МАТЕМАТИЧКИ ВЕСНИК
59 (2007), 161–170

UDK 519.24
оригинални научни рад
research paper

CORRELATION ANALYSIS: EXACT PERMUTATION PARADIGM

Justice Ighodaro Odiase and Sunday Martins Ogbonmwan

Abstract. For a general class of problems, the permutation of observations is the only possible method of truly constructing exact tests of significance. The exact sampling distribution of a test statistic for an experiment compiled by the permutation approach requires no reference to a population distribution and therefore no requirement that it should conform to a mathematically definable frequency distribution. Algorithms for the exact permutation distribution of correlation coefficients are presented and implemented. As an illustrative example, critical values for Pearson's product moment and Spearman's rank correlation coefficients are produced for Charles Darwin's data on the heights of cross and self fertilized plants.

1. Introduction

There are several experimental situations in which there is only one set of n experimental subjects and two-observations are made on each subject. The data consists of n pairs, such as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. In an attempt to ensure that the probability of a type I error is exactly α in analyzing the linear relationship for paired observations, an algorithm for obtaining exact permutation distribution of paired observations is presented.

A major problem of statistical inference is to obtain an exact test of significance when the form of the underlying probability distribution is unknown. The idea of a general method of dealing with this problem of obtaining an exact test of significance originated with Fisher [5]. The essential feature of the method is that all the distinct permutations of the observations is considered, with the property that each permutation is equally likely under the hypothesis to be tested. An exact test on the level of significance is constructed by choosing a proportion, α, of the permutation as critical region. It is elaborately shown in Scheffe [16] that for a general class of problems, the permutation approach is the only possible method of constructing exact tests of significance.

Several approaches which are computationally less demanding have been suggested as alternatives to the permutation approach. Permutation tests have received attention under the guise of bootstrap, see Efron [4]. Other approaches like

AMS Subject Classification: 62E15

Keywords and phrases: Permutation test, Monte Carlo test, p-value, algorithm, paired observations, correlation.

the Bayesian and the likelihood have also been found useful in obtaining approximate permutation distribution, see Bayarri and Berger [2], Spiegelhalter [18].

Permutation tests are attractive because the distribution of the observations under the null hypothesis need not be known to calculate the p-value. It is asymptotically as powerful as the best parametric test when based on the same statistic, see Hoeffding [10]. Available permutation procedures can sample from the permutation sample space rather than carrying out complete enumeration of all possible distinct rearrangements. These available procedures can perform Monte Carlo sampling without replacement within a sample, but none can avoid the possibility of drawing the same sample more than once, thereby reducing the power of the permutation test, see Opdyke [14]. This paper therefore presents an algorithm that makes it possible to obtain all the distinct permutations of an experiment without the problem of drawing a sample more than once.

2. Correlation analysis

Correlation coefficient has become the workhorse of quantitative research and analysis. Relationships among things are often examined in terms of whether they change together or separately. The world around us is understood through the multifold and interlaced correlations it manifests.

The permutation method discussed in this paper is applied to measure linear association in paired, exchangeable observations. Exchangeability is a generalization of the concept of independent, identically distributed random variables. Permutation analysis of correlation assumes that in the null hypothesis, two variables X and Y ($x_i \in X$; $y_i \in Y$) are independent within each individual unit and pairs (x_i, y_i), $i = 1(1)n$ are independent and identically distributed. Paired, exchangeable observations (x_1, y_1) have the same distribution as (y_1, x_1), and the marginal distributions of x_1 and y_1 are identical. A test of exchangeability of paired observations is given by Hollander [11]. Computational advances involving the use of permutation tests are well documented in Hilton and Gee [9], Hilton [8], Good [7] and Pesarin [15].

The two most commonly used correlation coefficients are the Pearson's correlation coefficient and the Spearman's rank correlation coefficient. Given the observations (x_i, y_i), $i = 1(1)n$, the Pearson's correlation coefficient is defined as

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}.$$

When r is calculated from sample data, the obtained value is only an estimate of a corresponding population correlation coefficient, denoted by ρ. To test the null hypothesis of no correlation, for example, H_0: $\rho = 0$, we assume that both variables are measured on an interval or ratio scale. The calculation is based on the actual values and both variables (X and Y) have a normal distribution. If all the assumptions are met and H_0: $\rho = 0$ is true, then, for n pairs of observations, $t = \frac{r\sqrt{n-1}}{\sqrt{1-r^2}}$ has the t distribution with $n - 2$ degrees of freedom. A more general way to test H_0: $\rho = \rho_0$ or construct confidence intervals for ρ is based on Fisher Z transformation, $Z = \frac{1}{2}\ln\frac{1+r}{1-r}$. Z is approximately normal with $\eta = (Z-\mu_Z)\sqrt{n-3}$ having approximately the standard normal distribution, see Freund [6].

