.642	5491.890	6054.509
10	2857.642	3134.424	3415.334	3750.402	4156.259	4565.365	5011.467	5496.401	6022.050
11	3051.277	3315.483	3609.506	3934.138	4290.421	4679.589	5103.030	5562.256	6058.883
12	3274.744	3527.395	3809.196	4120.398	4462.086	4834.779	5239.698	5678.069	6151.235
13		3768.851	4270.902	4543.242	4846.253	5179.816	5544.199	5939.933	6367.755
14			4298.702	4586.933	4903.963	5250.087	5700.567	6106.081	6543.190
15				4864.256	5170.263	5525.420	5868.513	6261.496	6684.522
16					5465.483	5789.339	6141.647	6522.722	6932.997
17						6102.359	6443.904	6813.876	7212.437
18							6774.867	7134.179	7521.820
19								7482.995	7860.071
20									8226.734

Table II. Values of a_{11}

n_1 \ n_2	3	4	5	6	7	8	9	10	11
3	-17.42222	-12.28025	-9.596343	-8.035714	-7.051246	-6.393113	-5.934705	-5.606172	-5.366267
4		-9.257143	-7.485268	-6.391002	-5.675182	-5.185227	-4.838500	-4.587325	-4.402622
5			-6.146032	-5.284563	-4.760989	-4.302042	-4.011488	-3.797802	-3.638250
6				-4.552189	-4.050905	-3.695762	-3.436892	-3.244014	-3.097979
7					-3.596404	-3.270873	-3.031202	-2.850809	-2.712712
8						-2.964103	-2.736552	-2.563963	-2.430725
9							-2.516776	-2.378330	-2.218869
10								-2.184520	-2.056031
11									-1.928458
12									
13									
14									
15									
16									
17									
18									
19									
20									

Table II (continued)

n_1 \ n_2	12	13	14	15	16	17	18	19	20
3	-5.189255	-5.058333	-4.962110	-4.892612	-4.844105	-4.812367	-4.794224	-4.787248	-4.789549
4	-4.265798	-4.164481	-4.090154	-4.036781	-3.999976	-3.976480	-3.963826	-3.960116	-3.963862
5	-3.518059	-3.427278	-3.358987	-3.308244	-3.271143	-3.245852	-3.229447	-3.220630	-3.218153
6	-2.986213	-2.912020	-2.833882	-2.783047	-2.744515	-2.712921	-2.695462	-2.681739	-2.673653
7	-2.605678	-2.522035	-2.456391	-2.404860	-2.364580	-2.333401	-2.309678	-2.292135	-2.279767
8	-2.326461	-2.244055	-2.178492	-2.126153	-2.084363	-2.051112	-2.024862	-2.004416	-1.988833
9	-2.198999	-2.034366	-1.968860	-1.915451	-1.872377	-1.837460	-1.809231	-1.786548	-1.768511
10	-1.945197	-1.872508	-1.806362	-1.752428	-1.708242	-1.671951	-1.642131	-1.617678	-1.597722
11	-1.826930	-1.745094	-1.678451	-1.623746	-1.578571	-1.541110	-1.509974	-1.484083	-1.462583
12	-1.725328	-1.643128	-1.575900	-1.520432	-1.474351	-1.435868	-1.403611	-1.376515	-1.353740
13		-1.560401	-1.492516	-1.436287	-1.389358	-1.349955	-1.316716	-1.288587	-1.264733
14			-1.423915	-1.366921	-1.318762	-1.278933	-1.244815	-1.215777	-1.190987
15				-1.309153	-1.260632	-1.219589	-1.184666	-1.154811	-1.129195
16					-1.211346	-1.169547	-1.133875	-1.103273	-1.076911
17						-1.127020	-1.090640	-1.059345	-1.032299
18							-1.053583	-1.021638	-0.993956
19								-0.989067	-0.960791
20									-0.931947

Table III. Values of a_{12}

n_1 \ n_2	3	4	5	6	7	8	9	10	11
3	0.829630	0.423457	0.234057	0.133929	0.075414	0.038513	0.013834	-0.003452	-0.016019
4		0.257143	0.162723	0.107359	0.072759	0.049858	0.033961	0.022487	0.013932
5			0.111746	0.078874	0.057157	0.042234	0.031587	0.023736	0.017783
6				0.058361	0.044032	0.033834	0.026380	0.020783	0.016479
7					0.034251	0.027032	0.021625	0.017495	0.014277
8						0.021795	0.017770	0.014640	0.012170
9							0.014718	0.012299	0.010364
10								0.010402	0.008862
11									0.007622
12									
13									
14									
15									
16									
17									
18									
19									
20									

