

TESTS FOR TIME-SPACE CLUSTERING OF DISEASE

by

Nguyen Van Dat

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1409

July 1982

TESTS FOR TIME-SPACE CLUSTERING OF DISEASE

by

Nguyen Van Dat

A dissertation submitted to the faculty of
the University of North Carolina at Chapel
Hill in partial fulfillment of the require-
ments for the degree of Doctor of Philosophy
in the Department of Biostatistics

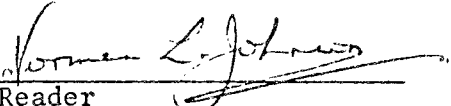
Chapel Hill

1982

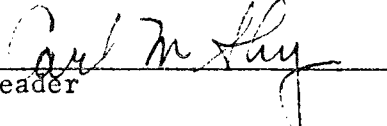
Approved by



Advisor



Reader



Reader

NGUYEN VAN DAT. Tests for Time-Space Clustering of Disease (Under the direction of ROGER C. GRIMSON.)

Several statistical tests for time-space clustering of disease have been suggested by different authors. Some commonly used tests are reviewed and their differences and similarities are discussed. Two alternative new tests that do not have some of the drawbacks of the others are suggested: the number-of-empty-cells test and the zero-one matrix test.

The exact distribution of the number-of-empty-cells test statistic is derived, together with the exact moments of the zero-one matrix test statistic. Since these moments are too complicated for most practical testing situations, a way to approximate the first and second moments is suggested, based on the exact moments calculated for a wide range of the parameters. Using these approximations and the asymptotic normal property of the test statistic, the approximate test can be carried out with relative ease.

The performance of the zero-one matrix test is compared with its counterparts in the literature: the EMM test and the scan test. Examples drawn from North Carolina mortality data are given to illustrate situations in which one test is better than the others.

Generalizations of the zero-one matrix test are also suggested to adjust for extraneous factors and for multivariate applications, using maximum likelihood procedures and Euclidean distances respectively.

Finally, a summarized step-by-step practical guide for the test of time or space clustering is presented together with some suggestions for further research.

ACKNOWLEDGEMENTS

I would like to express my appreciation to my advisor, Dr. Roger C. Grimson, for his initial suggestion of the topic and more importantly, for his enthusiastic guidance and continuous support in the course of this research. Gratitude is also expressed to Drs. Michael D. Hogan, Norman L. Johnson, Carl M. Shy and Michael J. Symons, who served on my advisory committee.

I would also like to thank Dr. Lawrence L. Kupper for his guidance and support during my first academic experience at the University of North Carolina, and Dr. William Mendenhall III for his advice during the first years in my statistics career at the University of Florida and his suggestion that I pursue a biostatistic training program at the University of North Carolina.

For the skillful typing of the manuscript, I must thank Ms. Ruth Bahr, whose work can only be appreciated by those who have seen my handwriting.

The valuable education which I obtained from the faculty of the Department of Biostatistics is gratefully acknowledged.

Last but not least I would like to thank my friends and my family from half way around the earth for their encouragement and moral support which helped me maintain a semblance of sanity which has made this work all possible.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	ii
LIST OF TABLES	v
LIST OF FIGURES AND PLOTS	vi
LIST OF APPENDICES	vii
CHAPTER	
I INTRODUCTION AND REVIEW OF THE LITERATURE	1
1.0 Introduction	1
1.1 Literature review	2
1.1.1 Pinkel and Nefzger's test	2
1.1.2 Knox's test	3
1.1.3 Barton and David's test	5
1.1.4 The EMM procedure	5
1.1.5 Grimson's cluster model	6
1.1.6 Mantel's test	7
1.1.7 The scan test	8
1.1.8 Bailar, Eisenberg and Mantel's test	9
II INTRODUCTION OF SOME NEW TESTS	11
2.0 Introduction	11
2.1 Number of empty cells test	11
2.2 Zero-one matrix test	15
2.2.1 Description of test	15
(i) Test for time clustering	16
(ii) Test for space clustering	17
(iii) Test for time-space interaction	18
(iv) General properties of the zero-one matrix test	18
2.2.2 Exact distribution of the test statistic for time clustering	20
2.2.3 Adjustment for better approximation (continuity correction)	24
2.2.4 Simulation results	24

CHAPTER

III	REGRESSION ESTIMATION FOR THE ZERO-ONE MATRIX TEST FOR TIME CLUSTERING	26
	3.0 Introduction	26
	3.1 Binomial estimates versus exact values	26
	3.2 Estimate the variance using the expected value	28
	3.3 Estimate the expected value	29
	3.4 Appropriateness of the regression estimation	31
IV	COMPARING THE ZERO-ONE MATRIX TEST FOR TIME CLUSTERING WITH THE EMM TEST AND THE SCAN TEST	43
	4.0 Introduction	43
	4.1 Situations in which the zero-one matrix test is more powerful than the EMM test	43
	4.2 Situations in which the EMM test is more power- ful than the zero-one matrix test	46
	4.3 Zero-one matrix test for time clustering ver- sus the scan test	48
V	GENERALIZED ZERO-ONE MATRIX TEST FOR TIME CLUSTER- ING WITH ADJUSTMENT FOR EXTRANEIOUS FACTORS; MULTI- VARIATE APPLICATIONS	50
	5.0 Introduction	50
	5.1 Zero-one matrix test for time clustering with adjustment for extraneous factors	50
	5.2 Zero-one matrix test in multivariate cases	56
	5.2.1 Classifying data points into two groups	56
	5.2.2 Assessing the significant level of the test	58
VI	SUMMARY, PRACTICAL GUIDE, AND SUGGESTIONS FOR FURTHER RESEARCH	68
	6.0 Summary	68
	6.1 Practical guide	69
	6.2 Suggestions for further research	75
	APPENDICES	77
	REFERENCES	93

LIST OF TABLES

Table	Page
3.1	Approximate and exact values of $E(A)$ and $\text{Var}(A)$ for $n=5(1)10,12,15,18,20(5)50$ and $m=5(1)10$ 33
3.2	Means of the ratio and of the difference between the approximate and the exact moments by number of cells, based on results presented in Table 3.1 37
3.3	Regressing the exact variance on the exact ex- pected value of the test statistic A - Test for time clustering 38
3.4	Regressing the exact expected value of A on the number of time units m and the decimal part of $\left(\frac{n}{m}\right)$ minus 0.5 39
3.5	Regressing α (probability that $n_i \geq \left[\frac{n}{m} - 0.5\right]$) on the decimal part of $\left(\frac{n}{m}\right) - 0.5^i$ 40
5.1	Mortality statistics for 1980 - North Carolina residents - Diseases of the heart 63
5.2	Mortality statistics for 1980 - North Carolina residents - Cancer 64
5.3	1980 heart disease and cancer death rate in North Carolina by Health Service Region 65
5.4	Euclidean distance between 2 regions in North Carolina based on heart disease and cancer death rates 66

LIST OF FIGURES AND PLOTS

	Page
Plot 3.1 RE1 = Residual of the expected value in %	41
Plot 3.2 RE2 = Residual of the variance in %	42
Fig. 5.1 North Carolina counties and eight health service regions	62
Fig. 5.2 Dendrogram from single linkage clustering on Euclidean distance of heart disease and cancer death rates in the eight health service regions in North Carolina	67

LIST OF APPENDICES

Appendix	Page
1.1 Program to use the IMSL subroutine for Incomplete Beta function	77
1.2 FORTRAN program to calculate Sum-2 which is being used for $I_p^2(n,r)$	80
1.3 SAS Program to calculate $I_p^2(r,n)$ from Sum-2	83
2.1 Standardized residuals from the regression estimate of the expected value and variance of A	86

CHAPTER I

INTRODUCTION AND REVIEW OF LITERATURE

1.0 Introduction

In recent literature, many epidemiologic studies are found attempting to attach significance to what seems to be a large number of cases of a given disease in some particular place and/or at a given time. Recent examples of "clusters" of diseases reported included the case of the town of Rutherford, New Jersey, where public health authorities became alarmed when it was found that the community of 20,000 had had 32 cases of leukemia, lymphoma and Hodgkin's disease in a five year period. Residents in Utah are claiming an unusually high cancer rate in that state related to nuclear testing in the area approximately twenty years ago. The Center for Disease Control has published numerous reports on "clusters" of disease. Residents of the area surrounding the Three-Mile Island nuclear facility in Pennsylvania are being followed for evidence of unusual "clusters" of health problems.

While evidence that a particular disease does cluster in time and/or space would not necessarily prove anything about its etiology, it could provide clues in the development of theories concerning the cause of that disease or support existing theories. Recently many important studies have sought to link disease clusters to environmental factors.

Unlike many infectious diseases where time-space clusters are easily recognized due to their relatively confined and isolated patterns, chronic

diseases with lower incidence rates and longer latent periods present special difficulties in ascertaining patterns.

Several different statistical tests have been suggested by different authors in an attempt to deal with this problem. As part of this study these tests will be reviewed and their similarities and differences will be discussed.

Unlike the branch of multivariate statistical methods generally known as cluster analysis, time-space clustering methods do not seek to define clusters. Rather they test for the presence of a time cluster, a space cluster or a time-space interaction effect. The analysis is an attempt to determine whether disease cases tend to be closer together in time and/or closer together in space than one would expect if the cases occurred randomly.

While the methods have typically been applied to disease data, many other applications have been conceived using these techniques.

1.1 Literature review

As a review of the literature, the following tests have been recommended by different authors:

1.1.1 Pinkel and Nefzger's test: Pinkel and Nefzger (1959) recommended a test using a classical combinatorial problem: if r cases are classified into n cells, the probability of placing k or more cases in cells that contain at least one other case is:

$$P(k) = \frac{\binom{n}{r-k} \binom{r-1}{k}}{\binom{n+r-1}{r}}$$

To perform the test the authors recommend dividing the study area and time period into time-space units (subjectively) and observing the maximum

number of cases in a time-space unit $k_0 + 1$, then calculating the P-value of the test by:

$$P = \sum_{k=k_0}^{r-1} P(k) .$$

One concludes that there is a time-space cluster if $P < \alpha$ where α is the level of the test.

This test has been criticized as more sensitive to clustering in time or space alone and requires a uniform population distribution (see Mantel (1967)). It is not clear how the authors derived the combinatorial formula for $P(k)$.

1.1.2 Knox's test: In his 1963 paper Knox suggested a test using a contingency table approach. In this paper he divided the space distances under study into 5 categories and the range of 800 days under study into 9 intervals and formed a 5×9 contingency table, the cell entries being the number of pairs falling within the appropriate time-space unit. Although Knox did analyze the table using the usual chi-square test he indicated that this is not really appropriate since the cell entries are not independent. Abe (1973) has derived the appropriate way to analyze such a table. Instead of using the Knox statistic,

$$K = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - \ell_{ij})^2}{\ell_{ij}} ,$$

as a chi-square statistic with $(r-1)(c-1)$ degrees of freedom, where

O_{ij} = observed entries of the contingency table,

ℓ_{ij} = expected value of O_{ij} ,

Abe argued that an appropriate chi-square statistic is:

$$Q = \underline{X}' \underline{V}^{-1} \underline{X} ,$$

where the elements of $\underline{\chi}^2$ (a row vector) are the deviations $(O_{ij} - e_{ij})$ of the elements of Knox's table from their expectation, and \underline{V}^{-1} is the inverse of the variance-covariance matrix. This statistic has been shown to have the correct expectation of a chi-square distribution. In a paper of the next year (1964-b) Knox analyzed the problem in a 2×2 contingency table by classifying each of the $\binom{n}{2}$ possible pairs of cases as close or far apart in space and close or far apart in time as the following 2×2 table shows.

		Space		Total
		close	far	
Time	close	a	b	a+b
	far	c	d	c+d
Total		a+c	b+d	$\frac{n(n-1)}{2}$

The statistic a is the number of pairs which are close in both time and space. Statistical significance is calculated by treating the null distribution of this statistic as Poisson, with expectation calculated from the marginal totals in the usual manner. The Poisson assumption is only appropriate when such pairs could be considered rare events. In this case, this test is considered a good approximation test, more sensitive to time-space clustering (interaction) than clustering in time or space alone. Alternatives to the Poisson approximation is the normal approximation for large numbers or the permutation distribution for small numbers. Knox's method was extended by Pike and Smith (1968) to take into account assumed periods of infectivity and susceptibility of cases. This allows for investigations of diseases with long latent periods.

Roberson (1979) has shown that the choice of boundaries for "far" and "close" criteria has some effect on the test outcome.

1.1.3 Barton and David's test: Barton, et al. (1965) and David and Barton (1966) introduced a procedure analogous to ANOVA methods. They divided the time span of study into subintervals and calculated the statistic as the average squared spatial distance between all points within subintervals divided by the overall average squared spatial distance between all points. Under the null hypothesis of no clustering, this statistic has an expected value of 1, while if clustering is present the expected value is less than 1. The significance level of this statistic is determined by comparing the observed value with the permutation test derived by pairing each location in time with all possible locations in space.

The authors also discussed the approximation to the permutation distribution, using a β -function which in turn can be approximated by the normal function when the number of time subintervals is large ($k-1 \geq 3\sigma_Q n$) or by the F distribution when the number of time subintervals is small ($k-1 < 3\sigma_Q n$) where

k = number of time subintervals,

n = number of observations,

Q = test statistic, and

σ_Q^2 = variance of Q .

The major drawbacks to using this test are the lack of criteria for determining the subintervals, the difficulty in computing the significance level of the test statistic, and the fact that the smaller distances which are of greater interest have less influence on the value of the statistic than the larger distances.

1.1.4 The EMM procedure: Ederer, Myers and Mantel (1964) recommended a procedure for time clustering. The time period under study is divided into

k subintervals over which n cases of the disease are observed. The test statistic is a , the maximum number of cases in a subinterval. The null distribution of a is derived by a combinatorial procedure; its mean and variance, along with the p -value, can be calculated, based on this distribution. The procedure was also extended to test for time-space clustering by dividing both time of study and geographic region into subintervals called time-space subunits. Each time unit has a fixed number of time-space subunits, m .

Let a_i be the maximum number of cases in a subunit of the i^{th} unit. Then the statistic is a chi-square statistic with 1 degree of freedom.

$$\chi^2 = \frac{\left(\left| \sum_{i=1}^k a_i - E\left(\sum_{i=1}^k a_i \right) \right| - .5 \right)^2}{\sum_{i=1}^k \text{Var}(a_i)}, \quad i=1,2,\dots,k,$$

where $E(a_i)$ and $\text{Var}(a_i)$ are determined by assuming that the null distribution of a_i is that of the maximum of a multinomial distribution.

Mantel, Kryscio and Myers (1976) tabulated tables for the expected value and variance of a_i for some selected values of n cases and m subunits in each unit.

The disadvantage of this test is its sensitivity to time or space clustering alone as well as time-space interaction.

1.1.5 Grimson's cluster model: Grimson (1979) suggested a model using the procedure similar to the EMM test. This model considers only the number of cases; it does not take into account the fact that cases are distinct. The model uses number of cases in a time-space subunit rather than the arrangement of cases in a unit.

Let a_j be the number of cases in subunit j ($j=1,2,\dots,k$); then, based on the composition of the number of cases r ($r = \sum_{j=1}^k a_j$),

$$P\{\max(a_j) \geq m \mid r = \sum_{j=1}^k a_j\} = 1 - \frac{\sum_{j=0}^k (-1)^j \binom{k}{j} \binom{k+r-jm-1}{k-1}}{\binom{r+k-1}{k-1}}$$

Based on this composition model the limit distribution of the test statistic, $\max(a_j)$, was proved to be geometric rather than Poisson, as in the multinomial model of the EMM test. Expected value and variance of the test statistic was computed for the case $k=5$ and $r=2$ through 400.

This approach was used in modeling hepatitis data.

1.1.6 Mantel's test: Mantel (1967) gave a generalized regression approach which contains both Knox and Barton-David procedures as special cases. He suggested a statistic of the form:

$$\sum_{i < j} h(x_{ij})g(y_{ij}) ,$$

where

x_{ij} = spatial distance between cases i and j

y_{ij} = temporal distance between cases i and j

$h(\cdot)$ and $g(\cdot)$ are transformation functions for spatial distance and temporal distance respectively .

The exact p-value for an observed statistic is obtained from the permutation distribution derived by evaluating the statistic for every possible pairing of time and space location. For even a moderately small sample size this is not simple to compute. Mantel gave some general formulae for the permutation mean and variance of the statistic. In most cases, however, it may be more practical to use Monte-Carlo methods to estimate p values as suggested by Besag and Diggle (1977), or to use the asymptotic distribution as an approximation. The choice of the functions $h(\cdot)$ and $g(\cdot)$ were also discussed. Mantel favored the choice of transformations which emphasize closeness rather than great distances, as this should increase the

power of the test to detect clustering if it is present. He suggested having the temporal and spatial transformation functions as the reciprocals of the absolute time and space distances respectively. To avoid problems of zero distances he suggested adding a small constant to each distance. No criteria were set by the author for selecting the values of these constants, and their values do effect the results of the analysis as discussed by Roberson (1979).

Mantel's procedure uses, where c and c' are the constants,

$$h(x_{ij}) = \frac{1}{c + |x_{ij}|} ,$$

$$g(y_{ij}) = \frac{1}{c' + |y_{ij}|} .$$

Note that if we let

$$h(x_{ij}) = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ and } j^{\text{th}} \text{ cases are close in space,} \\ 0 & \text{otherwise,} \end{cases}$$

$$g(y_{ij}) = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ and } j^{\text{th}} \text{ cases are close in time,} \\ 0 & \text{otherwise,} \end{cases}$$

then $a = \sum_{i < j} h(x_{ij})(y_{ij})$ is the Knox statistic. Mantel also pointed out that the Knox's test with the significance level evaluated using its permutation distribution or normal approximation probably is more applicable in many situations than the one evaluated by using the Poisson distribution as suggested by Knox.

Asymptotic properties as well as Monte-Carlo results on robustness of this test and the Knox test were also investigated by Roberson (1979).