To calculate the rank-correlation coefficient for n pairs of observations, find the sum of the squares of the differences, d, between the ranks of the $X's$ and $Y's$, and substitute into the formula

$$r_s = 1 - \frac{6(\sum d^2)}{n(n^2-1)}.$$

When there are ties, assign to each of the tied observations the mean of the ranks which they jointly occupy. When using r_s to test the null hypothesis of no correlation between two variables X and Y, we do not have to make any assumptions about the nature of the populations sampled. To test the null hypothesis, the statistic, $z = \frac{r_s - 0}{1/\sqrt{n-1}} = r_s\sqrt{n-1}$, which has approximately the standard normal distribution is employed.

3. Methodology

The p-value of a test statistic represents the probability of obtaining values of the test statistic that are equal to or more extreme than the observed value of the test statistic. In this paper, consideration is given to the permutation distribution of paired observations on which the correlation coefficient is to be computed, see Agresti [1] for conditional permutation. Also, see Odiase and Ogbonmwan [12] for an algorithm for generating unconditional exact permutation distribution for a two-sample (independent samples) experiment.

In this section, we will use the techniques of Odiase and Ogbonmwan [13]. Given a bivariate sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ for which $(x_1, x_2, \ldots, x_n) \sim F_X$ and $(y_1, y_2, \ldots, y_n) \sim F_Y$ are simultaneously tested in an experiment with R as the test statistic. Let $H_0 : F_X = F_Y$ against $H_1 : F_X \neq F_Y$ or $H_1 : F_X < F_Y$ or $H_1 : F_X > F_Y$. Each of the 2^n distinct permutations occurs with the probability $\frac{1}{2^n}$. For k distinct values of the test statistic R, the probability distribution of the test statistic $R = (R_1, R_2, \ldots, R_k)$ under the null hypothesis $H_0 : F_X = F_Y$ is given by

$$P(R_j = R_1|H_0) = \sum_{i=1}^{f_j}\left(\frac{1}{2^n}\right) = \frac{f_j}{2^n},$$

where f_j is the number of occurrences of R_j. Ordering all the distinct occurrences of R in ascending order of magnitude, if g is the position of the observed value of R, the following significance level for the left tail of the distribution of the test statistic is obtained.

$$\alpha = P(R_g \leq c|H_0) = \sum_{j=1}^{g}\sum_{i=1}^{f_j}\left(\frac{1}{2^n}\right) = \left(\frac{1}{2^n}\right)\sum_{j=1}^{g} f_j,$$

and for the right tail,

$$\alpha = P(R_g \geq c|H_0) = \left(\frac{1}{2^n}\right)\sum_{j=g}^{k} f_j.$$

Let a two-sample layout with n variates in each sample be represented by $\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$, where $x_i \in X$ and $y_j \in Y$ in an experiment. Choose the test statistic R, such as Spearman's coefficient of correlation and the acceptable significance level α. Let π_1, π_2, …, π_{2^n} be a set of all distinct permutations of the data set. The permutation test procedure is outlined as follows:

1. Compute the test statistic R_1 for the original arrangement π_1.
2. Obtain a distinct permutation π_i based on Algorithm 2.
3. Compute the test statistic for permutation π_i, $R_i = R(\pi_i)$.
4. Repeat Steps 2 and 3 for $i = 2, 3, \ldots, 2^n$; $n =$ sample size.
5. Construct an empirical cumulative distribution

$$p_0 = \hat{p}(R \leq R_i) = \frac{1}{2^n} \sum_{i=1}^{2^n} \theta(R_1 - R_i)$$

   where θ is a step-function, that is, $\theta = 1$, if $R_1 \geq R_i$, and $\theta = 0$ otherwise.
6. Under the empirical distribution $\hat{p}$ if $p_0 \leq \alpha$, reject the null hypothesis.