Table III (continued)

n_1 \ n_2	12	13	14	15	16	17	18	19	20
3	-0.025438	-0.032679	-0.038367	-0.042918	-0.046617	-0.049666	-0.052210	-0.054355	-0.056182
4	0.007380	0.002246	-0.001856	-0.005189	-0.007936	-0.010567	-0.012169	-0.013822	-0.015246
5	0.013160	0.009494	0.006535	0.004110	0.002094	0.000400	-0.001040	-0.002276	-0.003345
6	0.013097	0.010392	0.008190	0.006374	0.004856	0.003574	0.002479	0.001535	0.000716
7	0.011724	0.009663	0.007975	0.006575	0.005399	0.004400	0.003545	0.002806	0.002163
8	0.010190	0.008580	0.007252	0.006145	0.005211	0.004415	0.003731	0.003139	0.002621
9	0.008796	0.007511	0.006445	0.005552	0.004795	0.004148	0.003590	0.003105	0.002681
10	0.007601	0.006559	0.005689	0.004956	0.004391	0.003797	0.003334	0.002931	0.002577
11	0.006595	0.005739	0.005019	0.004409	0.003888	0.003439	0.003050	0.002817	0.002411
12	0.005751	0.005040	0.004438	0.003925	0.003485	0.003104	0.002773	0.002483	0.002227
13		0.004446	0.003938	0.003503	0.003127	0.002801	0.002517	0.002267	0.002046
14			0.003507	0.003135	0.002812	0.002531	0.002285	0.002068	0.001876
15				0.002815	0.002536	0.002293	0.002078	0.001888	0.001720
16					0.002294	0.002081	0.001893	0.001726	0.001578
17						0.001894	0.001728	0.001581	0.001450
18							0.001582	0.001451	0.001410
19								0.001335	0.001230
20									0.001137

Table IV. Values of a_{22}

n_1 \ n_2	3	4	5	6	7	8	9	10	11
3	-0.118519	-0.052932	-0.026006	-0.013393	-0.006856	-0.003209	-0.000337	0.000247	0.001068
4		-0.028571	-0.016272	-0.009760	-0.006063	-0.003835	-0.002426	-0.001499	-0.000871
5			-0.010159	-0.006573	-0.004397	-0.003017	-0.002106	-0.001484	-0.001046
6				-0.004489	-0.003145	-0.002226	-0.001649	-0.001223	-0.000915
7					-0.002283	-0.001690	-0.001272	-0.000972	-0.000751
8						-0.001282	-0.000987	-0.000771	-0.000608
9							-0.000775	-0.000615	-0.000493
10								-0.000495	-0.000403
11									-0.000331
12									
13									
14									
15									
16									
17									
18									
19									
20									

Table IV (continued)

n_1 \ n_2	12	13	14	15	16	17	18	19	20
3	0.001590	0.001922	0.002131	0.002259	0.002331	0.002365	0.002373	0.002363	0.002341
4	-0.000434	-0.000125	0.000098	0.000259	0.000378	0.000456	0.000529	0.000576	0.000610
5	-0.000731	-0.000500	-0.000327	-0.000196	-0.000073	-0.000017	0.000043	0.000091	0.000129
6	-0.000689	-0.000520	-0.000390	-0.000290	-0.000211	-0.000149	-0.000099	-0.000059	-0.000027
7	-0.000586	-0.000460	-0.000363	-0.000286	-0.000225	-0.000176	-0.000136	-0.000104	-0.000077
8	-0.000485	-0.000390	-0.000315	-0.000256	-0.000208	-0.000170	-0.000138	-0.000112	-0.000090
9	-0.000400	-0.000327	-0.000269	-0.000222	-0.000184	-0.000154	-0.000128	-0.000107	-0.000089
10	-0.000330	-0.000273	-0.000228	-0.000191	-0.000160	-0.000136	-0.000115	-0.000098	-0.000083
11	-0.000275	-0.000230	-0.000193	-0.000163	-0.000139	-0.000119	-0.000104	-0.000087	-0.000075
12	-0.000230	-0.000194	-0.000164	-0.000140	-0.000120	-0.000103	-0.000091	-0.000078	-0.000034
13		-0.000165	-0.000139	-0.000121	-0.000104	-0.000090	-0.000079	-0.000069	-0.000060
14			-0.000121	-0.000104	-0.000091	-0.000079	-0.000069	-0.000061	-0.000054
15				-0.000091	-0.000079	-0.000069	-0.000061	-0.000054	-0.000048
16					-0.000070	-0.000061	-0.000054	-0.000048	-0.000043
17						-0.000054	-0.000048	-0.000043	-0.000038
18							-0.000043	-0.000038	-0.000034
19								-0.000034	-0.000031
20									-0.000028