1.1.7 The Scan test: In this test, the maximum number of cases in a fixed interval of time within a fixed spatial unit was used as the test statistic.

This statistic was first investigated in detail by Naus (1965, 1966) as he derived the probability distribution and the expectation and variance of this statistic. Later, Wallenstein (1980) investigated this statistic and tabulated the lower tail of the distribution of this statistic for some selected values of n (scan statistic), N (total number of disease cases), and L (time period under study divided by the length of the time interval of the test statistic). Naus (1982) suggested some approximations for the distribution of the scan statistics.

This test is probably more powerful than the EMM test and does not need a criterion for dividing time under study into fixed intervals. However, the test statistic is not easily assessed and due to the complexity of its distribution, its significance level (p-value) is much more difficult to evaluate than that of the EMM statistic for most practical situations.

1.1.8 Bailar, Eisenberg and Mantel's test: Bailar, Eisenberg and Mantel (1970) introduced the following test:

Let n be the total number of disease cases,

N be the number of years (time units) over which
 n disease cases occur,

p_i be the probability that a case occurs in the i^{th} year
(time unit),

n_i be the number of cases observed in the i^{th} year
(time unit)

$$\left(\sum_{i=1}^N n_i = n \right) \quad i=1,2,\dots,N.$$

Then $(n_i \cdot n_{i+d})$ = number of pairs formed where the first case
occurs in i^{th} year and the second case occurs
in $(i+d)^{\text{th}}$ year.

The statistic is $t =$ total number of pairs occurring in years exactly
 d years apart,

$$t = \sum_{i=1}^{N-d} n_i n_{i+d}$$

(d=1 or d=period of infectivity and susceptibility.)

When $P_i = p_{i-d} + p_{i+d}$ and $p_t = 0$ for $t \leq 0$ or $t > N$, it was proved that

$$E(t) = \frac{n(n-1)}{2} \sum_{i=1}^N p_i P_i$$

$$\begin{aligned} \text{Var}(t) = & \frac{n(n-1)}{2} \sum_{i=1}^N p_i P_i + n(n-1)(n-2) \sum_{i=1}^N p_i P_i^2 \\ & - \frac{n(n-1)}{2} (2n-3) \left(\sum_{i=1}^N p_i P_i \right)^2 . \end{aligned}$$

The p-value of this test statistic is evaluated by using the normal distribution as an approximation, with the above expected value and variance.

This test has not been used widely and the appropriateness of the approximation has not been investigated. It seems to be appropriate for testing of clustering in time or for testing of a latent period of length d.

CHAPTER II

INTRODUCTION OF SOME POSSIBLE NEW TESTS

2.0 Introduction

Although many different tests have been suggested, these tests contain the following drawbacks:

- They involve some subjective steps to determine the criteria which would affect the outcome of the test.
- While they are mainly sensitive to time-space clustering, they tend to be somewhat sensitive to clustering in time or space.
- The distributions of the test statistics are complicated and the significance levels cannot be assessed easily in most practical situations even with approximations.
- They are only a one tail test of clustering.

In an attempt to overcome these drawbacks, the following tests are suggested:

- Number-of-empty-cells test for rare events.
- Zero-one matrix tests for events with larger frequencies.

2.1 Number-of-empty-cells test:

Most of the derivations of the statistical formulas of this section are known (see Feller, Johnson and Kotz, and Riordan). This section will show how an occupancy distribution of the number of empty cells may be used as a cluster test.

After dividing the time and space under study into time-space units, similar to what is done in the EMM test, the statistic of interest is X , the number of units without disease cases. Example: dividing 10 years of North

Carolina mortality data into 1 year time units and 100 counties (space units) will provide 1000 time-space subunits.

Let n be the total number of cases occurring in m time-space units. Each unit has a_j cases, $j = 1, 2, \dots, m$; $0 \leq a_j \leq n$; $\sum_{j=1}^m a_j = n$.

If X is the number of units with no cases, then under the null hypothesis of no clustering we have, for the special case of $X=0$,

$$P(X=0) = \sum_{i=0}^{m-1} (-1)^i \binom{m}{i} \left[1 - \frac{i}{m}\right]^n \quad (\text{for } n \geq m) \quad (2.1)$$

$$\{\text{when } n < m \quad P(X=0) = 0.\}$$

To see this, assume equal probability of having cases in any unit.

If E_j is the event that unit j has no cases, then

$$P(E_j) = \left(1 - \frac{1}{m}\right)^n = \left(\frac{m-1}{m}\right)^n,$$

$$P(E_j \cap E_i) = \left(1 - \frac{2}{m}\right)^n,$$

.....

$$\begin{aligned} \text{therefore, } P(X=0) &= 1 - P\left(\bigcup_{j=1}^m E_j\right) \\ &= 1 - \sum_{j=1}^m P(E_j) + \sum_{j < j'} P(E_j \cap E_{j'}) - \dots \end{aligned}$$

$$\text{Note that } P\left(\bigcap_{j=1}^m E_j\right) = 0.$$

$$\text{Therefore, } P(X=0) = \sum_{i=0}^{m-1} (-1)^i \binom{m}{i} \left(1 - \frac{i}{m}\right)^n. \quad \text{This establishes equation (2.1).}$$

Next, let $m \geq k \geq \text{Max}(0, m-n)$. Then

$$P(X=k) = \binom{m}{k} \sum_{j=0}^{m-k-1} (-1)^j \binom{m-k}{j} \left(1 - \frac{k+j}{m}\right)^n \quad (2.2)$$

$$= \sum_{j=0}^{m-k-1} (-1)^j \binom{k+j}{k} \binom{m}{k+j} \left(1 - \frac{k+j}{m}\right)^n. \quad (2.3)$$

Proof of (2.3) is as follows:

$$\begin{aligned}
 P(X=k) &= \binom{m}{k} P\left(\begin{array}{l} \text{First } k \text{ cells} \\ \text{are empty} \end{array} \cap \begin{array}{l} \text{Last } m-k \text{ cells} \\ \text{are occupied} \end{array}\right) \\
 &= \binom{m}{k} P\left(\begin{array}{l} \text{All cases are in the last } m-k \text{ cells} \\ \text{and all these } m-k \text{ cells are occupied} \end{array}\right)
 \end{aligned}$$

Applying the inclusion-exclusion principle to the last $m-k$ cells,

$$P(X=k) = \binom{m}{k} \sum_{j=0}^{m-k-1} (-1)^j \binom{m-k}{j} \left(1 - \frac{k+j}{m}\right)^n.$$

Note that $\binom{m}{k+j} \binom{k+j}{k} = \binom{m}{k} \binom{m-k}{j}$, so that

$$P(X=k) = \sum_{j=0}^{m-k-1} (-1)^j \binom{k+j}{k} \binom{m}{k+j} \left(1 - \frac{k+j}{m}\right)^n,$$

which is equation (2.3). The expected value of X is given by

$$E(X) = m \left(1 - \frac{1}{m}\right)^n.$$

To see this let

$$X_j = \begin{cases} 0 & \text{if } j^{\text{th}} \text{ cell is occupied,} \\ 1 & \text{if } j^{\text{th}} \text{ cell is empty.} \end{cases}$$

$$\text{Then } E(X_j) = \left(\frac{m-1}{m}\right)^n.$$

$$E(X) = \sum_{i=1}^m E(X_i) = m \left(1 - \frac{1}{m}\right)^n. \quad (2.4)$$

Similarly,

$$\text{Var}(X) = m(m-1) \left(1 - \frac{2}{m}\right)^n + m \left(1 - \frac{1}{m}\right)^n - m^2 \left(1 - \frac{1}{m}\right)^{2n}. \quad (2.5)$$

The descending factorial moments of X are given by:

$$\begin{aligned}
 \mu_{(R)}(X) &= E\{X(X-1) \dots (X-R+1)\} = \binom{m-R}{m}^n m^{(R)} \\
 &= E(X^{(R)}). \quad (2.6)
 \end{aligned}$$

Equation (2.6) is derived from equation 2.2, where for any

$$m \geq k \geq \max(0, m-n),$$

$$1 = \sum_{k=\max(0, m-n)}^m \binom{m}{k} \sum_{j=0}^{m-k} (-1)^j \binom{m-k}{j} \left(\frac{m-k-j}{m}\right)^n.$$

If $n \geq m$ then $P(X=0) > 0$,

If $n < m$ then $P(X=0) = 0$.

Therefore,

$$\begin{aligned} E(X^{(R)}) &= \sum_{k=\max(0, m-n)}^m \binom{m}{k} \sum_{j=0}^{m-k} (-1)^j \binom{m-k}{j} \left(\frac{m-k-j}{m}\right)^n k^{(R)} \\ &= \sum_{k=\max(0, m-n)}^m m^{(R)} \binom{m-R}{k-R} \sum_{j=0}^{m-k} (-1)^j \binom{m-k}{j} \left[\left(\frac{m-k-j}{m-R}\right) \left(\frac{m-R}{m}\right)\right]^n. \end{aligned}$$

Note that when $0 \leq k < R$, $k^{(R)} = 0$, so that no contribution is made to the sum.

Let $k' = k - R$, $m' = m - R$. Then

$$E(X^{(R)}) = m^{(R)} \left(\frac{m-R}{m}\right)^n \sum_{k'=\max(0, m'-n)}^{m'} \binom{m'}{k'} \sum_{j=0}^{m'-k'} (-1)^j \binom{m'-k'}{j} \left(\frac{m'-k'-j}{m'}\right)^n$$

The double sum simplifies to unity (see (2.2)), so that

$$E(X^{(R)}) = m^{(R)} \left(\frac{m-R}{m}\right)^n.$$

Based on (2.2), the exact p-value for this test statistic can be computed readily.

This test is only good in situations where the disease incidence rates are not so high that we would expect some time-space units to have no disease cases ($E(X) > 0$). In order for $E(X) \doteq \frac{m}{2}$ (to have approximately half of the number of cells empty) we should have $0.65m < n < 0.7m$ for m between 10 and 1000; since $E(X) = m(1 - \frac{1}{m})^n$, in order for $E(X) = \frac{m}{2}$ we must have

$$m(1 - \frac{1}{m})^n = \frac{m}{2}$$

or

$$(1 - \frac{1}{m})^n = \frac{1}{2}$$

or

$$n = \frac{\log 2}{\log m - \log(m-1)}$$

for m = 10, n = 6.5788

for m = 100, n = 68.9675

for m = 1000, n = 692.8006

The test can be generalized by using the number of cells containing less than a certain number of cases, say a, as a test statistic.

This test will not be very powerful in detecting clustering if n is too large (or too small) relative to m, since in these situations most likely the test statistic will assume a value of zero (or (m-n)) under a random allocation of cases.

2.2 ZERO-ONE matrix test:

2.2.1 Description of the test: Divide the area and the time under study into time-space subunits to form a two-way table as follows:

		Time							Total	
		1	2	3	•	•	• j •	•		c
Space	1	n ₁₁	n ₁₂	n ₁₃	•	•	• n _{1j} •	•	n _{1c}	n _{1•}
	2	n ₂₁	n ₂₂	n ₂₃	•	•	• n _{2j} •	•	n _{2c}	n _{2•}
	3	n ₃₁	n ₃₂	n ₃₃	•	•	• n _{3j} •	•	n _{3c}	n _{3•}
	•	•	•	•	•	•	•	•	•	•
	•	•	•	•	•	•	•	•	•	•
	•	•	•	•	•	•	•	•	•	•
	i	n _{i1}	n _{i2}	n _{i3}	•	•	• n _{ij} •	•	n _{ic}	n _{i•}
	•	•	•	•	•	•	•	•	•	•
	•	•	•	•	•	•	•	•	•	•
	•	•	•	•	•	•	•	•	•	•
r	n _{r1}	n _{r2}	n _{r3}	•	•	• n _{rj} •	•	n _{rc}	n _{r•}	
Total	n _{•1}	n _{•2}	n _{•3}	•	•	• n _{•j} •	•	n _{•c}	n _{••}	

Let n_{ij} be the observed entry of the $(ij)^{\text{th}}$ time-space unit, where

$$\begin{cases} i = 1, 2, \dots, r \\ j = 1, 2, \dots, c, \end{cases}$$

and let

$$\begin{cases} n_{..} = \sum_{i=1}^r \sum_{j=1}^c n_{ij}, \\ n_{.j} = \sum_{i=1}^r n_{ij}, \\ n_{i.} = \sum_{j=1}^c n_{ij}. \end{cases}$$

(i) Test for time clustering:

$$\text{Let } a_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \frac{n_{i.}}{c}, \\ 1 & \text{if } n_{ij} \geq \frac{n_{i.}}{c}. \end{cases}$$

a_{ij} is distributed approximately like a Bernoulli distribution with $p = \frac{1}{2}$. The test statistic, denoted by A , is defined by

$$A = \sum_{i=1}^r \sum_{j=1}^c a_{ij}.$$

Note that a_{ij} and $a_{i'j}$ are not independent; however, under the null hypothesis H_0 of no time-clustering, a plausible approximate distribution of A is a binomial distribution with parameters $p = \frac{1}{2}$ and $n = rc$. From this binomial distribution the significance level of the test statistic could be assessed. When n is large, an approximate distribution of A is a normal distribution with mean equal to $\frac{rc}{2}$ and variance equal to $\frac{rc}{4}$.

Like the EMM test, this test may be referred to as a test of time-space clustering but it is sensitive to time clustering within spatial units. However, this test, by its nature, is more powerful than the EMM test in

situations where within each time unit there is more than one large cluster in different time-space subunits, since the EMM test uses only the maximum frequency in a single time-space subunit as the test statistic.

When the observed statistic is large, the test indicates cluster avoidance in some time-space subunits. (At certain times, some places have unusually few cases.) When the observed statistic is small it indicates some clustering in some time-space subunits. (At certain times some places have unusually many cases.)

(ii) Test for space clustering:

$$\text{Let } b_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \frac{n_{.j}}{r} , \\ 1 & \text{if } n_{ij} \geq \frac{n_{.j}}{r} . \end{cases}$$

The test statistic, denoted by B, is defined by:

$$B = \sum_{i=1}^r \sum_{j=1}^c b_{ij} .$$

Under the null hypothesis H_0 of no space clustering, the approximate distribution of B can be evaluated in the same manner as that of A.

As in test (i), this test is a type of time-space clustering test; however, it is sensitive to space clustering within temporal units.

The test requires that the populations under study in the time-space subunits are equal in size; this can be a severe restriction. However, when this requirement is not met, the frequencies n_{ij} 's need to be adjusted or the rates should be used instead of the frequencies. Example:

$$b_{ij} = \begin{cases} 0 & \text{if } r_{ij} < \frac{\sum_{i=1}^r r_{ij}}{r} \\ 1 & \text{if } r_{ij} \geq \frac{\sum_{i=1}^r r_{ij}}{r} \end{cases} \quad \text{where } r_{ij} \text{ is the observed disease rate at time-space subunit } ij.$$

(iii) Test for time-space 'interaction':

$$\text{Let } c_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} , \\ 1 & \text{if } n_{ij} \geq \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} . \end{cases}$$

The test statistic, denoted by C, is defined by

$$C = \sum_{i=1}^r \sum_{j=1}^c c_{ij} .$$

Under the null hypothesis of no time-space interaction, the approximate distribution of C can be evaluated in the same manner as that of A.

The test for time-space "interaction" is not sensitive to time or space clustering alone. This test, unlike the test for space clustering, does not require that the populations under study within time-space units be equal in size.

(iv) General properties of the zero-one matrix test:

The zero-one matrix test has the following advantages:

- The test can be used to test either time or space clustering alone or time-space interaction.
- The test statistic and its approximate distribution (to be derived) are simple; therefore the p-value of the test can be evaluated without complicated computation.
- The test is sensitive to cases with more than one cluster, which gives it an advantage over the EMM test, which depends only on the maximum.
- The rates (adjusted or unadjusted) can be used for this test instead of the numerators themselves, as required by other tests.
- The test can be used as a one- or two-tail test; for instance, when A is small it shows time clustering; when A is large it shows

cluster avoidance (vacuity) at some time-space subunit(s).

Example: If we have one space and 6 time units with a total of 30 disease cases, then we may have one of the following situations ($r=1$, $c=6$, $n_{..}=30$):

Time	1	2	3	4	5	6	Total
Expectation	5	5	5	5	5	5	30
Observation 1	4	6	7	2	8	3	30
Observation 2	10	0	6	10	4	0	30
Observation 3	10	0	10	0	0	10	30
Observation 4	10	4	4	4	4	4	30
Observation 5	6	6	6	6	6	0	30

$$A \sim \text{Bin}\left(\frac{1}{2}, 6\right)$$

$$E(A) = 3$$

$$\text{Var}(A) = 1.5$$

In observations 1, 2, 3, since $A=3=E(A)$, we do not reject H_0 of clustering or cluster avoidance, while in observation 4 we would reject H_0 in favor of clustering ($A=1$), and in observation 5 we would reject H_0 in favor of avoidance ($A=5$). Depending on how one defines clustering, one may argue that observations 2 and 3 should be considered as clustering. This test is not intended for detecting clustering of this type.

This test is not designed for testing clustering of extremely rare diseases, since in these cases most time-space units have no cases ($n_{ij}=0$ for most ij) and we will have a small test statistic (A is small) and hence the probability of committing a type II error is large. For the applications to rare diseases, the numer-of-empty-cells test is recommended.

However, as discussed in Section 2.1, the number-of-empty-cells test is also not very powerful in situations with extremely rare diseases.

In the above example, binomial and normal distributions are considered approximate distributions of the test statistic. Since a_{ij} 's are not independent of each other, the appropriateness of these approximations needs to be investigated further. As part of this study, this issue will be investigated further along with the exact distribution of the test statistic.

2.2.2 Exact distribution of the test statistic for time clustering:

The test statistic is

$$A = \sum_{i=1}^r \sum_{j=1}^c a_{ij} = \sum_{i=1}^r a_{i\cdot} ,$$

where

$$a_{i\cdot} = \sum_{j=1}^c a_{ij} .$$

Each $a_{i\cdot}$ is distributed independently (since a_{ij} only depends on the marginal $n_{i\cdot}$). Thus, for simplicity here we shall investigate the distribution of $a_{i\cdot}$. This is equivalent to the special case where $r=1$, $A = \sum_{j=1}^c a_j$.