The six steps compute the cumulative distribution of the test statistic R exactly, under the null hypothesis. The nonparametric analogue is obtained if we choose to use ranks in the permutations or configurations rather than the actual observations. The number of distinct permutations is obtained as $\sum_{i=0}^{n} \binom{n}{i}$, this is the expression that really facilitates the distinct enumeration, see Odiase and Ogbonmwan [13]. In a way, permutation of paired observations is a constrained permutation since the concept of exchangeability of observations is only applicable within pairs.

As an illustration, examine an experiment with five pairs of observations, that is, $\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ x_5 & y_5 \end{pmatrix}$. By following the permutation procedure described so far, the experiment results in $2^5 = 32$ permutations, see Table 1.

Observe that permutation (1) is the original arrangement $\binom{5}{0}$, permutations (2) to (6) are obtained by switching one pair of observations $\binom{5}{1}$. Permutations (7) to (16) are obtained by switching two pairs of observations $\binom{5}{2}$. Permutations (17) to (26) are obtained by switching three pairs of observations $\binom{5}{3}$. Permutations (27) to (31) are obtained by switching four pairs of observations $\binom{5}{4}$, while permutation (32) is obtained by switching all the five pairs of observations $\binom{5}{5}$ in the experiment.

TABLE 1. Permutations of a 2 × 5 Paired Sample Experiment

1	2	3	4	5	6	7	8
x_1 y_1	x_1 y_1	x_1 y_1	x_1 y_1	x_1 y_1	y_1 x_1	x_1 y_1	x_1 y_1
x_2 y_2	x_2 y_2	x_2 y_2	x_2 y_2	y_2 x_2	x_2 y_2	x_2 y_2	x_2 y_2
x_3 y_3	x_3 y_3	x_3 y_3	y_3 x_3	x_3 y_3	x_3 y_3	x_3 y_3	y_3 x_3
x_4 y_4	x_4 y_4	y_4 x_4	x_4 y_4	x_4 y_4	x_4 y_4	y_4 x_4	x_4 y_4
x_5 y_5	y_5 x_5	x_5 y_5	x_5 y_5	x_5 y_5	x_5 y_5	y_5 x_5	y_5 x_5
9	10	11	12	13	14	15	16
x_1 y_1	y_1 x_1	x_1 y_1	x_1 y_1	y_1 x_1	x_1 y_1	y_1 x_1	y_1 x_1
y_2 x_2	x_2 y_2	x_2 y_2	y_2 x_2	x_2 y_2	y_2 x_2	x_2 y_2	y_2 x_2
x_3 y_3	x_3 y_3	y_3 x_3	x_3 y_3	x_3 y_3	y_3 x_3	y_3 x_3	x_3 y_3
x_4 y_4	x_4 y_4	y_4 x_4	y_4 x_4	y_4 x_4	x_4 y_4	x_4 y_4	x_4 y_4
y_5 x_5	y_5 x_5	x_5 y_5	x_5 y_5	x_5 y_5	x_5 y_5	x_5 y_5	x_5 y_5
17	18	19	20	21	22	23	24
x_1 y_1	x_1 y_1	y_1 x_1	x_1 y_1	y_1 x_1	y_1 x_1	x_1 y_1	y_1 x_1
x_2 y_2	y_2 x_2	x_2 y_2	y_2 x_2	x_2 y_2	y_2 x_2	y_2 x_2	x_2 y_2
y_3 x_3	x_3 y_3	x_3 y_3	y_3 x_3	y_3 x_3	x_3 y_3	y_3 x_3	y_3 x_3
y_4 x_4	y_4 x_4	y_4 x_4	x_4 y_4	x_4 y_4	x_4 y_4	y_4 x_4	y_4 x_4
y_5 x_5	y_5 x_5	y_5 x_5	y_5 x_5	y_5 x_5	y_5 x_5	x_5 y_5	x_5 y_5
25	26	27	28	29	30	31	32
y_1 x_1	y_1 x_1	x_1 y_1	y_1 x_1	y_1 x_1	y_1 x_1	y_1 x_1	y_1 x_1
y_2 x_2	y_2 x_2	y_2 x_2	x_2 y_2	y_2 x_2	y_2 x_2	y_2 x_2	y_2 x_2
x_3 y_3	y_3 x_3	y_3 x_3	y_3 x_3	x_3 y_3	y_3 x_3	y_3 x_3	y_3 x_3
y_4 x_4	x_4 y_4	y_4 x_4	y_4 x_4	y_4 x_4	x_4 y_4	y_4 x_4	y_4 x_4
x_5 y_5	x_5 y_5	y_5 x_5	y_5 x_5	y_5 x_5	y_5 x_5	x_5 y_5	y_5 x_5