Table V. Values of $a_{11,11}$

n_1 \ n_2	3	4	5	6	7	8	9	10	11
3	0.079012	0.039699	0.023117	0.014881	0.010284	0.007489	0.005675	0.004438	0.003560
4		0.021429	0.013018	0.008611	0.006063	0.004474	0.003425	0.002698	0.002177
5			0.008127	0.005477	0.003908	0.002913	0.002246	0.001780	0.001443
6				0.003741	0.002696	0.002024	0.001570	0.001250	0.001017
7					0.001957	0.001478	0.001152	0.000921	0.000751
8						0.001122	0.000878	0.000704	0.000576
9							0.000689	0.000553	0.000454
10								0.000446	0.000366
11									0.000301
12									
13									
14									
15									
16									
17									
18									
19									
20									

Table V (continued)

n_1 \ n_2	12	13	14	15	16	17	18	19	20
3	0.002915	0.002428	0.002053	0.001757	0.001520	0.001328	0.001169	0.001038	0.000927
4	0.001791	0.001497	0.001270	0.001090	0.000945	0.000827	0.000729	0.000648	0.000579
5	0.001191	0.000999	0.000850	0.000731	0.000635	0.000556	0.000491	0.000437	0.000391
6	0.000843	0.000709	0.000604	0.000520	0.000452	0.000397	0.000351	0.000313	0.000280
7	0.000624	0.000527	0.000449	0.000387	0.000337	0.000296	0.000262	0.000234	0.000210
8	0.000479	0.000404	0.000346	0.000299	0.000261	0.000557	0.000203	0.000181	0.000162
9	0.000378	0.000320	0.000274	0.000237	0.000207	0.000182	0.000161	0.000144	0.000129
10	0.000306	0.000259	0.000222	0.000192	0.000168	0.000148	0.000131	0.000117	0.000105
11	0.000252	0.000214	0.000183	0.000159	0.000139	0.000122	0.000109	0.000097	0.000087
12	0.000211	0.000179	0.000154	0.000133	0.000117	0.000103	0.000091	0.000082	0.000073
13		0.000152	0.000131	0.000113	0.000099	0.000088	0.000078	0.000070	0.000063
14			0.000112	0.000098	0.000085	0.000075	0.000067	0.000060	0.000054
15				0.000085	0.000074	0.000066	0.000058	0.000052	0.000047
16					0.000065	0.000058	0.000051	0.000046	0.000041
17						0.000051	0.000045	0.000041	0.000037
18							0.000040	0.000036	0.000033
19								0.000032	0.000029
20									0.000026

ACKNOWLEDGMENTS

The author is indebted to Doctor John E. Freund for his encouragement and guidance during the preparation of this dissertation and to Doctor Frank Wilcoxon for suggesting the problem.

The author is grateful to Doctor Boyd Harshbarger for his counsel and interest during the author's graduate study.

The assistance and advice of Doctor R. A. Bradley and Doctor R. L. Wine are greatly appreciated.

The author is particularly appreciative of the interest that Mrs. Dotty Welcher took in preparing the final manuscript. He also wishes to thank Mrs. Janice Belcher, Mrs. Patricia Vandevender and Miss Mary McGahey for their help in preparing the final manuscript.

The vita has been removed from the scanned document

ABSTRACT

ON THE ANALYSIS OF PAIRED RANKED OBSERVATIONS

by

Leo Lynch, B.Sc., M.S.