$P(A=a)$ = Probability that among m cells a of them have frequencies of $\frac{n_{i\cdot}}{c}$ or larger.

$$\text{Let } I_p^a(\rho, \nu) = \frac{\Gamma(\nu+1)}{\Gamma^a(\rho)\Gamma(\nu+1-a\rho)} \int_0^p \cdots \int_0^p (1 - \sum_{j=1}^a x_j)^{\nu-a\rho} \prod_{j=1}^a x_j^{\rho-1} dx_j$$

= Incomplete (type I) Dirichlet Integral .

It has been shown (see Harter et al (1975)) that $I_p^a(\rho, \nu)$ is the probability that the minimum frequency of the first a cells of a multinomial which has $a+1$ cells is at least ρ , provided that the first a cells have a common cell probability of $p = \frac{1}{m}$ ($m > a$) and the total of the cell frequencies is ν .

Note the special case $I_p^1(\rho, \nu) = \text{Incomplete beta function } I_p(\rho, \nu - \rho + 1)$.

Applying the inclusion and exclusion principles it can be shown that for the case of 1 row ($r=1, m=r \times c=c$):

$$P(A=a) = \binom{m}{a} \sum_{j=0}^{m-a} (-1)^j \binom{m-a}{j} I_p^{(a+j)}(\rho, \nu) \quad (2.7)$$

$$\text{where } \rho = \frac{\nu}{c} = \frac{\sum_{j=1}^c n_j}{c} .$$

From this exact distribution we get the descending factorial moments of A:

$$E(A^{(k)}) = m^{(k)} I_p^k(\rho, \nu) . \quad (2.8)$$

$$E(A) = m I_p^1(\rho, \nu) = m I_p(\rho, \nu - \rho + 1) = m\alpha . \quad (2.9)$$

Note that the theory derived in this section does not apply if the population sizes are unequal and have been adjusted for (e.g., the rates were used instead of the frequencies).

The higher moments of A can be derived from (2.8) as follows:

$$\begin{aligned} E(A(A-1)) &= EA^2 - E(A) = m(m-1) I_p^2(\rho, \nu) = m(m-1)\beta . \\ E(A^2) &= m(m-1)\beta + m\alpha = m(\alpha + (m-1)\beta) \end{aligned} \quad (2.10)$$

Similarly,

$$\begin{aligned} E\{A(A-1)(A-2)\} &= E(A^3) - 3E(A^2) + 2E(A) = m(m-1)(m-2)\gamma \\ E(A^3) &= m(m-1)(m-2)\gamma + 3(m(m-1)\beta + m\alpha) - 2m\alpha \\ &= m\alpha + 3m(m-1)\beta + m(m-1)(m-2)\gamma . \end{aligned} \quad (2.11)$$

Similarly,

$$\begin{aligned} E\{A(A-1)(A-2)(A-3)\} &= E(A^4) - 6E(A^3) + 11E(A^2) - 6E(A) \\ &= m(m-1)(m-2)(m-3)\delta \end{aligned}$$

$$\begin{aligned} E(A^4) &= m(m-1)(m-2)(m-3)\delta + 6(m(m-1)(m-2)\gamma + 3m(m-1)\beta + m\alpha) \\ &\quad - 11(m(m-1)\beta + m\alpha) + 6m\alpha \\ &= m\alpha + 7m(m-1)\beta + 6m(m-1)(m-2)\gamma + m(m-1)(m-2)(m-3)\delta, \end{aligned} \tag{2.12}$$

where $\alpha = I_p^1(\rho, \nu)$, $\beta = I_p^2(\rho, \nu)$, $\gamma = I_p^3(\rho, \nu)$, $\delta = I_p^4(\rho, \nu)$.

In an attempt to find a good approximate distribution for the statistic A we need to compute the first 4 moments of A . To do this we need to have $I_p^{(b)}(\rho, \nu)$ for $b = 1, 2, 3, 4$. Tables are available for only some selected values of b, p, ρ and ν (Harter and Owen (1975)).

To compute $I_p^{(b)}(\rho, \nu)$ for $b = 1, 2, 3, 4$ note that for any $\nu > b\rho$ we have

$$I_p^{(b)}(\rho, \nu) = \frac{\nu!}{[(\rho-1)!]^b [v-b\rho]!} \int_0^p \cdots \int_0^p (1 - \sum_{i=1}^b x_i)^{\nu-b\rho} \prod_{i=1}^b x_i^{\rho-1} dx_i.$$

Let $m' = \nu - b\rho$ and $s = \rho - 1$; then the integrand may be expanded as follows:

$$\begin{aligned} (1 - \sum_{i=1}^b x_i)^{m'} \prod_{i=1}^b x_i^s &= \sum_{k=0}^{m'} (-1)^k \binom{m'}{k} \left(\sum_{i=1}^b x_i \right)^k \prod_{i=1}^b x_i^s \\ &= \sum_{k=0}^{m'} (-1)^k \binom{m'}{k} \sum_{i_1+i_2+\dots+i_b=k} x_1^{i_1} x_2^{i_2} \cdots x_b^{i_b} \left(\frac{k!}{i_1! i_2! \cdots i_b!} \right) x_1^s x_2^s \cdots x_b^s \\ &= \sum_{k=0}^{m'} (-1)^k \binom{m'}{k} \sum_{i_1+i_2+\dots+i_b=k} x_1^{s+i_1} x_2^{s+i_2} \cdots x_b^{s+i_b} \frac{k!}{i_1! i_2! \cdots i_b!} \\ &= \sum_{k=0}^{m'} (-1)^k \sum_{i_1+i_2+\dots+i_b=k} x_1^{s+i_1} x_2^{s+i_2} \cdots x_b^{s+i_b} \frac{m'!}{(m'-k)! i_1! i_2! \cdots i_b!} \end{aligned}$$

$$\begin{aligned}
I_p^b(\rho, \nu) &= \frac{\nu!}{[(\rho-1)!]^b (\nu-b\rho)!} \int_0^p \dots \int_0^p \\
&\quad \left(\sum_{k=0}^{m'} (-1)^k \sum_{i_1+i_2+\dots+i_b=k} x_1^{s+i_1} x_2^{s+i_2} \dots x_b^{s+i_b} \frac{m'!}{(m'-k)! i_1! i_2! \dots i_b!} \right) \prod_{i=1}^b dx_i \\
&= \frac{\nu!}{[(\rho-1)!]^b (\nu-b\rho)!} \sum_{k=0}^{m'} \\
&\quad \left((-1)^k \sum_{i_1+i_2+\dots+i_b=k} \frac{p^{k+b\rho} m'!}{(i_1+\rho)(i_2+\rho)\dots(i_b+\rho) \cdot (m'-k)! i_1! i_2! \dots i_b!} \right) \\
&= \frac{\nu!}{[(\rho-1)!]^b} \sum_{k=0}^{\nu-b\rho} (-1)^k p^{k+b\rho} \sum_{i_1+i_2+\dots+i_b=k} \frac{1}{(\nu-b\rho-k)! \prod_{j=1}^b [(i_j+\rho) i_j!]} \\
\end{aligned} \tag{2.13}$$

To compute the first 4 moments of A, (2.13) gives

$$I_p^1(\rho, \nu) = \frac{\nu! p^\rho}{(\rho-1)!} \sum_{x=0}^{\nu-\rho} \frac{(-p)^x}{(\rho+x)(\nu-\rho-x)! x!} = \text{Incomplete beta} = \alpha.$$

$$I_p^2(\rho, \nu) = \frac{\nu! p^{2\rho}}{[(\rho-1)!]^2} \sum_{x=0}^{\nu-2\rho} \left(\frac{(-p)^x}{(\nu-2\rho-x)!} \sum_{i=0}^x \frac{1}{(\rho+i)(\rho+x-i) i! (x-i)!} \right) = \beta.$$

$$I_p^3(\rho, \nu) = \frac{\nu! p^{3\rho}}{[(\rho-1)!]^3} \sum_{x=0}^{\nu-3\rho} \left(\frac{(-p)^x}{(\nu-3\rho-x)!} \sum_{\substack{i+j+k=x \\ i,j,k \geq 0}} \frac{1}{(\rho+i)(\rho+j)(\rho+k) i! j! k!} \right) = \gamma.$$

$$I_p^4(\rho, \nu) = \frac{\nu! p^{4\rho}}{[(\rho-1)!]^4} \sum_{x=0}^{\nu-4\rho} \left(\frac{(-p)^x}{(\nu-4\rho-x)!} \sum_{\substack{i+j+k+l=x \\ i,j,k,l \geq 0}} \frac{1}{(\rho+i)(\rho+j)(\rho+k)(\rho+l) i! j! k! l!} \right) = \delta.$$

If $m =$ number of cells $(m = \frac{1}{p})$

$$\mu_1' = E(A) = m\alpha$$

$$\mu_2' = E(A^2) = m\alpha + m(m-1)\beta$$

$$\mu_3' = E(A^3) = m\alpha + 3m(m-1)\beta + m(m-1)(m-2)\gamma$$

$$\mu_4' = E(A^4) = m\alpha + 7m(m-1)\beta + 6m(m-1)(m-2)\gamma + m(m-1)(m-2)(m-3)\delta$$

Using these 4 moments a Pearson-type curve could be selected to be a good approximate distribution of A (see Pearson et al (1962)).

2.2.3 Adjustment for better approximation (continuity correction):

As presented by (2.7), the exact distribution is not in a simple form and it cannot be used conveniently in most practical situations. Usually we will have to rely on some form of approximation.

It was found from the following simulation results that the distribution of $A = \sum_{i=1}^r \sum_{j=1}^c a_{ij}$ is closer to the binomial distribution when a_{ij} is defined as follows:

$$a_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \left[\frac{n_{i\cdot}}{c} - 0.5 \right] \\ 1 & \text{if } n_{ij} \geq \left[\frac{n_{i\cdot}}{c} - 0.5 \right] \end{cases},$$

where $[x]$ = smallest integer larger than x . (Note that $[a-0.5]$ is the "nearest integer to a " with rounding up of 0.5.)

This adjustment will be explored more thoroughly in Chapter III but first the simulation results will be described.

2.2.4 Simulation results:

Results of 100 tests using random numbers of one, two and three digits to form tables of 10 columns and 50 rows gave the means very similar to the binomial means, 250. However, the standard deviations obtained are smaller than the binomial expected standard deviation, 11.18. This is due to the fact that the a_{ij} 's are negatively correlated to each other. Therefore, binomial approximations should be considered as a conservative first step of testing. Exact tests and/or other approximations should be performed to confirm a non-significant result from the binomial test.

The following proof shows why the variance obtained through the simulation is much smaller than that of the binomial approximation; again, since the rows are independent of each other, for simplicity we prove for the case of $r=1$ without loss of generality

$$\text{Var}(A) = \sum_{i=1}^m \text{Var}(a_i) + \sum_{i=1}^m \sum_{j=1, j \neq i}^m \text{Cov}(a_i, a_j) .$$

While the binomial approximation assumes that $\text{Cov}(a_i, a_j) = 0$, the true $\text{Cov}(a_i, a_j)$ has a negative value as shown below:

$$E(a_i) = E(a_j) = P(n_i \geq \rho) = I_p^1(\rho, v) = \alpha .$$

For $i \neq j$:

$$E(a_i a_j) = P(n_i \geq \rho, n_j \geq \rho) = I_p^2(\rho, v) = \beta .$$

$$\text{Cov}(a_i, a_j) = E(a_i a_j) - E(a_i) E(a_j) = \beta - \alpha^2 .$$

Note that $\beta = P(n_i \geq \rho, n_j \geq \rho) = P(n_i \geq \rho | n_j \geq \rho) P(n_j \geq \rho)$.

Since $P(n_i \geq \rho | n_j \geq \rho) < P(n_i \geq \rho) = \alpha$, we have $\beta < \alpha^2$.

Hence

$$\text{Cov}(a_i, a_j) < 0 .$$

If we let $P(n_i \geq \rho | n_j \geq \rho) = \alpha'$, then

$$\begin{aligned} \text{Var}(A) &= m \times \text{Var}(a_i) + m(m-1)(\beta - \alpha^2) \\ &= m \times \text{Var}(a_i) + m(m-1)\alpha(\alpha' - \alpha) . \end{aligned}$$

Note that $(\alpha' - \alpha) < 0$.

Therefore

$$\text{Var}(A) < m \times \text{Var}(a_i) .$$

Hypothetical clustering data are also generated by using random numbers from the tables referred to above, with the first column changed to zero, then changed to the sum of the first and second columns, and finally changed to the sum of the first, second and third columns. Results are presented in the following Table 2.1, which indicates that the test performs well on either side of clustering (cluster avoidance when column 1 = 0 or clustering when column 1 is larger than expected by the random process). We have larger means with cluster avoidance data and smaller means with clustering data, and the larger clustering data gives smaller means. The standard deviations were not significantly affected by the size of clustering.

Table 2.1

Simulation results of
100 tests using random numbers of 1, 2 and 3 digits

10 columns, 50 rows => $m=500$, $E(A)=250$, $Var(A)=125$, $SD(A)=11.18$

		1-Digit	2-Digit	3-Digit
Random Completely	Mean SD	249.75 6.29	250.13 6.44	250.66 5.88
Column 1 = 0	Mean SD	252.09 5.58	251.49 6.17	251.57 6.89
Column 1 = Col 1 + Col 2	Mean SD	244.74 5.71	246.63 6.89	243.29 6.30
Column 1 = Col 1 + Col 2 + Col 3	Mean SD	238.17 6.62	233.21 5.89	235.54 7.96

CHAPTER III

REGRESSION ESTIMATION FOR THE ZERO-ONE MATRIX TEST FOR TIME CLUSTERING

3.0 Introduction

As shown by the simulation results in Section 2.2.4, the binomial approximation tends to give a conservative estimate of the null distribution of the statistic A , in testing for time clustering. The binomial estimate of the variance is much larger than the variance obtained by the simulation procedure.

In this chapter we will attempt to compare the binomial estimates with the exact mean and variance for a range of situations with different values of n , the number of cases to be distributed into m time units. Based on the results of this comparison we will use a linear regression technique to estimate the mean and the variance of the statistic A .

3.1 Binomial estimates vs the exact values

For the statistic A , since the rows (space units) are treated as independent from each other, we shall, without loss of generality, consider the distribution of cases in only one row: n disease cases are distributed into m time units, each unit contains n_j cases ($\sum_{j=1}^m n_j = n$). Let $p = \frac{1}{m}$ and $\rho = \lceil \frac{n}{m} \rceil =$ smallest integer larger than $(\frac{n}{m} - 0.5)$. To test the hypothesis that some units have unusually more (or fewer) cases than others, the test

statistic is

$$A = \sum_{j=1}^m a_j,$$

where

$$a_j = \begin{cases} 0 & \text{if } n_j < \rho, \\ 1 & \text{if } n_j \geq \rho. \end{cases}$$

As shown in the derivation of equations (2.9) and (2.10), the exact expected value of A and A^2 are:

$$E(A) = m\alpha,$$

$$E(A^2) = m\alpha + m(m-1)\beta,$$

where

$$\begin{aligned} \alpha &= I_p^1(\rho, n) = I_p(\rho, n-\rho+1) \\ &= \text{Incomplete Beta function} \end{aligned}$$

$$\beta = I_p^2(\rho, n) = \text{Incomplete dirichlet function.}$$

From these results we have the exact variance of A :

$$\begin{aligned} \text{Var}(A) &= E(A^2) - [E(A)]^2, \\ &= m\alpha + m(m-1)\beta - m^2\alpha^2, \\ &= m(m-1)\beta - m\alpha(m\alpha-1). \end{aligned} \tag{3.1}$$

For m from 5 to 10 and some selected values of n from 5 to 50, the α was computed using the Incomplete Beta program of the Institutes of Mathematical Statistics Library (IMSL) and β was computed using a computer program written by the author especially for this purpose. Details of these programs can be found in Appendices 1.1 to 1.3.

From these values of α 's and β 's, the exact expected value and variance of A are calculated using equations (2.9) and (3.1) respectively.

Under the binomial approximate distribution of A, $[A \sim \text{Bin}(m, \frac{1}{2})]$, the (approximate) expected value and variance of A are:

$$\text{Bin } E(A) = \frac{m}{2}$$

$$\text{Bin } \text{Var}(A) = \frac{m}{4}$$

The exact and binomial approximate mean and variance of A for each combination of $m = 5(1)10$ and $n = 5(1)10, 12, 15, 18, 20(5)50$ are presented in Table 3.1.

Consistent with the simulation results, the expected values of A under the binomial approximation are very similar to those under the exact distribution, while the exact variances are much smaller than those obtained from the binomial approximation. In fact, the approximate variances are about 3 times larger than the exact ones.

The ratios and the differences between the approximate and the exact variances as well as those of the expected values are also presented in Table 3.1.

Over all, the ratios between the variances are about 3, while the ratios between the expected values are close to unity. Means of the ratios and means of the differences between the approximate and the exact moments are presented in Table 3.2 by number of cells m. The difference between the variances becomes larger as m is increased, while the ratio between these values remains about the same for all values of m from 5 to 10.

3.2 Estimate the variance using the expected value

Since approximate variance of A is a linear function of m ($\text{Bin } \text{Var}(A) = \frac{m}{4}$), the exact variance of the statistic A should be highly correlated with m, which is in turn highly correlated with the exact expected value of this statistic. Therefore it is believed that the expected value and the

variance of the statistic A are highly correlated.

A general least squares model fitting the exact variance on the exact expected value showed that the variance can be closely estimated as a function of the exact expected value as follows:

$$\text{Var}(A) = 0.155 \times E(A) . \quad (3.2)$$

This model gives an R-square of 99.4%; the details of the regression results are presented in Table 3.3.

Some other models were also tried but none gave a better fit; the same model as in (3.2) with an intercept term gives an R-square of 90.4%; the model fitting the variance on the squared term of the expected value gives an R-square of 87.8% with an intercept term and 93.8% without an intercept term; the model as in (3.2) with an addition of the square term of the expected value only improves the R-square value from 99.4 to 99.5%; the model fitting the standard deviation (instead of the variance) on the expected value gave an R-square value of 90.1% with an intercept term and 98.1% without.