Numbers 1 – 32 on top of the permutations represent the permutation numbers

4. Permutation algorithms for correlation

When implementing the Spearman's rank correlation coefficient, the ranking of the two samples is done independently and the ranks so obtained retain the positions of their respective observations. Therefore, any exchange of observations in any pair will result in a fresh ranking of the two samples. It is therefore constrained to a given data set. When ties exist, the mean rank of the tied observations is assigned to each of the tied observations. Algorithm 1 depicts the procedure for generating ranks for the tied and untied observations as required by the Spearman's rank correlation coefficient. After independently sorting each sample in ascending order of magnitude, Algorithm 1 ranks the observations and also takes care of tied observations.

ALGORITHM 1: RANK OBSERVATIONS

1. Assign ranks to the variates of first sample (T)
2. Again assign ranks to the variates of second sample (S)

$jt \leftarrow 1$
While $jt < k$ do (k is sample size)
  $it \leftarrow jt$
  $jt \leftarrow it + 1$
  $SumRanks \leftarrow ndx_{it}$
  $Counter \leftarrow 1$
  While $T_{it} \leftarrow T_{jt}$ do
    $SumRanks \leftarrow SumRanks + ndx_{jt}$
    $Counter \leftarrow Counter + 1$
    $jt \leftarrow jt + 1$
  end while
  if $Counter > 1$ then
    $RankMean \leftarrow SumRanks/Counter$
    for $jj \leftarrow it, jt - 1$ do
      $ndx_{jj} \leftarrow RankMean$
    end for
  end if
end while

The Algorithm(Paired-Permutation) of Odiase and Ogbonmwan [13] is adapted for correlation coefficient in Algorithm 2.

ALGORITHM 2: PERMUTATION DISTRIBUTION OF CORRELATION COEFFICIENT

Input data (X_{ij}; $i = 1(1)n$, $j = 1, 2$)
$i_1 \leftarrow 1, 2$ do
  $T_1 \leftarrow X_{1,i_1}$
  if $i_1 \leftarrow 1$ then
    $S_1 \leftarrow X_{1,2}$
  else
    $S_1 \leftarrow X_{1,1}$
  end if
  for $i_2 \leftarrow 1, 2$ do
    $T_2 \leftarrow X_{2,i_2}$
    if $i_2 \leftarrow 1$ then
      $S_2 \leftarrow X_{2,2}$
    else
      $S_2 \leftarrow X_{2,1}$
    end if
    for $i_3 \leftarrow 1, 2$ do
      $T_3 \leftarrow X_{3,i_3}$
      if $i_3 \leftarrow 1$ then
        $S_3 \leftarrow X_{3,2}$
      else
        $S_3 \leftarrow X_{3,1}$
      end if

```
        for $i_4 \leftarrow 1, 2$ do
          $T_4 \leftarrow X_{4,i_4}$
          if $i_4 \leftarrow 1$ then
            $S_4 \leftarrow X_{4,2}$
          else
            $S_4 \leftarrow X_{4,1}$
          end if
          for $i_5 \leftarrow 1, 2$ do
            $T_5 \leftarrow X_{5,i_5}$
            if $i_5 \leftarrow 1$ do
              $S_5 \leftarrow X_{5,2}$
            else
              $S_5 \leftarrow X_{5,1}$
            end if
            ...
            for $i_{15} \leftarrow 1, 2$ do
              $T_{15} \leftarrow X_{15,i_{15}}$
              if $i_{15} \leftarrow 1$ do
                $S_{15} \leftarrow X_{15,2}$
              else
                $S_{15} \leftarrow X_{15,1}$
              end if
              Call $Algorithm 1$
              Compute $r$ and $r_s$
              Update Frequency Distribution of $r$ and $r_s$
            end for
          end for
        end for
      end for
    end for
      ...
  end for
```

The algorithm presented in this paper can carry out a complete enumeration of all the possible distinct n-paired permutations by making the necessary adjustments to reflect the number of pairs.

5. Results

The algorithms were implemented in Intel Visual FORTRAN. The paired permutation p-values generated for the Pearson's and the Spearman's correlation coefficients are presented in Table 2 along with their classical results for the heights of cross and self fertilized plants, see Darwin [3]. The algorithms can be applied to any sample size. The statistic of interest is computed each time a new permutation is generated and fused into the frequency distribution of the previously computed values of the test statistic. These results were generated via Algorithms 1 and 2 presented in this paper.