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute
in candidacy for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

August, 1957

The problem considered in this dissertation is the following: let π_1 and π_2 be two bivariate populations having unknown cumulative distribution functions $F_1(x_1,x_2)$ and $F_2(x_1,x_2)$, respectively. Assume that F_1 and F_2 are continuous and identical except possibly in location parameters. It is desired to test the null hypothesis

$$H_o: \quad F_1(x_1,x_2) \equiv F_2(x_1,x_2)$$

against the alternative

$$H_o: \quad F_1(x_1,x_2) \neq F_2(x_1,x_2)$$

It cannot be assumed that the variables x_1 and x_2 are statistically independent.

Suppose there are n_1 pairs of observations $(x_{11},x_{21}),\ldots,(x_{1n_1},x_{2n_1})$ from population π_1 and n_2 pairs of observations $(x_{1n_1+1},x_{2n_1+1}),\ldots,(x_{1N},x_{2N})$ from population π_2, where $N = n_1 + n_2$. The x_{1i} $(i = 1,2,\ldots,N)$ are ranked according to magnitude, the largest being assigned rank 1 and the smallest assigned rank N. In a similar manner, ranks are assigned to the observations x_{2i} $(i = 1,2,\ldots,N)$. It is assumed that there are no ties in ranks.

Let u_{1i} and u_{2i} denote the ranks assigned to x_{1i} and x_{2i} if these observations belong to population π_1, and let u'_{1i} and u'_{2i} denote the ranks of these same observations if they belong to population π_2. Since the sum of the first N integers is $\frac{N(N+1)}{2}$, it follows that

$$\sum_{k=1}^{n_1} u_{1k} + \sum_{k=n_{1+1}}^{N} u'_{ik} = \frac{N(N+1)}{2}$$

If the N pairs of ranks are plotted on a plane, it is likely that the n_1 points from population π_1 and the n_2 points from population π_2 will be interspersed forming a circular or elliptical pattern under the assumption that $F_1(x_1,x_2)$ and $F_2(x_1,x_2)$ are identical. Under the alternative hypothesis, it is likely that there will be a segregation of the points into two groups. A test statistic, S_1^2 is constructed to measure the extent of this segregation.

The S_1^2-statistic proposed here, is based on the Euclidean distance between the centroids of the ranks belonging to π_1 and π_2, in particular,

$$S_1^2 = (\bar{u}_1 - \bar{u}'_1)^2 + (\bar{u}_2 - \bar{u}'_2)^2$$

where

$$\bar{u}_i = n_1^{-1} \sum_{k=1}^{n_1} u_{ik} \quad , \quad u'_i = n_2^{-1} \sum_{k=n_{1+1}}^{N} u'_{ik}$$

The first two moments of S_1^2 are derived under the following conditional randomization procedure: keeping the ranks paired as given in the sample, n_1 pairs are selected at random (with equal probabilities) from among the $N = n_1 + n_2$ pairs and assigned to population π_1; the remaining n_2 pairs are assigned to population π_2. It is shown that

$$E(S_1^2) = \frac{N^2(N+1)}{6\, n_1 n_2}$$

and

$$\sigma^2_{S_1^2} = a_{oo} + a_{11}A_{11} + a_{12}A_{12} + a_{21}A_{21} + a_{22}A_{22} + a_{11,11}A_{11}^2$$

where $A_{rs} = \sum_{k=1}^{N} u_{1k}^r u_{2k}^s$ are parameters depending on the sample, and the coefficients a_{oo}, a_{11}, a_{12}, a_{22} and $a_{11,11}$ have been tabulated for values of n_1 and n_2 up to 20.

The exact sampling distribution of S_1^2 is unknown. However, it is shown that the distribution of $\frac{kE(S_1^2)}{\sigma^2_{S_1^2}}$ is approximately χ^2 with $\frac{2[E(S_1^2)]^2}{\sigma^2_{S_1^2}}$ degrees of freedom.

A rank analogue of Wald's modification of Hotelling's T^2 is given and the first two moments obtained. Also, a multivariate extension is considered and a statistic, $S_1^2(k,2)$, constructed. The expectation and variance of $S_1^2(k,2)$ are derived. A multi-population extension for the case of bivariate populations is given and the expectation is derived for a statistic, $S_1^2(2,p)$. A statistic, $S_1^2(k,p)$ is constructed for the most general case and its expectation is given.

An alternative approach to the problem, also investigated, is by means of discriminant analysis. In this case simplified formulas are given for the calculation of the components of a vector which provides

optimum discrimination. It is shown that this method is not a fruitful one for the construction of tests of significance pertaining to the original null hypothesis.