From these results, it is concluded that the model of equation (3.2) gives the "best" estimate of variance using the expected value; the word "best" here is used to indicate the relative simplicity and lack of large errors of the estimate.

3.3 Estimate the expected value

It was observed that even though the ratios between the binomial approximate and the exact expected value of (A) remain about the same with different values of m , (Table 3.2), within each value of m these values depend on the decimal part on the value $\frac{n}{m}$. This is intuitively logical since the expected value of A depends on the probability that each

$a_i = 1$, i.e. the probability that $n_i >$ integer part of $(\frac{n}{m} - 0.5)$, this probability depends on the value $\frac{n}{m}$; yet n_i can only be an integer; therefore the decimal part of $\frac{n}{m}$ was neglected in the binomial approximation. For a better estimate of the expected value of the statistic A, this decimal part should be taken into account.

A general least squares model fitting the exact expected value of A on the number of time units (m) and $\frac{n}{m}$ minus the smallest integer larger than $(\frac{n}{m} - 0.5)$ (Decimal) showed that the expected value of the statistic A can be estimated as follows:

$$E(A) = 0.6 \times m + 2.2 \times \text{Decimal} \quad (3.3)$$

This model gives an R-square value of 99.5%. The details of the regression results are presented in Table 3.4.

The same model as in (3.3) with intercept term gives an R-square value of 92.0%.

Due to the simplicity of the model and its good fit, equation (3.3) can be chosen to estimate the expected value of A. However, since $E(A) = m\alpha$, where α is the probability that $n_i \geq [\frac{n}{m} - 0.5]$, and α is independent of i , a more general estimate of $E(A)$ can be obtained from an estimate of α multiplied by m .

A general least squares model fitting the exact value of α on the difference between $\frac{n}{m}$ and $[\frac{n}{m} - 0.5]$ (Decimal) showed that the value of α can be estimated as follows:

$$\alpha = 0.6 + 0.3 \times \text{Decimal} \quad (3.4)$$

This model gives an R-square value of 80.8%. The details of the regression results are presented in Table 3.5.

From equation (3.4) the expected value of A can be estimated by:

$$E(A) = m\alpha$$

$$E(A) = m(0.6 + 0.3 \times \text{Decimal}) \quad (3.5)$$

3.4 Appropriateness of the regression estimation

To check the appropriateness of the regression estimation, equation (3.5) was used to estimate the expected value of A and this estimate was then used to estimate the variance of A, using equation (3.2). The residuals were calculated as the difference between the estimates and the exact values. These residuals were standardized in percents as follows:

$$\text{Standardized residual} = \frac{\text{Residual} \times 100}{\text{exact values}}$$

These standardized residuals were plotted against n in Plots 3.1 and 3.2, with Plot 3.1 presenting the residuals of the expected values and Plot 3.2 presenting the residuals of the variances. Univariate procedure applied on the residuals showed that in estimating the expected value the residual is not more than 15% of the value being estimated; 90% of the time the residual is not more than 10% and 50% of the time the residual is not more than 6%. In estimating the variance, the residual is not more than 1/3 of the estimated value (33%); 90% of the time the residual is not more than 16%, and 50% of the time the residual is not more than 7%. More details of these results can be found in Appendix 2.

In summary, the results of the study presented in this chapter lead to the conclusion that a good estimate of the expected value of the test statistic A, test for time-clustering, is

$$E(A) = m(0.6 + 0.3 \times \text{Decimal})$$

where m is the number of time units and Decimal is $\frac{n}{m}$ minus the smallest

integer larger than $(\frac{n}{m} - 0.5)$; n is the number of disease cases to be distributed into m time units. A good estimate of the variance of A is

$$V(A) = 0.155 \times E(A) .$$

These estimates are simple and relatively accurate with the error of the estimate usually less than 10% for $E(A)$ and less than 15% for $V(A)$.

Table 3.1*
Approximate and exact values of E(A) and Var(A) for n = 5(1)10,12,15,18,20(5)50 and m = 5(1)10

OBS	H	M	R	L1	L2	EEA	EVA	UEA	DVA	REA	RVA	UEA	DVA	EV	DECIMAL
1	5	5	1	0.672120	0.422400	3.16160	0.50925	4.5	1.25	0.74369	2.45861	-0.8616	0.74075	6.60113	0.00000
2	6	5	1	0.598122	0.327912	3.58873	0.54769	4.0	1.50	0.83595	2.73079	-0.3067	0.95231	8.52524	-0.16667
3	7	5	1	0.537336	0.260606	3.76135	0.55904	3.5	1.75	0.93052	3.13037	-0.2613	1.19096	6.72623	-0.28571
4	8	5	1	0.487091	0.211487	4.89673	0.55550	4.0	2.00	1.02500	3.60037	0.1033	1.44506	7.01634	-0.37500
5	9	5	1	0.445071	0.174770	4.00564	0.54394	4.5	2.25	1.12342	4.13647	0.4944	1.70606	7.36409	-0.44444
6	10	5	1	0.409510	0.146700	4.09510	0.52827	5.0	2.50	1.22097	4.73246	0.9049	1.97173	7.75196	-0.50000
7	5	6	1	0.737856	0.522368	3.68928	0.52585	2.5	1.25	0.67764	2.73466	-1.1893	0.72415	7.01580	0.20000
8	6	6	1	0.665102	0.417996	3.99061	0.60550	1.0	1.50	0.75176	2.47729	-0.9906	0.89450	6.59061	0.00000
9	7	6	1	0.603430	0.339671	4.22401	0.64792	3.5	1.75	0.82860	2.70095	-0.7240	1.10208	6.51934	-0.14286
10	8	6	1	0.551205	0.280388	4.40964	0.66646	4.0	2.00	0.90710	3.00094	-0.4096	1.33354	6.61853	-0.25000
11	9	6	1	0.506730	0.234837	4.56057	0.67005	4.5	2.25	0.98672	3.15747	-0.0606	1.57995	6.80634	-0.33333
12	10	6	1	0.468559	0.199262	4.68559	0.66443	5.0	2.50	1.06710	3.76263	0.3144	1.81557	7.05206	-0.40000
13	5	7	1	0.790285	0.608563	3.95142	0.50894	2.5	1.25	0.63268	2.45610	-1.4514	0.74106	7.76407	0.40000
14	6	7	1	0.720918	0.500364	4.32551	0.62641	3.0	1.50	0.69356	2.39460	-1.3255	0.87359	6.90524	0.16667
15	7	7	1	0.660083	0.415031	4.62058	0.70210	3.5	1.75	0.75748	2.49252	-1.1266	1.04790	6.58107	0.00000
16	8	7	1	0.607304	0.348092	4.85843	0.74722	4.0	2.00	0.82331	2.67660	-0.8584	1.25278	6.50204	-0.12500
17	9	7	1	0.561538	0.295258	5.05384	0.77110	4.5	2.25	0.89041	2.91790	-0.5538	1.47890	6.55403	-0.22222
18	10	7	1	0.521703	0.253121	5.21703	0.78056	5.0	2.50	0.95840	3.20283	-0.2170	1.71944	6.68371	-0.30000
19	5	8	2	0.496683	0.204252	2.48342	0.40110	2.5	1.25	1.00668	3.11644	0.0166	0.84890	6.19155	-0.40000
20	6	8	1	0.767432	0.573882	4.60459	0.61880	3.0	1.50	0.65152	2.42403	-1.6046	0.68120	7.44111	0.14286
21	7	8	1	0.708643	0.485046	4.96050	0.72588	3.5	1.75	0.70557	2.41088	-1.4605	0.80412	6.83360	0.00000
22	8	8	1	0.658391	0.412895	5.25113	0.79890	4.0	2.00	0.76174	2.50344	-1.2511	1.20110	6.57233	0.00000
23	9	8	1	0.610256	0.354431	5.49230	0.84596	4.5	2.25	0.81933	2.65969	-0.9923	1.40404	6.49237	-0.11111
24	10	8	1	0.569533	0.306838	5.69533	0.87398	5.0	2.50	0.87791	2.85057	-0.6953	1.62602	6.53654	-0.20000
25	5	9	2	0.563792	0.278750	2.81896	0.44742	2.5	1.25	1.09328	3.27212	0.2560	1.04158	5.98589	0.50000
26	6	9	2	0.457341	0.174805	2.74405	0.45842	3.0	1.50	1.03138	2.42126	-1.7519	1.02505	7.24441	0.28571
27	7	9	1	0.750265	0.548931	5.25186	0.72495	3.5	1.75	0.66643	2.41395	-1.5947	1.17528	6.78384	0.12500
28	8	9	1	0.699362	0.473769	5.59474	0.82471	4.0	2.00	0.71496	2.42508	-1.5947	1.17528	6.78384	0.12500
29	9	9	1	0.653561	0.411281	5.88204	0.89582	4.5	2.25	0.76504	2.51166	-1.3258	1.35418	6.56613	0.00000
30	10	9	1	0.612579	0.355204	6.12095	0.94436	5.0	2.50	0.81622	2.64730	-1.1258	1.55564	6.48673	-0.10000
31	5	10	2	0.624190	0.352062	3.12095	0.48470	2.5	1.25	0.80104	2.57892	-0.6210	0.76530	6.44884	0.00000
32	6	10	2	0.515483	0.232562	3.09290	0.50373	3.0	1.50	0.96996	2.97777	-0.0929	0.99627	6.13995	-0.33333
33	7	10	1	0.785942	0.606455	5.50159	0.70519	3.5	1.75	0.63618	2.48160	-2.0016	1.04481	7.80158	0.42857
34	8	10	1	0.736924	0.530162	5.89540	0.82880	4.0	2.00	0.72249	2.43727	-1.7285	1.32683	6.74688	0.11111
35	9	10	1	0.692054	0.465121	6.22848	0.92316	4.5	2.25	0.76767	2.51808	-1.5132	1.50710	6.56031	0.00000
36	10	10	1	0.651321	0.410017	6.51321	0.99282	5.0	2.50	0.68954	2.42347	-1.1258	0.73421	7.02926	0.40000
37	5	12	2	0.725122	0.501761	3.62561	0.51579	2.5	1.25	0.68954	2.42347	-1.1258	0.73421	7.02926	0.40000
38	6	12	2	0.618667	0.354872	3.71200	0.57919	3.0	1.50	0.80819	2.50891	-0.7120	0.92081	6.40891	0.00000
39	7	12	2	0.528198	0.251831	3.69739	0.60364	3.5	1.75	0.94661	2.89906	-0.1974	1.14636	6.12510	-0.28571
40	8	12	2	0.453296	0.181120	3.62637	0.61857	4.0	2.00	1.10303	3.23320	0.3736	1.38143	5.86251	-0.50000
41	9	12	1	0.756684	0.562377	6.81016	0.92302	4.5	2.25	0.66078	2.43766	-2.3102	1.32695	7.37815	0.33333
42	10	12	1	0.717570	0.503861	7.17570	1.03243	5.0	2.50	0.69680	2.42146	-2.1757	1.46757	6.95028	0.20000
43	5	15	3	0.601977	0.326149	3.00988	0.47347	2.5	1.25	0.83060	2.64006	-0.5099	0.77653	6.15702	0.00000
44	6	15	3	0.467775	0.185379	2.80665	0.49074	3.0	1.50	1.06809	3.05662	0.1934	1.00926	5.71525	-0.50000
45	7	15	2	0.653370	0.405732	4.57359	0.69661	3.5	1.75	0.76536	2.51216	-1.0736	1.05339	6.56549	0.14286
46	8	15	2	0.575922	0.309979	4.60738	0.73826	4.0	2.00	0.86817	2.70907	-0.6074	1.26174	6.24085	-0.12500
47	9	15	2	0.508696	0.238096	4.57427	0.76068	4.5	2.25	0.94250	2.95787	-0.0783	1.46932	6.01862	-0.33333
48	10	15	2	0.450357	0.184488	4.50957	0.77733	5.0	2.50	1.10875	3.21612	0.4904	1.72267	5.80133	-0.40000
49	5	18	4	0.498974	0.207787	2.49487	0.42623	2.5	1.25	1.00206	2.93267	0.0051	0.82377	5.85130	-0.40000
50	6	18	3	0.597345	0.327630	3.58967	0.56740	3.0	1.50	0.81704	2.64365	-0.3047	0.93266	6.31664	0.00000
51	7	18	3	0.485469	0.208063	3.39628	0.58663	3.5	1.75	1.02993	2.97301	-1.4173	1.16137	5.71321	-0.42857
52	8	18	2	0.677160	0.441705	5.41720	0.80582	4.0	2.00	0.71838	2.48194	-1.4173	1.16137	5.71321	-0.42857
53	9	18	2	0.609934	0.354266	5.48941	0.86298	4.5	2.25	0.81976	2.60724	-0.4964	1.10702	6.16049	0.00000
54	10	18	2	0.549716	0.284463	5.49716	0.90260	5.0	2.50	0.70750	2.78936	-0.4974	1.00174	6.13343	-0.20000
55	5	20	4	0.588551	0.309213	2.94275	0.40722	2.5	1.25	0.84954	2.67542	-0.4428	0.78278	6.24849	0.00000

*See page 36 for definitions of column labels

Table 3.1 (continued)

OBS	M	N	B	I1	I2	EEA	EVA	UEA	UVA	PEA	BVA	DEA	DVA	EV	DECIMAL
56	6	20	3	0.671341	0.426564	4.02804	0.598834	3.0	1.50	0.74478	2.50069	-1.0280	0.90017	6.71527	0.33333
57	7	20	3	0.559610	0.287302	3.9127	0.630976	3.5	1.75	0.89348	2.73976	-0.4173	1.11102	5.13054	-0.14296
58	8	20	3	0.464691	0.192197	3.71753	0.660526	4.0	2.00	1.07598	3.02789	-0.2825	1.33947	5.62811	-0.50000
59	9	20	2	0.668092	0.431149	6.01283	0.903424	4.5	2.25	0.74040	2.49605	-1.5328	1.34858	6.67037	0.22222
60	10	20	2	0.608253	0.354136	6.08253	0.957644	5.0	2.50	0.82203	2.61057	-1.0815	1.54236	6.35156	0.00000
61	5	25	5	0.579326	0.297854	2.89663	0.463248	2.5	1.25	0.86307	2.69834	-0.3966	0.78675	6.25287	0.00000
62	6	25	4	0.618433	0.354495	3.71060	0.576903	3.0	1.50	0.80889	2.60009	-0.7106	0.92110	6.43193	0.16667
63	7	25	4	0.488066	0.210867	3.41646	0.600668	3.5	1.75	1.02445	2.51342	-0.0835	1.14933	5.68777	-0.42857
64	8	25	3	0.620391	0.365042	4.96313	0.772852	4.0	2.00	0.80594	2.58702	-0.9631	1.22715	6.42183	0.12500
65	9	25	3	0.536248	0.267665	4.82623	0.805621	4.5	2.25	0.93240	2.57288	-0.3242	1.44438	5.99069	-0.22222
66	10	25	3	0.462906	0.195869	4.62906	0.86244	5.0	2.50	1.08013	3.01542	-0.3709	1.67651	5.58243	0.50000
67	5	30	6	0.572887	0.289581	2.86244	0.329071	2.5	1.25	0.87338	2.71444	-0.3524	0.74951	5.21594	0.00000
68	6	30	5	0.575661	0.303089	3.45397	0.56745	3.0	1.50	0.86657	2.69423	-0.4840	0.94326	6.20186	0.00000
69	7	30	4	0.638271	0.385147	4.46790	0.831959	3.5	1.75	0.78337	2.56613	-0.9679	1.06804	6.55156	0.00000
70	8	30	4	0.526614	0.254456	4.21292	0.713815	4.0	2.00	0.94946	2.80185	-0.2129	1.28618	5.99197	-0.23000
71	9	30	3	0.562799	0.423742	5.88648	0.942228	5.0	2.25	0.84940	2.65329	-1.4652	1.35286	6.69191	0.33333
72	10	30	3	0.588648	0.330072	2.83579	0.458504	2.5	1.25	0.88159	2.72626	-0.3358	0.77150	6.24741	0.00000
73	5	35	6	0.567158	0.259759	3.23887	0.541354	3.0	1.50	0.82238	2.69291	-0.5120	0.95865	5.94291	-0.16667
74	6	35	6	0.539812	0.303190	4.01200	0.649854	3.5	1.75	0.90078	2.77083	-0.2389	1.10015	6.17370	0.00000
75	7	35	4	0.652175	0.407203	5.21900	0.744376	4.0	2.00	0.76043	2.54980	-1.2190	1.21562	6.63370	0.00000
76	8	35	4	0.555077	0.288697	4.99570	0.824899	4.5	2.25	0.98235	2.82849	-0.0539	0.94968	5.75862	-0.33333
77	9	35	4	0.469015	0.201761	2.81427	0.851116	5.0	2.50	1.06606	2.72761	-0.4957	1.42510	6.05613	-0.11111
78	10	35	4	0.562854	0.278142	4.69015	0.56948	2.5	1.25	0.88833	2.71530	-0.3143	1.64888	5.51054	-0.50000
79	5	40	8	0.571305	0.240153	3.05390	0.530317	3.0	1.50	0.98235	2.82849	-0.1166	1.2676	5.80293	0.00000
80	6	40	7	0.508983	0.226757	4.99570	0.824899	4.5	2.25	0.96776	2.80792	-0.5704	1.25723	6.15323	0.00000
81	7	40	6	0.516657	0.240153	3.61660	0.623237	3.5	1.75	0.96776	2.80792	-0.7687	1.36482	6.74003	0.44444
82	8	40	5	0.62907	0.304665	4.57084	0.742771	4.0	2.00	0.87519	2.69262	-0.5704	1.56623	6.17782	0.00000
83	9	40	4	0.576869	0.316032	5.96616	0.885176	4.5	2.25	0.85675	2.54187	-1.4662	1.36482	6.74003	0.44444
84	10	40	4	0.576869	0.316032	5.96616	0.885176	4.5	2.25	0.85675	2.54187	-1.4662	1.36482	6.74003	0.44444
85	5	45	9	0.59283	0.274043	2.79682	0.457320	2.5	1.25	0.75425	2.69262	-0.2904	0.79262	6.17782	0.00000
86	6	45	8	0.481957	0.199783	2.89174	0.523077	3.0	1.50	0.85675	2.54187	-0.2904	0.79262	6.17782	0.00000
87	7	45	6	0.636914	0.383228	4.45860	0.676662	3.5	1.75	0.87504	2.86765	-0.1083	0.97692	5.52833	-0.50000
88	8	45	6	0.499431	0.226318	3.99545	0.705632	4.0	2.00	1.00114	2.83434	-0.0946	1.07334	6.58082	0.42857
89	9	45	5	0.569904	0.305758	5.12914	0.835625	4.5	2.25	0.87734	2.69262	-0.6291	1.29437	5.66223	-0.37500
90	10	45	5	0.472862	0.205494	4.72862	0.802289	5.0	2.50	1.05739	2.89590	-0.2714	1.41438	6.13809	0.00000
91	5	50	10	0.556259	0.270259	2.78130	0.450864	2.5	1.25	0.84886	2.77245	-0.2813	0.75914	5.47745	-0.50000
92	6	50	8	0.608941	0.342142	3.65364	0.568782	3.0	1.50	0.82110	2.63721	-0.6536	0.93122	6.42362	0.33333
93	7	50	7	0.584337	0.316560	4.09016	0.654827	3.5	1.75	0.95567	2.67246	-0.5904	1.09517	6.24647	0.14286
94	8	50	6	0.606513	0.347331	4.85210	0.759734	4.0	2.00	0.82438	2.63250	-0.8521	1.24027	6.36658	0.25000
95	9	50	6	0.486133	0.216046	4.37519	0.780185	4.5	2.25	1.02853	2.85466	-0.1248	1.46181	5.55097	-0.44444
96	10	50	5	0.568801	0.306599	5.68801	0.928464	5.0	2.50	0.87904	2.69262	-0.6880	1.57134	6.12626	0.00000