TABLE 2: P-values for correlation coefficients (1-tailed)

Correlation	Coefficient	Asymptotic p-value	Permutation p-value
Pearson	−0.33518	0.11356	0.09241
Speraman rank	−0.34375	0.09919	0.09308

The scatter diagram and the distribution of Spearman's correlation coefficient for heights of paired fertilized plants are displayed in Figures 1 and 2.

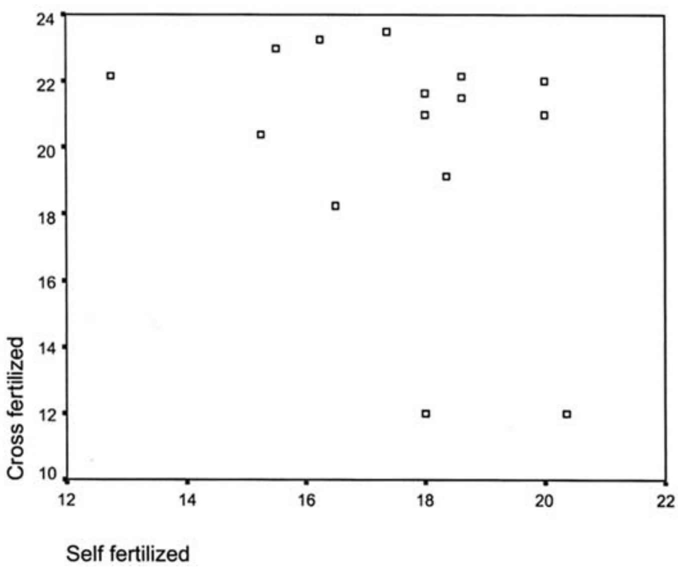

Fig. 1. Scatter diagram of the heights of paired fertilized plants

Critical values for the permutation distribution of the Pearson's and Spearman's rank correlation coefficient for the heights of paired fertilized plants are presented in Table 3 and Table 4.

6. Conclusion

Statistical test is based on calculating the test statistic of interest, comparing the calculated test statistic with a critical value and accepting or rejecting the null hypothesis based on the outcome of the comparison. The critical values are usually determined by cutting off the most extreme $100\alpha\%$ of the theoretical frequency distribution of the test statistic, where α is the level of significance, see Siegel and Castellan [17].

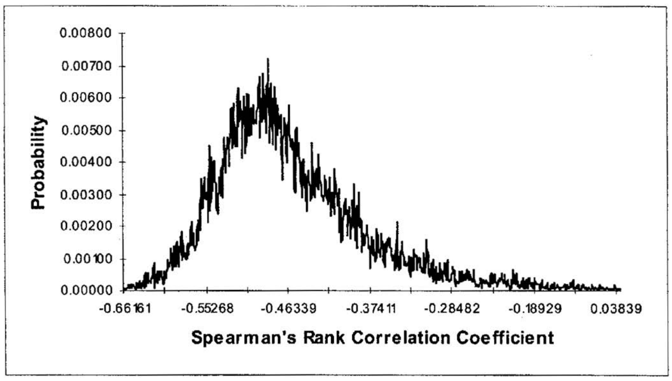

Fig. 2. Distribution of Spearman's correlation coefficient for heights of paired fertilized plants

TABLE 3: Lower critical values C_α for r and r_s
(If $\alpha' \leq \alpha$, then $C_\alpha = C_\alpha^{\geq}$; if $\alpha' > \alpha$, then $C_\alpha = C_\alpha^{>}$)

Correlation coefficient	α							
	0.001		0.0025		0.005		0.01	
	C_α	α'	C_α	α'	C_α	α'	C_α	α'
Pearson(r)	-0.5501	0.0010	-0.5352	0.0025	-0.5207	0.0050	-0.5051	0.0100
Spearman(r_s)	-0.6375	0.0010	-0.6241	0.0025	-0.6161	0.0049	-0.6027	0.0101

Correlation coefficient	α					
	0.025		0.05		0.1	
	C_α	α'	C_α	α'	C_α	α'
Pearson(r)	-0.4819	0.0250	-0.4616	0.0500	-0.4397	0.1000
Spearman(r_s)	-0.5848	0.0247	-0.5670	0.0510	-0.5491	0.0999