Table 3.1 (continued)

VARIABLE	N	MEAN	STD DEV	SUM	MINIMUM	MAXIMUM
I1	96	0.58740220	0.08690365	54.39661140	0.40950910	0.79628480
I2	96	0.32891416	0.10839478	31.57575980	0.14670000	0.60856320
EEA	96	4.37896320	1.09789159	420.38046700	2.48341800	7.17570300
EVA	96	0.68261654	0.15995691	65.51118800	0.40109810	1.03243400
BEA	96	3.75000000	0.85839506	360.00000000	2.50000000	5.00000000
BVA	96	1.87500000	0.42919754	180.00000000	1.25000000	2.50000000
DECIMAL	96	-0.06991981	0.27263506	-6.71230159	-0.50000000	0.44444444

CORRELATION COEFFICIENTS / PROB > |R| UNDER H0:RHO=0 / N = 96

	I1	I2	EEA	EVA	BEA	BVA	DECIMAL
I1	1.00000	0.99301	0.43236	0.23472	-0.17985	-0.17985	0.89898
I2	0.00000	0.00001	0.00001	0.0213	0.0795	0.0795	0.00001
EEA	0.99301	1.00000	0.46255	0.27997	-0.11295	-0.11295	0.87449
EVA	0.00001	0.00000	0.00001	0.0057	0.2732	0.2732	0.00001
BEA	0.43236	0.46255	1.00000	0.95068	0.79946	0.79946	0.36396
BVA	0.00001	0.00001	0.00000	0.00001	0.00001	0.00001	0.00003
DECIMAL	0.23472	0.27997	0.95068	1.00000	0.86991	0.86991	0.18743
	0.0213	0.0057	0.0001	0.0000	0.0001	0.0001	0.0675
	-0.17985	-0.11295	0.75946	0.86991	1.00000	1.00000	-0.19471
	0.0795	0.2732	0.0001	0.0001	0.0000	0.0000	0.0573
	-0.17985	-0.11295	0.75946	0.86991	1.00000	1.00000	-0.19471
	0.0795	0.2732	0.0001	0.0001	0.0000	0.0000	0.0573
	0.89898	0.87449	0.36396	0.18743	-0.19471	-0.19471	1.00000
	0.00001	0.00001	0.00003	0.0675	0.0573	0.0573	0.0000

Labels of the variable names in Table 3.1:

M = number of time units

N = number of disease cases to be distributed into M units

R = smallest integer larger than $(\frac{N}{M} - 0.5)$

I1 = $I_{1/M}^1(R, N)$ = incomplete dirichlet of first degree

= α = incomplete beta function $I_{1/M}(R, N-R+1)$

I2 = $I_{1/M}^2(R, N)$ = incomplete dirichlet of second degree

EEA = exact expected value of A = $M \times I1$

EVA = exact variance of A = $M(M-1) I2 - M(I1)(M \times I1 - 1)$

BEA = binomial approximate expected value of A = $\frac{M}{2}$

BVA = binomial approximate variance of A = $\frac{M}{4}$

RVA = BVA/EVA

REA = BEA/EEA

DEA = BEA - EEA

DVA = BVA - EVA

EV = EEA/EVA

Decimal = $\frac{N}{M} - R$

Table 3.2

Means of the ratio and of the difference between the approximate and the exact moments by number of cells, based on results presented in Table 3.1.

Number of obs.	m	BinVarA/Var(A)	BinVar-VarA	Bin E(A)/E(A)	Bin E(A)-E(A)
16	5	2.68	0.78	0.84	-0.53
16	6	2.72	0.94	0.87	-0.53
16	7	2.69	1.09	0.84	-0.75
16	8	2.76	1.28	0.87	-0.69
16	9	2.79	1.43	0.86	-0.84
16	10	2.98	1.64	0.94	-0.44
96	5-10	2.77	1.19	0.87	-0.63

Table 3.3

Regressing the exact variance on the exact expected value of the test statistic A - Test for time clustering

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: EVA = exact variance of A

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	1	46.89698209	46.89698209	16724.97	0.0001	0.999352	7.7573
ERROR	95	0.26638102	0.00280401		STD DEV		EVA MEAN
UNCORRECTED TOTAL	96	47.16336311			0.05295291		0.68261654

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
EVA	1	46.89698209	16724.97	0.0001	1	46.89698209	16724.97	0.0001

PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > T	STD ERROR OF ESTIMATE
EVA	0.15486795	129.33	0.0001	0.00119751

EEA = exact expected value of (A)

Table 3.4

Regressing the exact expected value of A on the number of time units m and the difference between $\binom{n}{m}$ and $[\frac{n}{m} - 0.5]$

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: EEA											
SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.				
MODEL	2	1945.98441884	972.99220942	9775.74	0.0001	0.995215	7.2046				
ERROR	94	9.35594008	0.09953128				EEA MEAN				
UNCORRECTED TOTAL	96	1955.34035892					4.37896320				
SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F			
M	1	1912.61568566	19216.23	0.0001	1	1897.82825884	19067.66	0.0001			
DECIMAL	1	33.36873318	335.26	0.0001	1	33.36873318	335.26	0.0001			
PARAMETER	ESTIMATE	T FOR H0: PARAMETER=0	PR > T	STD ERROR OF ESTIMATE							
M	0.60309615	138.09	0.0001	0.00436755							
DECIMAL	2.19626911	18.31	0.0001	0.11994868							

Table 3.5

Regressing α (probability that $n_i \geq \lfloor \frac{n}{m} - 0.5 \rfloor$) on the difference between $\lfloor \frac{n}{m} \rfloor$ and $\lfloor \frac{n}{m} - 0.5 \rfloor$

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: II

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	F-SQUARE	C.V.
MODEL	1	0.57983311	0.57983311	396.02	0.0001	0.800171	6.5141
ERROR	94	0.13763010	0.00146415		STD DEV		II MEAN
CORRECTED TOTAL	95	0.71746321			0.03826421		0.58740220

SOURCE	DF	TYPE I SS	F VALUE	PR > F	DF	TYPE IV SS	F VALUE	PR > F
DECIMAL	1	0.57983311	396.02	0.0001	1	0.57983311	396.02	0.0001

PARAMETER ESTIMATE

INTERCEPT 0.60743807

DECIMAL 0.28655499

PR > |T|

0.0001
0.0001

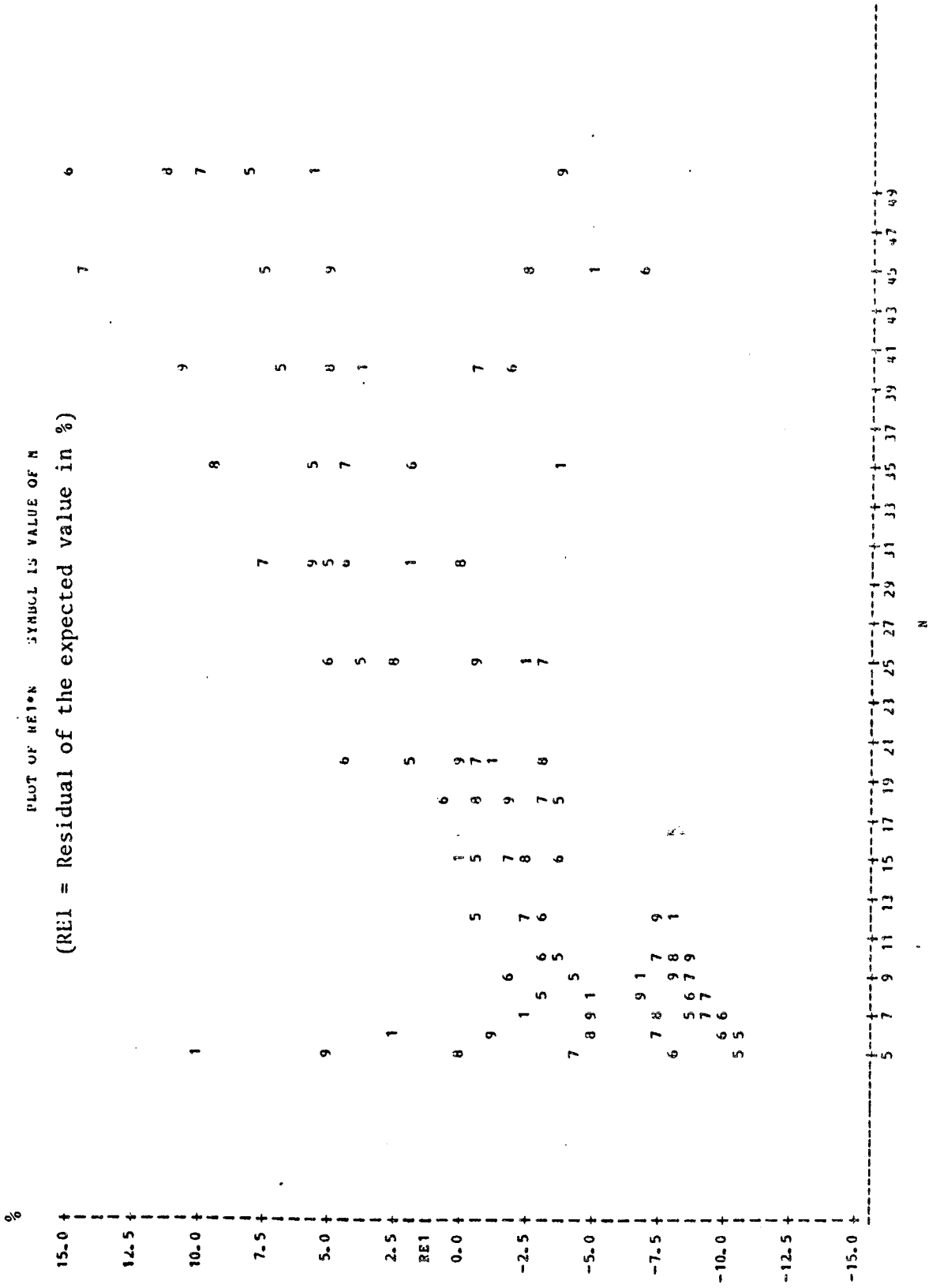
STD ERROR OF ESTIMATE

0.00403302
0.01439956

T FOR HO: PARAMETER=0

150.62
19.90

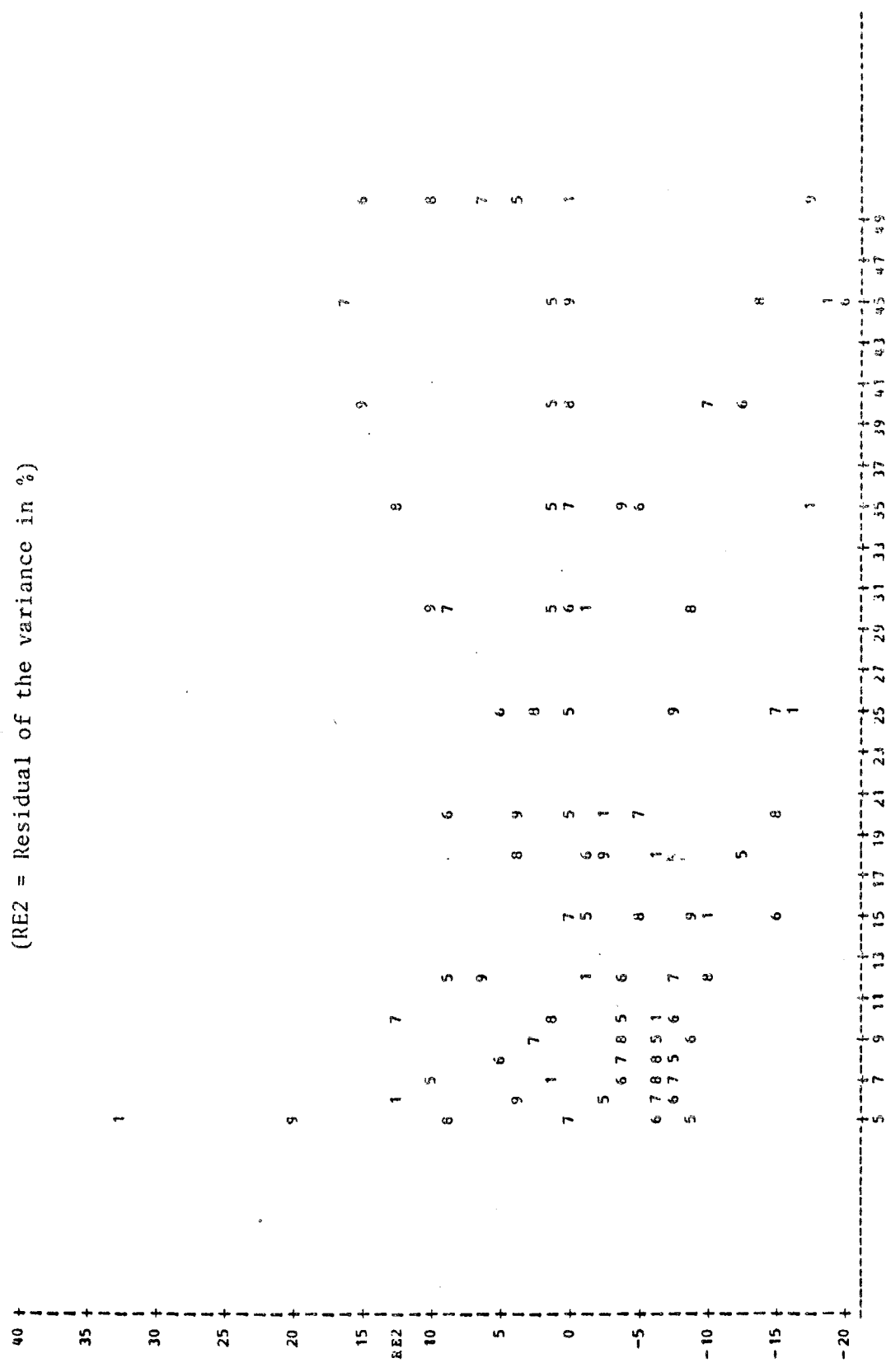
Plot 3.1



NOTE: 7 OBS HIDDEN

Plot 3.2

PLOT OF RE2*
(RE2 = Residual of the variance in %)



NOTE: 9 OBS HIDDEN

CHAPTER IV

COMPARING THE ZERO-ONE MATRIX TEST FOR TIME CLUSTERING WITH THE EMM TEST AND THE SCAN TEST

4.0 Introduction

The zero-one matrix test is based on the number of disease cases observed in each time-space unit, without regarding the arrangements of cases within each unit. In this respect, more than any other test, it is similar to the EMM test and the scan test, both of which are based on the maximum number of disease cases in an appropriately defined time-space unit.

In this chapter we will compare the zero-one matrix test with these other two tests, in an attempt to define situations in which each test is more powerful than the others and to determine the alternative hypotheses for which one test is more appropriate than the others.

4.1 Situations in which the zero-one matrix test is more powerful than the EMM test

For a fixed number of time units and conditional on the total frequency in all units, the EMM test depends only on the statistic a , the maximum frequency in a time unit, regardless of the distribution of the other frequencies in the individual units, other than the maximum. On the other hand, the zero-one matrix test statistic depends on the frequencies of all the individual units. Therefore, the difference in power of the two tests depends on the frequencies of the individual units, other than the maximum. To test for time clustering, the zero-one matrix test is more powerful in situations in

which there is less variation between the frequencies of the individual time units, other than the maximum. The following examples, (4.1) and (4.2), illustrate this point:

Example 4.1: Annual numbers of cervix cancer deaths in Nash County of North Carolina reported from 1976 to 1980 are as follows:

Year	1976	1977	1978	1979	1980	Total
Number of cervix cancer deaths	2	2	2	2	6	14

With a total frequency of 14, for 5 time units (average 2.8 deaths per year), the zero-one matrix test statistic is $A=1$, which shows significant clustering. (Under the null hypothesis of no clustering, the approximate expected value and variance of A are, by equations (3.5) and (3.2), 2.70 and 0.42 respectively.) The EMM test statistic in this case is $a=6$; under the null hypothesis, conditional on total frequency of 14 for 5 time units, expected value and variance of a are 4.86 and 0.93 respectively. Therefore, the EMM test shows that the clustering of cervix cancer deaths during 1980 in Nash County of North Carolina is not significant ($Z = \frac{6-4.86}{\sqrt{.93}} = 1.18$), while the zero-one matrix test shows that it is significant ($Z = \frac{1-2.7}{\sqrt{.42}} = -2.62$). The significance declared by the zero-one matrix test is based largely on the fact that there is little variation in the annual frequencies of cervix cancer deaths in Nash County from 1976 to 1979 (2 deaths each year) and in 1980 the number of deaths was 6. If in 1976 and 1977 the number of deaths had been 1 and 3 respectively (instead of 2 each year), as in the following array:

Year	1976	1977	1978	1979	1980	Total
Number of cervix cancer deaths	1	3	2	2	6	14

then the zero-one matrix test statistic would be $A=2$, which shows insignificant clustering for the same maximum and total frequencies of 6 and 14, respectively, for 5 time units. The EMM test would remain unchanged in this case. It can be concluded by this example that while the EMM test is independent of the variation of the frequencies, the zero-one matrix test takes into account this variation, and clusters are more likely to be declared significant by this test when there is a sudden change among otherwise stable frequencies. The following example restates the same point:

Example 4.2: Annual number of arteriosclerosis deaths in Catawba County in North Carolina reported from 1976 to 1980 are as follows:

Year	1976	1977	1978	1979	1980	Total
Number of arteriosclerosis deaths	15	11	11	10	11	58

With a total frequency of 58 for 5 time units (average 11.6 deaths per year), the zero-one matrix test statistic is $A=1$, which shows significant clustering, (under the null hypothesis of no clustering, the approximate expected value and variance of A are, by equations (3.5) and (3.2), 2.40 and .37, respectively, which give a Z value of $Z = \frac{1-2.40}{\sqrt{.37}} = -2.30$.) However, the EMM test statistic $a=15$ is not significant, as its expected value and variance are 15.7 and 3.38 respectively, which give a Z value of $Z = \frac{15-15.7}{\sqrt{3.38}} = -0.38$.

Therefore, the EMM test shows that the clustering of arteriosclerosis deaths during 1976 in Catawba County in North Carolina is not significant, while the zero-one matrix test showed that it is significant, due to the fact that there is little variation in the annual frequencies of arteriosclerosis deaths in Catawba County from 1977 to 1980. If in 1979 and 1980 the number of deaths had been 9 and 12 respectively, instead of 10 and 11, the zero-one

matrix test statistic would be $A=2$, which shows insignificant clustering for the same maximum and total frequencies of 15 and 58, respectively, for 5 time units. The EMM test would remain the same with this frequency change.

The above two examples show that for a given number of time units and conditional on total frequency, the EMM test will not detect a cluster unless maximum frequency in a time unit is quite large, while the zero-one matrix test does not require that large a maximum; rather it depends on the variation between units.

Note that examples (4.1) and (4.2) are used here for illustration purposes. It is assumed that the population of the two counties (Nash and Catawba of North Carolina) did not change substantially from 1976 to 1980, and that the frequencies of deaths before 1976 and after 1980 were ignored. The reasons for the clustering detected are not in the scope of this research and hence are not discussed here.

4.2 Situations in which the EMM test is more powerful than the zero-one matrix test

For comparison purposes, suppose that the annual number of cervix cancer deaths in Nash County in North Carolina reported from 1976 to 1980 which were used in example (4.1) were modified as follows:

Year	1976	1977	1978	1979	1980	Total
Number of cervix cancer deaths	2	0	3	2	7	14

For this data set, the EMM test will show significant clustering (test statistic $A=7$, $E(A) = 4.86$, $\text{Var}(A) = 0.93$, which gives a Z value of $Z = \frac{7-4.86}{\sqrt{.93}} = 2.22$), while the zero-one matrix test will not declare the clustering significant, (test statistic $A=2$, $E(A) = 2.7$, $\text{Var}(A) = 0.42$, which gives a Z value of $Z = \frac{2-2.7}{\sqrt{.42}} = -1.08$).

The significance declared by the EMM test is based on the large maximum frequency in a time unit (7 for 1980) given the fact that there are only 14 cases in 5 time units (from 1976 to 1980). The clustering in 1980 is not significant according to the zero-one matrix test, due to the fact that there is a large variation between time units; it can be argued that, for example, if the frequency can vary by chance from 0 to 3 as in 1977 and 1978, so also it could be by chance that we observed 7 cases in 1980. However, in example (4.1), the frequency observed in 1980 is only 6, but since in the previous 4 years there were only 2 each, this is declared as a significant cluster.

Similar examples can be obtained by assuming that the arteriosclerosis data set of Catawba County used in example (4.2) is modified as follows:

Year	1976	1977	1978	1979	1980	Total
Number of arterio- sclerosis deaths	20	13	4	10	11	58

For this data set, the EMM test shows significant clustering, (test statistic $A=20$, $E(A) = 15.70$, $\text{Var}(A) = 3.38$, which gives a Z value of $Z = \frac{20-15.7}{\sqrt{3.38}} = 2.34$), while the zero-one matrix test does not declare the clustering significant (test statistic $A=2$, $E(A) = 2.4$, $\text{Var}(A) = 0.38$, which gives a Z value of $Z = \frac{2-2.4}{\sqrt{.37}} = -0.66$).

Similar to the above case, the EMM test declares the clustering significant, based on the large maximum frequency in a time unit (20 for 1976), given the fact that there are a total of 58 cases in 5 time units (from 1976 to 1980). This clustering is not significant according to the zero-one matrix test, due to the fact that there is a larger frequency variation between time units compared to the original data in example 4.2.

It can be concluded from these illustrations that the EMM test is more

powerful than the zero-one matrix test in cases in which one large maximum frequency is observed in one unit, together with a large frequency variation between the other units. However, it may be undesirable to call this type of clustering "significant", since a large frequency variation between units leads to the logical expectation of a large maximum in one unit.

4.3 Zero-one matrix test for time-clustering versus the scan test

As in the EMM test, the scan test statistic is the maximum frequency of disease cases in a fixed interval of time; therefore, similar to the case of the EMM test, the scan test is more powerful than the zero-one matrix test in situations in which one large maximum is observed in a fixed interval of time, and the rest of the cases are distributed rather unevenly during the rest of the time intervals under consideration. The zero-one matrix test is more powerful than the scan test in situations in which a small cluster of events is observed in a fixed time unit and the rest of the events are distributed uniformly (evenly) during the rest of the time under consideration.

Due to the nature of the scan test, a criterion for dividing the time under study into fixed intervals (time units) is not needed and the test statistic does not depend on this subjective step. In this respect the zero-one matrix test has a drawback compared to the scan test; however, the scan test does not get this advantage without paying a high price, as neither the exact nor the approximate distribution of the scan test can be assessed easily, and extensive tables on the significant points of this statistic are not available. Furthermore, the scan test statistic requires more detailed data than those required by the zero-one matrix test, when the length of the time unit is the same. For example, if the length of the time unit is one year, then the zero-one matrix test requires annual data, i.e., number of cases occurring in each year with an appropriate starting date, while the scan test requires data with a more exact time at which each case occurred, or at least the semi-

annual data, i.e., number of cases occurring in each six months with an appropriate starting date, so that the scan statistic can be computed as the maximum of all the sums of two consecutive semi-annual frequencies.

Moreover, the scan statistic depends on how detailed the available data are; for example, if the length of the "window" for the scan test is one year, the scan statistic can be the maximum of all the sums of two consecutive semi-annual frequencies or the maximum of all the sums of 12 consecutive monthly frequencies. The latter statistic can be either equal or greater than the former and the two statistics are not necessarily the same. Therefore, in assessing the significant level of the scan test statistic, this fact should be taken into account, and this may not be feasible in most practical situations.

Note that when "time" defines a "unit", "clustering" between neighboring units can occur without being detected by the zero-one matrix test. Example: 5, 7, 8, 8, 5, 7 might be considered less "clusterful" than 5, 5, 7, 8, 8, 7; the zero-one matrix test does not allow for this consideration.

CHAPTER V

GENERALIZED ZERO-ONE MATRIX TEST FOR TIME CLUSTERING WITH ADJUSTMENT FOR EXTRANEOUS FACTORS; MULTIVARIATE APPLICATIONS

5.0 Introduction

A major weakness of most of the tests for time-space clustering is that they do not take into account the population changes due to time and space, nor the difference in distribution of extraneous factors among study groups. The above factors may affect the outcome of the study.

None of the tests found in the literature can be used in a multivariate problem such as cases in which one needs to test the hypothesis of an unusual pattern occurring when a group of different diseases are considered simultaneously.

In this chapter we discuss first the possibility of using the zero-one matrix test for time clustering adjusting for extraneous factors, and then we discuss the way to use this test in multivariate situations.

5.1 Zero-one matrix test for time clustering with adjustment for extraneous factors

One of the most important factors that should be taken into account when investigating disease patterns over a period of time is the change in population size. As discussed earlier in Chapter II, to adjust the zero-one matrix test for this change, the rates of disease in each unit should be used, rather than the frequencies themselves. However, unless the rates are known to have the same precision, an apparent gap within them may be

misleading as a basis for classification as zero or one in the zero-one matrix test. The larger the population base for the rate, the more stable the rate will be and also the more reliable will be the classification of the corresponding unit.

Symons, Grimson and Yuan (1982) recommended a way to classify units into high or normal risk of sudden infant death syndrome, adjusting for population sizes, which can be used in the zero-one matrix test. Instead of classifying units into "high" or "normal" risk, we classify them into one or zero; unusual "cluster" is found when one observes a few "high" risk units among many "normal" risk units, and "avoidance" (or "vacuity") is found when one observes a few "normal" risk units among many "high" risk units. The following discussion summarizes their method of classification:

The number of events, n_i , in the i^{th} unit having a population size of N_i , is assumed to have a Poisson distribution. The distribution of the number of events is

$$P(n_i | N_i, \lambda_g) = \exp(-N_i \lambda_g) (N_i \lambda_g)^{n_i} / n_i! , \quad (5.1)$$

where $i = 1, \dots, m$ indexes the units and λ_g is the population rate in the g^{th} group, $g=0$ or 1 for the zero or one value in the test matrix, which correspond to "normal" or "high" risk respectively.

If a randomly selected unit is "normal" risk with probability π_0 and "high" risk with probability $\pi_1 = 1 - \pi_0$, the unconditional distribution of n_i is then

$$f(n_i | \pi_g, \lambda_g, N_i) = \sum_{g=0}^1 \pi_g P(n_i | N_i, \lambda_g), \quad (5.2)$$

The cluster test is formulated as a problem of estimating the unknown mixture component origin of each n_i , denoted by $a_i = g$, or equivalently

n_i come from the g^{th} component of the mixture. The a priori probability of a_i equaling g is π_g . A likelihood of the data is required to estimate the m components of the vector $\underline{a} = (a_1, \dots, a_m)$ by a maximum likelihood or a Bayesian approach. Let the vector \underline{n} denote the m observations n_i , and \underline{N} be the vector of corresponding population size, and $\underline{\theta}$ be the vector of parameters $(\pi_0, \pi_1, \lambda_0, \lambda_1)$. The likelihood of the data for the m units, a specified allocation \underline{a} , parameters $\underline{\theta}$, and population size \underline{N} is given by

$$L(\underline{n} | \underline{\theta}, \underline{a}, \underline{N}) = \left\{ \prod_{g=0}^1 \pi_g^{m_g} \right\} \exp \left\{ - \sum_g \sum_{C_g} N_i \lambda_g + \sum_g \sum_{C_g} n_i \ln(N_i \lambda_g) \right\} \quad (5.3)$$

The m_g denotes the number of observations allocated to the g^{th} component by \underline{a} ; C_g is the collection of these observations. The first factor of (5.3) represents the likelihood of the *number* of observations assigned to each group and the remainder represents the likelihood of *which* of these numbers of observations are assigned to each group.

A maximum likelihood approach determines the maximum likelihood estimate of \underline{a} , $\hat{\underline{a}}$, as the allocation that maximizes (5.3). This is accomplished by replacing the parameters $\underline{\theta}$ by their maximum likelihood estimators, giving an allocation of the m n_i to the 2 components, and then seeking the maximum of (5.3) over the 2^m possible allocations. Given an allocation \underline{a} , the maximum likelihood estimates of the parameters are

$$\hat{\pi}_g = \frac{m_g}{\hat{m}} \quad (5.4)$$

and

$$\hat{\lambda}_g = \frac{\sum_{C_g} n_i}{\sum_{C_g} N_i} \quad (5.5)$$

The allocation $\hat{\underline{a}}$ that maximizes (5.3) is equivalent to the partition of the m observations into 2 groups which minimizes the negative of the

natural logarithm of (5.3), especially,

$$\sum_{i=1}^m n_i - \sum_{g=0}^1 \sum_{C_g} n_i \ln(N_i, \hat{\lambda}_g) - \sum_{g=0}^1 m_g \ln(m_g) \quad (5.6)$$

A Bayesian approach requires the specification of a prior distribution, $p(\underline{\theta})$, for $\underline{\theta}$ and the averaging over the uncertainty in the unknown parameters, since the allocation vector \underline{a} is of primary interest. Generally,

$$L(\underline{n}|\underline{a}, \underline{N}) = \int L(\underline{n}|\underline{\theta}, \underline{a}, \underline{N}) p(\underline{\theta}) d\underline{\theta}, \quad (5.7)$$

where the integration is over the parameter space of $\underline{\theta}$. The mode of (5.7) is taken as the Bayes estimate of the optimal allocation $\tilde{\underline{a}}$.

If the prior is chosen to delineate the parameter space as follows:

$$p(\underline{\theta}) = p(\pi_0, \pi_1, \lambda_0, \lambda_1) \propto \left[\prod_{g=0}^1 \pi_g^{-1} \right] \left[\prod_{g=0}^1 \lambda_g^{-1} \right], \quad (5.8)$$

then the resulting Bayesian optimal allocation, $\tilde{\underline{a}}$, is equivalent to the partition of the data into 2 groups that minimizes the criterion

$$\sum_{g=0}^1 \left(\sum_{C_g} n_i \right) \ln \left(\sum_{C_g} N_i \right) - \sum_{g=0}^1 \ln \left[\Gamma \left(\sum_{C_g} n_i \right) \right] - \sum_{g=0}^1 \ln \left[\Gamma(m_g) \right]. \quad (5.9)$$

A "marginal" or unconditional likelihood can be written as

$$L(\underline{n}|\underline{\theta}, \underline{N}) = \prod_{i=1}^m \left\{ \sum_{g=0}^1 \pi_g \exp(-N_i \lambda_g) (N_i \lambda_g)^{n_i/n_i!} \right\}. \quad (5.10)$$

An alternative large sample likelihood ratio procedure would be to assign n_i to the mixture component g when

$$\hat{\pi}_g P(n_i | N_i, \hat{\lambda}_g) = \max_{0 \leq j \leq 1} \{ \hat{\pi}_j P(n_i | N_i, \hat{\lambda}_j) \}, \quad (5.11)$$

where $\hat{\lambda}_j$ and $\hat{\pi}_j$ are the large sample maximum likelihood estimates or known parameter values.

The required maximum likelihood parameter estimates for (5.10) with the likelihood ratio procedure (5.11) can be obtained by an iterative scheme with initial estimates determined by preliminary descriptive analysis of the data, or more precise initial estimates by the method of moments. Alternatively, initial estimates based upon $\hat{\underline{a}}$ or $\tilde{\underline{a}}$ from criteria (5.6) or (5.9) respectively, may be employed. These are biased parameter estimates; however, they are found by Symons (1973) to work quite well as initial estimates for an iterative maximum likelihood scheme for a mixture of multivariate normals, probably because the bias in the estimates of the means tends to exaggerate slightly the separation of the mixture components.

It has been shown by Lindley (1965) and Cox and Hinkley (1975) that the Bayes estimates of the parameters in likelihood (5.10) with vague prior information will converge to the maximum likelihood estimates with large samples. Due to the computation difficulties in optimizing criteria (5.6) or (5.9) over 2^m allocations, given the desirable large sample properties of maximum likelihood estimates and the general optimality of the likelihood ratio criterion, the likelihood ratio procedure in (5.11) is preferable with large samples. With generally smaller sample sizes, the criteria (5.6) or (5.9) may be more reliable, as has been shown by Symons (1973) for a mixture of multivariate normals.

A computer program that performs the classification is available from Symons, Grimson and Yuan (1982); this program makes the classification using one of the three optimal criteria: maximum likelihood (5.6), Bayesian optimal allocation (5.9), and large sample likelihood ratio (5.11). The results of the following example are obtained using this computer program.

Example: Annual number of deaths due to large intestinal cancer in North Carolina from 1970 to 1978 for white males between the ages of 75 and 84, with the corresponding population size, are as follows:

<u>Year</u>	<u>Number of deaths</u>	<u>Population size</u>
1970	54	37286
1971	39	38053
1972	47	38815
1973	63	39582
1974	66	40431
1975	81	41188
1976	64	42077
1977	64	43029
1978	67	44370

Source: EPA data: combined population data from Census Bureau and mortality data from National Center for Health Statistics

When classifying into two groups, all 3 options of the classification program applied to these 9 time units (years) give the same results, as follows:

Group 1: 2 units 1971 and 1972

Group 2: 7 units 1970, 1973, 1974, 1975, 1976, 1977 and 1978

If one assigns a zero to each of the units in Group 1 and a one to each of the units in Group 2, the test statistic is $A=7$. The approximate expected value and variance of this statistic are (applying equations (3.5) and (3.2) and assuming the Decimal to be equal to zero):

$$E(A) = 0.6 \times m = 0.6 \times 9 = 5.4 ,$$

$$\text{Var}(A) = 0.155 \times E(A) = 0.155 \times 5.4 = .837 .$$

Under normal approximation, the Z value for the test statistic A is

$$Z = \frac{7-5.4}{\sqrt{.837}} = 1.75 ,$$

which corresponds to a p-value of 0.08 (only significant at the level $\alpha=.10$).