TABLE 4: Upper critical values C_α for r and r_s
(If $\alpha' \leq \alpha$, then $C_\alpha = C_\alpha^{\geq}$; if $\alpha' > \alpha$, then $C_\alpha = C_\alpha^{>}$)

Correlation coefficient	α							
	0.001		0.0025		0.005		0.01	
	C_α	α'	C_α	α'	C_α	α'	C_α	α'
Pearson(r)	0.0137	0.0010	-0.0739	0.0025	-0.1335	0.0050	-0.1911	0.0100
Spearman(r_s)	-0.0116	0.0010	-0.0848	0.0025	-0.1232	0.0049	-0.1821	0.0100

Correlation coefficient	α					
	0.025		0.05		0.1	
	C_α	α'	C_α	α'	C_α	α'
Pearson(r)	-0.2561	0.0250	-0.3011	0.0500	-0.3398	0.1000
Spearman(r_s)	-0.2420	0.0250	-0.2991	0.0499	-0.3500	0.1001

The p-values presented in Table 2 are consistently smaller for the exact permutation approach, both for the Pearson and for the Spearman, indicating that the probability of a type I error is more than α for the classical approaches. It is therefore advisable that the permutation test should be employed whenever possible.

The critical values displayed in Tables 2 and 3 clearly reveal that correlation analysis can easily be handled by the permutation approach. Without difficulty, a nonparametric confidence interval can be constructed for the exact permutation distribution generated. This can be obtained exactly the same way bootstrap confidence intervals are obtained, see Efron [4].

REFERENCES

[**1**] Agresti, A., *A survey of exact inference for contingency tables*, Statistical Science, **7** (1992), 131–177.

[**2**] Bayarri, M. J. and Berger, J. O., *The interplay of Bayesian and frequentist analysis*, Statistical Science, **19** (2004), 58–80.

[**3**] Darwin, C., *The Effects of Cross and Self Fertilization in the Vegetable Kingdom*, 2nd edition, London: John Murray, 1878.

[**4**] Efron, B., *The jackknife, the bootstrap, and other resampling plans*, CBMS-NSF Reg. Conf. Ser. Appl. Math. **38** (1982), 92.

[**5**] Fisher, R. A., *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.

[**6**] Freund, J. E., *Modern Elementary Statistics* (5th edition), Prentice-Hall Inc., Englewood Cliff, N. J., 1979.

[**7**] Good, P., *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses* (2nd edition), Springer Verlag, New York. 2000.

[**8**] Hilton, J. F., *A new asymptotic distribution for Hollander's bivariate symmetric statistic*, Computational Statistics and Data Analysis, **32** (2000), 455–463.

[**9**] Hilton, J. F. and Gee, L., *The size and power of the exact bivariate symmetry test*, Computational Statistics and Data Analysis, **26** (1997), 53–69.

[**10**] Hoeffding, W., *Large sample power of tests based on permutations of observations*, The Annals of Mathematical Statistics, **23** (1952), 169–192.

[**11**] Hollander, M., *A nonparametric test for bivariate symmetry*, Biometrika, **58** (1971), 203–212.

[**12**] Odiase, J. I. and Ogbonmwan, S. M., *An algorithm for generating unconditional exact permutation distribution for a two-sample experiment*, Journal of Modern Applied Statistical Methods, **4** (2005), 319–332.

[**13**] Odiase, J. I. and Ogbonmwan, S. M., *Exact permutation algorithm for paired observations: The challenge of R. A. Fisher experiment*, Journal of Mathematics and Statistics, **3**(3) (2007), 116–121.

[**14**] Opdyke, J. D., *Fast permutation tests that maximize power under conventional Monte Carlo sampling for pairwise and multiple comparisons*, Journal of Modern Applied Statistical Methods, **2** (2003), 27–49.

[**15**] Pesarin, F., *Multivariate Permutation Tests*, Wiley, New York, 2001.

[**16**] Scheffe H., *Statistical inference in the nonparametric case*, The Annals of Mathematical Statistics, **14** (1943), 305–332.

[**17**] Siegel, S. and Castellan, N. J., *Nonparametric Statistics for the Behavioral Sciences* (2nd edition), McGraw-Hill, New York, 1988.

[**18**] Spiegelhalter, D. J., *Incorporating Bayesian ideas into health-care evaluation*, Statistical Science, **19** (1988), 156–174.

(received 20.03.2007, in revised form 26.08.2007)

Department of Mathematics, University of Benin, Benin City, Nigeria

E-mail: `justiceodiase@yahoo.com`