Note that the regular zero-one matrix test applied to this mortality data without adjusting for the population size would have given the statistic $A=6$ (zeros are assigned to years 1970, 1971 and 1972, and ones are assigned to years 1973, 1974, 1975, 1976, 1977 and 1978) which corresponds to a Z-value of $Z = \frac{6-5.4}{\sqrt{.837}} = 0.66$ and a p-value of 0.51.

For the data of this example, adjusting for population size using the regular zero-one matrix test on the death rates (instead of the death frequencies) would have given the same results as those obtained through the maximum likelihood methods. However, the maximum likelihood procedures would be more reliable in situations where population sizes change more drastically from one unit to another so that the rates do not have the same precision.

5.2 Zero-one matrix test in multivariate cases

There are situations in which the hypothesis is tested through several variables simultaneously. Examples are situations in which unusual patterns for a group of diseases are tested, but these diseases cannot be combined because of their differences in certain characteristics.

In this section we suggest a procedure for doing the test in these situations, using the zero-one matrix test.

Since the zero-one matrix test involves classifying data points into two groups, zero and one, in order to apply this test to a multivariate data set, we first must classify the data into two groups using a multivariate technique; then the significant level of the test is assessed, based on the number of data points in each group conditional on total number of data points used.

5.2.1 Classifying data points (time-space units) into two groups: There are

several methods to classify multivariate data into two groups using cluster analysis concepts. Classifying involves choosing a similarity or resemblance measure. Two units (data points) are similar if their profiles across variables are "close" or if they show "many" aspects in common, relative to those which other pairs of units share.

Generally, the most common type of measure for non-dichotomous data is the distance type measures. The Euclidean distance between two points i and j in a space of r dimensions is

$$d_{ij} = \left[\sum_{t=1}^r (X_{it} - X_{jt})^2 \right]^{1/2}$$

where X_{it} and X_{jt} are the projection points of i and j on variable t ($t=1,2,\dots,r$). The variables can be standardized to zero mean and unit standard deviation before applying the formula, since some researchers feel that the use of Euclidean distance measure should be restricted to orthogonal, standardized variables. This has the effect of assigning equal weights to all variables.

The ordinary squared Euclidean distances are similar to the Mahalanobis D^2 in the context of discriminant analysis. The difference is that the Mahalanobis D^2 takes both differences in axis lengths and the correlatedness of axes into account. The Mahalanobis D^2 is the same as ordinary Euclidean d^2 if the latter is computed in a space in which the configuration has been transformed into a hypersphere, as it has been shown by Green (1978).

Three different amalgamation rules for building up two clusters are single linkage, complete linkage and average linkage. Each rule can be described briefly as follows (see Green (1978)):

Single linkage: The single linkage or minimum distance rule starts out by finding the two points with the shortest distance. At the next stage, a third point joins the already formed cluster of two if its shortest distance

to the members of the cluster is smaller than the two closest unlinked points. Otherwise, the two closest unclustered points are placed into a cluster.

The process continues until all points end up in two clusters. The distance between two clusters is defined as the shortest distance from a point in the first cluster to a point in the second.

Complete linkage: The complete linkage rule uses a similar way of first clustering the two closest points. However, the criterion for joining points to clusters or clusters to clusters involves maximum (rather than minimum) distance. In other words, the distance between two clusters is the longest distance from a point of first cluster to a point in the second cluster.

Average linkage: The average linkage rule starts out the same way as the other two. However, in this case the distance between two clusters is the average distance from points in the first cluster to points in the second cluster.

5.2.2 Assessing the significant level of the test: The test statistic A is the size of the smaller cluster; conditional on the total number of data points m , the significant level (or P-value) of the test statistic A is

$$P = \sum_{j=0}^{A-1} \binom{m-2}{j} \left(\frac{1}{2}\right)^{m-2} \quad (5.12)$$

Proof of equation (5.12) is as follows:

Since we have m data points and we force them into two clusters, and since each cluster contains at least one point, there are only $(m-2)$ data points to be classified into the two clusters. Under the null hypothesis of no special pattern in the data, each of these $(m-2)$ points has equal probability (one half) of being classified into either of the two clusters. There-

fore, the problem becomes one of assigning $(m-2)$ "balls" randomly into two "urns". When the smaller cluster contains A of m points, it is equivalent to the fact that one of the two "urns" contains $(A-1)$ "balls" of the total $(m-2)$ "balls". The probability that one of the "urns" contains exactly $(A-1)$ "balls" is a binomial probability,

$$P(U = A-1) = \binom{m-2}{A-1} \left(\frac{1}{2}\right)^{A-1} \left(\frac{1}{2}\right)^{m-A-1} = \binom{m-2}{A-1} \left(\frac{1}{2}\right)^{m-2}$$

and the P-value of the statistic A is the probability that one of the "urns" contains $(A-1)$ balls or less:

$$P = P\{U \leq (A-1)\} = \sum_{j=0}^{A-1} \binom{m-2}{j} \left(\frac{1}{2}\right)^{m-2}$$

which is equation (5.12).

Example: One hundred counties in the state of North Carolina are divided into 8 health service regions, as in Figure 5.1. In this example each region is considered a space unit and we would like to test for space clustering of heart disease and cancer, based on the 1980 death rates reported by the North Carolina State Center for Health Statistics. The 1980 death rates by county in North Carolina for heart disease and cancer are presented in Tables 5.1 and 5.2 respectively. From these rates, the rate for each region is calculated as the average rate of all counties in that region. The results are presented in Table 5.3. The Euclidean distance between any two regions i and j , d_{ij} , is calculated as

$$d_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$$

$i = 1, 2, \dots, 8$
 $j = 1, 2, \dots, 8$
 $i \neq j$

where X_i is the heart disease death rate in region i , and Y_i is the cancer death rate in region i .

The matrix of distances between two regions is presented in Table 5.4.

Based on these distances and applying the single linkage procedure as described earlier in this section, we have the dendrogram as in Figure 5.2. From this dendrogram, when forced into two groups, we have Group 1 consisting of Regions 1, 2, 3, 4, 5, 6 and 8 - and Group 2 consisting of only Region 7 (the statistic A equals 1). The P-value for this observation under the null hypothesis of no clustering of heart disease and cancer is

$$P = \sum_{j=0}^{A-1} \binom{m-2}{j} \left(\frac{1}{2}\right)^{m-2} = \binom{6}{0} \left(\frac{1}{2}\right)^6 = 0.0156$$

which is significant at $\alpha = .05$.

These results indicate that there is significant clustering of heart disease and cancer in region 7, which consists of 16 counties of the north-east area of North Carolina. These results justify the concerns of the state health officials about the inadequate health care of residents of these low income counties which consist mostly of small farms. It is also noted that if one considers only the cancer death rates in the above 8 service regions and applies the regular univariate zero-one matrix test, using equations (3.5) and (3.2) and assuming that the Decimal equals zero, we have the observed statistic $A=2$ (regions 1 and 7 have high rates of cancer) and under the null hypothesis the approximate expected value and variance of A are

$$E(A) = m \times 0.6 = 8 \times 0.6 = 4.8,$$

$$\text{Var}(A) = 0.155 \times E(A) = 0.155 \times 4.8 = .744$$

which gives a Z value of

$$Z = \frac{2-4.8}{\sqrt{.744}} = -3.25.$$

These results show a significant clustering of cancer deaths in regions 1 and 7 while heart disease combined with cancer show a significant cluster only in region 7. Therefore it can be concluded that while region 7 is having a general health problem (which is represented by both heart and cancer diseases) region 1 may be having a cancer problem which justifies the concern of Schneider (1982). However, since there are many other possible reasons for the cancer problem in this region, it is presumptuous to conclude, like Schneider, that it is caused by the extensive use of agent white on the forest areas of western North Carolina.

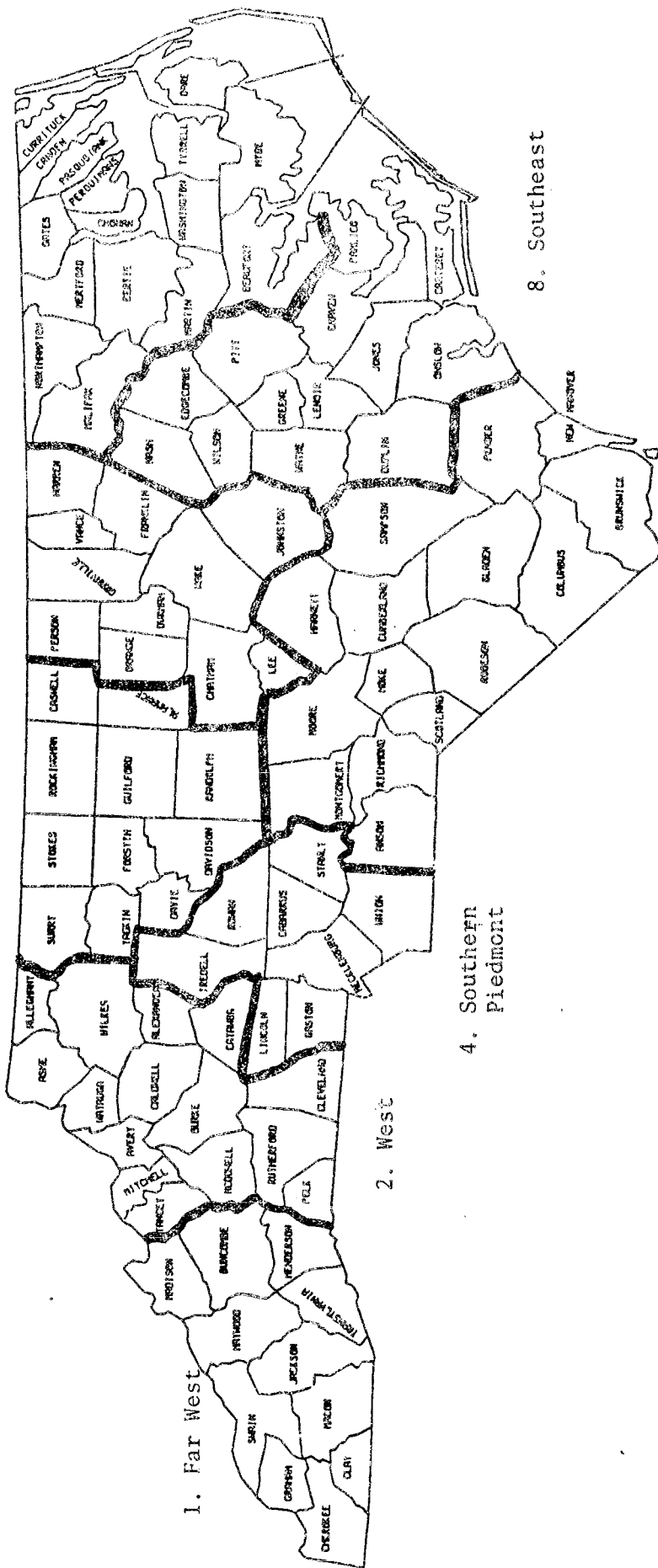
Figure 5.

NORTH CAROLINA COUNTIES AND EIGHT HEALTH SERVICE REGIONS

7. Northeast

5. Capitol

3. Piedmont



6. Cardinal

4. Southern Piedmont

2. West

TABLE 5.1

MORTALITY STATISTICS FOR 1980

NORTH CAROLINA RESIDENTS

DISEASES OF HEART

GEOGRAPHICAL AREA	NUMBER OF DEATHS 1980	DEATH RATE* 1980	COUNTIES (CONT'D)	NUMBER OF DEATHS 1980	DEATH RATE* 1980
NORTH CAROLINA	17579	299.24	42 HALIFAX	210	379.84
REGIONS			43 HARNETT	217	364.26
DHR I WESTERN	6236	309.81	44 HAYWOOD	160	344.13
DHR II N. CENTRAL	3836	304.22	45 HENDERSON	232	396.03
DHR III S. CENTRAL	3688	271.87	46 HERTFORD	80	342.39
DHR IV EASTERN	3819	306.95	47 HOKE	45	220.82
HSA I WESTERN	3329	327.30	48 HYDE	22	374.59
HSA II PIEDMONT	3246	291.20	49 IREDELL	248	300.47
HSA III S. PIEDMONT	2907	291.94	50 JACKSON	74	286.69
HSA IV CAPITAL	2394	292.81	51 JOHNSTON	310	439.11
HSA V CARDINAL	2550	284.09	52 JONES	40	412.24
HSA VI EASTERN	3153	305.61	53 LEE	147	400.37
COUNTIES			54 LENOIR	200	334.33
1 ALAMANCE	321	323.80	55 LINCOLN	146	344.57
2 ALEXANDER	64	256.01	56 MCDOWELL	103	293.16
3 ALLEGHANY	38	396.32	57 MACON	81	401.36
4 ANSON	97	379.48	58 MADISON	57	338.78
5 ASHE	78	349.36	59 MARTIN	104	400.80
6 AVERY	55	381.89	60 MECKLENBURG	981	242.66
7 BEAUFORT	157	389.90	61 MITCHELL	58	401.96
8 BERTIE	89	423.32	62 MONTGOMERY	96	427.25
9 BLADEN	119	390.81	63 MOORE	209	413.82
10 BRUNSWICK	128	357.83	64 NASH	215	320.17
11 BUNCOMBE	610	379.04	65 NEW HANOVER	304	293.80
12 BURKE	200	275.84	66 NORTHAMPTON	91	402.94
13 CABARRUS	315	366.73	67 ONSLOW	141	125.00
14 CALDWELL	175	258.31	68 ORANGE	135	175.19
15 CAMDEN	25	428.74	69 PAMLICO	50	480.86
16 CARTERET	145	352.89	70 PASQUOTANK	103	361.88
17 CASWELL	80	386.39	71 PENDER	78	351.12
18 CATAWBA	275	261.38	72 PERQUIMANS	27	284.75
19 CHATHAM	104	311.20	73 PERSON	112	384.00
20 CHEROKEE	56	295.79	74 PITT	241	288.10
21 CHOWAN	37	294.67	75 POLK	66	508.43
22 CLAY	34	513.44	76 RANDOLPH	222	241.67
23 CLEVELAND	292	349.98	77 RICHMOND	161	354.03
24 COLUMBUS	156	305.66	78 ROBESON	277	272.71
25 CRAVEN	154	216.76	79 ROCKINGHAM	258	309.25
26 CUMBERLAND	398	161.02	80 ROWAN	342	344.79
27 CURRITUCK	34	306.69	81 RUTHERFORD	184	342.08
28 DARE	54	403.70	82 SAMPSON	169	340.12
29 DAVIDSON	292	258.03	83 SCOTLAND	96	297.46
30 DAVIE	70	284.54	84 STANLY	175	360.67
31 DUPLIN	130	317.45	85 STOKES	87	262.97
32 DURHAM	455	297.81	86 SURRY	197	331.36
33 EDGECOMBE	183	326.85	87 SWAIN	36	350.19
34 FORSYTH	759	311.47	88 TRANSYLVANIA	53	226.32
35 FRANKLIN	121	402.58	89 TYRRELL	18	452.83
36 GASTON	517	318.02	90 UNION	183	260.02
37 GATES	36	405.58	91 VANCE	156	424.52
38 GRAHAM	20	277.23	92 WAKE	653	217.06
39 GRANVILLE	143	420.66	93 WARREN	58	357.31
40 GREENE	44	273.00	94 WASHINGTON	63	425.64
41 GUILFORD	876	276.20	95 WATAUGA	80	252.55
			96 WAYNE	268	276.14
			97 WILKES	205	349.48
			98 WILSON	192	304.11
			99 YADKIN	84	295.75
			100 YANCEY	43	287.75

* Per 100,000

TABLE 5.2
MORTALITY STATISTICS FOR 1980
NORTH CAROLINA RESIDENTS
CANCER

GEOGRAPHICAL AREA	NUMBER OF DEATHS 1980	DEATH RATE* 1980	COUNTIES (CONT'D)	NUMBER OF DEATHS 1980	DEATH RATE* 1980
NORTH CAROLINA	9698	165.08	42 HALIFAX	117	211.62
REGIONS			43 HARNETT	83	139.32
DHR I WESTERN	3382	168.02	44 HAYWOOD	83	178.52
DHR II N. CENTRAL	2158	171.14	45 HENDERSON	127	216.79
DHR III S. CENTRAL	1969	145.15	46 HERTFORD	44	188.31
DHR IV EASTERN	2189	175.94	47 NOKE	22	107.95
HSA I WESTERN	1764	173.43	48 HYDE	10	170.27
HSA II PIEDMONT	1899	170.36	49 IREDELL	134	162.35
HSA III S. PIEDMONT	1618	162.49	50 JACKSON	33	127.85
HSA IV CAPITAL	1268	155.08	51 JOHNSTON	130	184.14
HSA V CARDINAL	1393	153.19	52 JONES	21	216.42
HSA VI EASTERN	1756	170.20	53 LEE	77	209.71
COUNTIES			54 LENOIR	107	178.86
1 ALAMANCE	191	192.66	55 LINCOLN	60	141.60
2 ALEXANDER	39	156.00	56 MCDOWELL	46	130.92
3 ALLEGHANY	21	219.02	57 MACON	58	237.39
4 ANSON	40	156.48	58 MADISON	38	225.85
5 ASHE	41	183.64	59 MARTIN	49	188.83
6 AVERY	14	97.20	60 MECKLENBURG	618	152.86
7 BEAUFORT	87	216.06	61 MITCHELL	31	214.84
8 BERTIE	48	228.31	62 MONTGOMERY	45	200.27
9 BLADEN	51	167.49	63 MOORE	117	231.66
10 BRUNSWICK	81	226.44	64 NASH	121	180.19
11 BUNCOMBE	329	204.43	65 NEW HANOVER	205	198.12
12 BURKE	117	161.36	66 NORTHAMPTON	53	234.67
13 CABARRUS	163	189.76	67 ONSLOW	84	74.47
14 CALDWELL	91	134.32	68 ORANGE	84	109.01
15 CAMDEN	12	205.79	69 PAMLICO	18	173.11
16 CARTERET	78	189.83	70 PASQUOTANK	60	210.80
17 CASHWELL	31	149.72	71 PENDER	45	202.57
18 CATAWBA	144	136.87	72 PERQUIMANS	24	253.11
19 CHATHAM	43	128.67	73 PERSON	51	174.86
20 CHEROKEE	46	242.97	74 PITT	117	139.86
21 CHOWAN	42	334.50	75 POLK	21	161.77
22 CLAY	8	120.80	76 RANDOLPH	151	164.58
23 CLEVELAND	125	149.82	77 RICHMOND	68	193.50
24 COLUMBUS	102	199.85	78 ROBESON	149	146.69
25 CRAVEN	107	150.61	79 ROCKINGHAM	146	175.00
26 CUMBERLAND	225	91.03	80 ROWAN	181	182.48
27 CURRITUCK	31	279.63	81 RUTHERFORD	101	187.77
28 DARE	18	134.56	82 SAMPSON	95	191.19
29 DAVIDSON	174	153.75	83 SCOTLAND	45	139.43
30 DAVIE	38	154.46	84 STANLY	76	156.63
31 DUPLIN	76	185.58	85 STOKES	38	114.86
32 DURHAM	260	170.17	86 SURRY	131	220.34
33 EDGECOMBE	94	167.89	87 SWAIN	16	155.64
34 FORSYTH	391	160.45	88 TRANSYLVANIA	43	183.61
35 FRANKLIN	50	166.35	89 TYRRELL	11	276.72
36 GASTON	284	174.69	90 UNION	102	144.92
37 GATES	16	180.26	91 VANCE	66	179.60
38 GRAHAM	13	180.20	92 WAKE	415	137.95
39 GRANVILLE	51	150.02	93 WARREN	41	252.58
40 GREENE	22	136.50	94 WASHINGTON	33	222.95
41 GUILFORD	569	179.40	95 WATAUGA	44	138.90
			96 WAYNE	143	147.34
			97 WILKES	106	180.71
			98 WILSON	113	178.98
			99 YADKIN	39	137.13
			100 YANCEY	29	194.20

* Per 100,000

Table 5.3

1980 Heart Disease and Cancer Death Rate*
in North Carolina by Health Service Region

Region	1980 Heart Disease Death Rate	1980 Cancer Death Rate
1	313	193
2	332	163
3	306	164
4	354	163
5	334	169
6	348	173
7	371	221
8	330	163

*Per 100,000

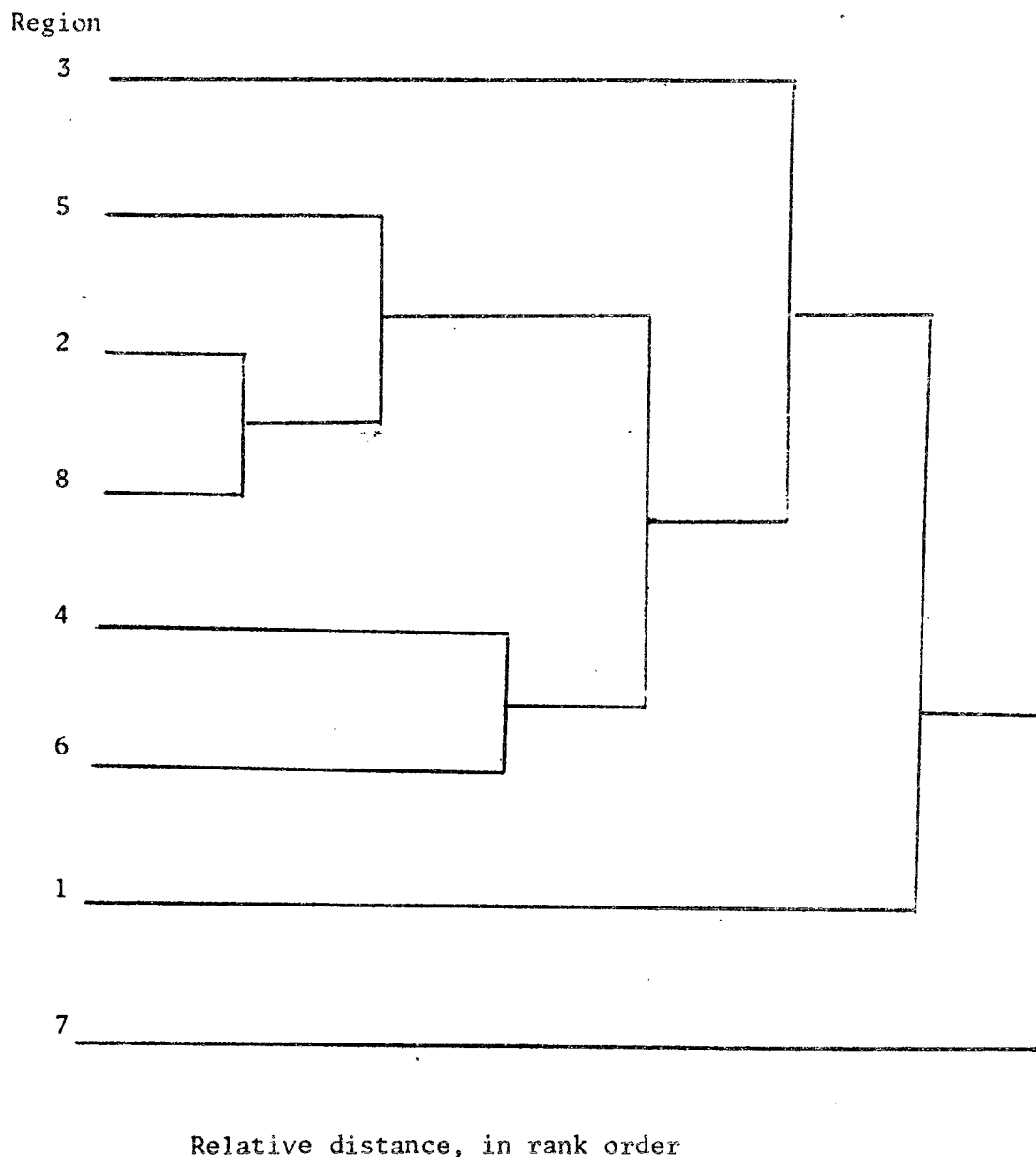
Table 5.4

Euclidean distance between 2 regions in North Carolina
based on heart disease and cancer death rates

Region	1	2	3	4	5	6	7	8
1	0	1261	890	2581	1017	1625	4148	1189
2	1261	0	677	484	40	356	4885	4
3	890	677	0	2305	809	1845	7474	577
4	2581	484	2305	0	436	136	3653	576
5	1017	40	809	436	0	212	4073	52
6	1625	356	1845	136	212	0	2833	424
7	4148	4885	7474	3653	4073	2833	0	5045
8	1189	4	577	576	52	424	5045	0

Figure 5.2

Dendrogram from single linkage clustering on Euclidean distance of heart disease and cancer death rates in the 8 Health Services Regions of the State of North Carolina



CHAPTER VI

SUMMARY, PRACTICAL GUIDE, AND SUGGESTIONS FOR FURTHER RESEARCH

6.0 Summary

Several tests have been suggested by different authors for testing or modeling the temporal and/or spatial clustering of disease. These included the Pinkel-Nefzger test, the Knox test, the Barton-David test, the Ederer-Myers-Mantel (EMM) test, the Grimson model modification of the EMM test, the Mantel test, the scan test and the Bailar-Eisenberg-Mantel test. Since these tests are sensitive to different types of clustering, it is important to have a clear understanding of the various concepts of temporal and spatial clusters and of "time-space interaction". In this research these concepts are deliberated when the tests are reviewed with the discussion of their differences and similarities.

Some weaknesses of the conventional tests are:

- They involve some subjective steps to determine the criteria.
- They are sensitive to time or space clustering as well as to time-space interaction.
- The distributions of the test statistics are so complicated that the significance levels cannot be assessed in most practical situations, even with approximation.
- They are a one-tail test of clustering, not sensitive to "vacuity".

Due to these weaknesses we suggest two new tests:

- Number-of-empty-cells test for rare events, and
- Zero-one matrix test for events with larger frequencies.

The number-of-empty-cells test is recommended for testing clusters of relatively rare events, i.e., the total number of diseases events is between one half and twice the total number of time-space units ($\frac{1}{2}m < n < 2m$). The zero-one matrix test is recommended for testing clusters of events with larger frequencies, i.e., when the number of events is greater than twice the number of time-space units ($n > 2m$).

In the next section of this chapter a step-by-step guide for testing is presented which summarizes the work done in this research on the above two tests.

6.1 Practical guide

Suppose we have n disease cases in m time-space subunits, and that each subunit has n_{ij} cases, where j indexes the time unit, $j = 1, 2, \dots, c$ and i indexes the space unit, $i = 1, 2, \dots, r$; $\sum_{j=1}^c n_{ij} = n_{i\cdot}$; $\sum_{i=1}^r n_{ij} = n_{\cdot j}$; $\sum_{j=1}^c \sum_{i=1}^r n_{ij} = \sum_{j=1}^c n_{\cdot j} = \sum_{i=1}^r n_{i\cdot} = n$.

The following steps are recommended for a test of clustering:

6.1.1 Step 1, specifying the alternative hypothesis: Depending on the circumstances, one may want to test for time clustering, for space clustering, or for time-space "interaction". When time (or space) clustering is being tested it is also necessary to specify whether the time (or space) clustering is being tested across all space (or time) units *simultaneously* or whether the clustering is being tested through individual space (or time) units. The difference between these two hypotheses is that the former seeks to answer the question "is there an unusual pattern over time where the pattern is the same for all space units?" while the latter seeks to answer the question

"is there some unusual pattern over time which may not be necessarily the same across all space units?" To answer the first question we can ignore the individual time-space frequencies n_{ij} and do the test on the marginal frequency on time $n_{.j}$, conditional on the total n and the number of time units c , as if we have only one space unit. To answer the second question the test should be done on each space unit separately, i.e., n_{ij} should be considered for each value of i separately ($i=1,2,\dots,r$) conditional on the number of time units c and $n_{i.}$; then the results are combined as in Step 5. This test is sensitive to both time clustering and time-space "interaction".

To test for time-space "interaction" n_{ij} should be considered, conditional on both marginal frequencies $n_{i.}$ and $n_{.j}$ and total number of time-space subunits $m = n \times c$.

Note that since the procedure to test for space patterns is similar to that for time patterns with the word "space" and "time" interchanged, and the subscripts i and j interchanged, in the following steps we discuss only the tests for time clustering and for time-space interaction.

6.1.2 Step 2, choosing a test: After deciding on the alternative hypothesis a test must be chosen based on the total frequency " n " and the total number of units " m " that the test uses as conditional factors. For example, if one tests for time clustering across all space units, then " n " is the above n and " m " is the number of time units c ; $n_{.j}$ are used as our observations. If the test for time clustering is done for each space unit i separately, then " n " is $n_{i.}$ and " m " is the number of time units c , and n_{ij} are used as our observations. If one tests for time-space "interaction", then " n " is the above total n and m is the above m , $m = r \times c$, and n_{ij} are again used as our observations.

Following are the recommended tests under different conditions:

- The number-of-empty-cells test is used when $\frac{1}{2}m < n < 2m$.
(Step 3)
- The zero-one matrix test is used when $n > 2m$.
(Step 4)
- For $n < \frac{1}{2}m$ the Knox test is suggested (Chapter 2, Section 2).

6.1.3 Step 3, the number-of-empty-cells test: In this step the notations n and m should be understood as the total frequencies and the total number of units (cells) respectively, which are the conditional factors on which the test is based. Depending on the hypothesis tested, n and m in this step correspond to the values specified in Step 2 for n and m .

- a) Test statistic: x = number of cells that contain no cases
- b) Exact distribution:

$$P(X=k) = \sum_{j=0}^{m-k-1} (-1)^j \binom{k+j}{k} \binom{m}{k+j} \left(1 - \frac{k+j}{m}\right)^n \quad \text{for } m \leq k \leq \max(0, m-n).$$

- c) Exact moments: The descending factorial moments of X are

$$E(X^{(R)}) = \left(\frac{m-R}{m}\right)^n m^{(R)}$$

which gives

$$E(X) = m \left(\frac{m-1}{m}\right)^n,$$

$$\text{Var}(X) = m(m-1) \left(\frac{m-2}{m}\right)^n + m \left(\frac{m-1}{m}\right)^n - m^2 \left(\frac{m-1}{m}\right)^{2n}.$$

- d) P-value of the test statistic x :

$$P = P(X \geq x) = \sum_{k=x}^m P(X=k).$$

6.1.4 Step 4, the zero-one matrix test:

a) Test statistic

(i) For time clustering across all space units:

$$A = \sum_{j=1}^c a_j$$

where

$$a_j = \begin{cases} 0 & \text{if } n_{\cdot j} < \left[\frac{n}{c} - 0.5 \right], \\ 1 & \text{if } n_{\cdot j} \geq \left[\frac{n}{c} - 0.5 \right], \end{cases}$$

(ii) For time clustering on an individual space unit:

$$A_i = \sum_{j=1}^c a_{ij}$$

where

$$a_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \left[\frac{n_{i \cdot}}{c} - 0.5 \right], \\ 1 & \text{if } n_{ij} \geq \left[\frac{n_{i \cdot}}{c} - 0.5 \right]. \end{cases}$$

(iii) For time-space "interaction":

$$C = \sum_{i=1}^c \sum_{j=1}^r a_{ij}$$

where

$$a_{ij} = \begin{cases} 0 & \text{if } n_{ij} < \left[\frac{n_{i \cdot} \cdot n_{\cdot j}}{m} - 0.5 \right] \\ 1 & \text{if } n_{ij} \geq \left[\frac{n_{i \cdot} \cdot n_{\cdot j}}{m} - 0.5 \right] \end{cases}$$

$[x]$ = smallest integer larger than x .

b) Exact distribution: Since the distribution of the test statistic C has not been investigated and since the statistic A and A_j are similar except for the differences in notations we only discuss the distribution of the statistic A . Note that A is equivalent to a special case of the

statistic C with $r = 1$, $m = c$. In order to make the notations consistent with those in the previous chapters we use r as 1 and m as c .

The exact distribution of A is

$$P(A=a) = \binom{m}{a} \sum_{j=0}^{m-a} (-1)^j \binom{m-a}{j} I_p^{a+j}(\rho, \nu)$$

where $p = \frac{1}{m}$, $\rho = \frac{n}{m}$ and $\nu = n$.

c) Exact moments: The descending factorial moment of A is

$$E(A^{(k)}) = m^{(k)} I_p^k(\rho, \nu)$$

which gives $E(A) = m I_p^1(\rho, \nu)$ and

$$\text{Var}(A) = m(m-1) I_p^2(\rho, \nu) - m I_p^1(\rho, \nu) \{m I_p^1(\rho, \nu) - 1\}$$

d) Approximate moments: Due to the complexity of the incomplete dirichlet integral involved in the exact distribution and moments of A , approximations are needed in their evaluations.

As discussed in Chapter 3, expected value and variance of A can be approximated as follows:

$$E(A) = m(0.6 + 0.3 \times \text{Decimal})$$

where Decimal is the difference between $\frac{n}{m}$ and $[\frac{n}{m} - 0.5]$ and $\text{Var}(A) = .155 \times E(A)$.

e) P-value: The exact distribution of A can be assessed based on its exact probability distribution function (2.7). However, due to its complexity we suggest approximating the P-value by first calculating the Z-value from the expectation and the variance of the test statistic, $Z = \frac{A - E(A)}{\sqrt{\text{Var}(A)}}$, then using the standard normal probability distribution to get the P-value corresponding to this Z .

6.1.5 Step 5, combining the time clustering results from all the space units: When the time clustering test is done separately on each individual space unit like the test statistic A_i in Step 4, the combined P-value of the test can be assessed using the procedure similar to that used for the EMM test: the combined statistic is a chi-square statistic with one degree of freedom:

$$\chi^2 = \frac{[\sum_{i=1}^r A_i - \sum_{i=1}^r E(A_i) - 0.5]^2}{\sum_{i=1}^r \text{Var}(A_i)}$$

Note that since this test is sensitive to both time clustering and time-space "interaction", it is not recommended; instead, the statistic A or C may be more appropriate, depending on the hypothesis being tested.

6.1.6 Step 6, adjusting for extraneous factors: In this research we discuss only the adjustment for a single, non-dichotomous variable such as population size. This adjustment is done by classifying the time-space subunit as zero or one, using the classification procedure recommended by Symons, Grimson and Yuan (1982). This procedure uses one of the following three criteria: maximum likelihood, large sample likelihood ratio and Bayes optimal allocation, based on the Poisson assumption regarding the distribution of the number of disease cases in a time-space subunit. In adjusting the population sizes, the regular zero-one matrix test summarized in Section 6.1.4 (Step 4) can be used with the *rates* of events replacing the *frequencies* of events; in this case the approximate first moment can be calculated using equation (3.5) and assuming the Decimal to be equal to zero. The procedure is only appropriate for situations in which the variation in the population size is not too large, i.e., the rates are calculated with comparable precision.

6.1.7 Step 7, multivariate procedure: To test for unusual patterns on several variables (diseases) simultaneously, the time-space subunits are classified into two groups based on their relative Euclidean distance from one to another. The subunits are linked into groups by using one of three linkage procedures: single linkage, complete linkage, and average linkage, until we have only two groups. The statistic is the size of the smaller group, namely A , and the P -value conditional on the total m subunits is

$$P = \sum_{j=0}^{A-1} \binom{m-2}{j} \left(\frac{1}{2}\right)^{m-2}$$

6.2 Suggestions for further research

There are several issues related to the zero-one matrix test for time-space clustering that need investigation before the test can be useful in most practical situations. Some of these issues are:

- Extensive tables are needed on the exact distributions and/or on the exact moments of the test statistic A , the test for time clustering.
- Exact distribution and moments of the test statistic C , the test for time-space "interaction", are still not known; this problem needs to be examined further so that an approximation procedure can be investigated based on the information from the exact distribution and/or moments.
- The computer program that classifies data into two groups, adjusting for population sizes, based on the maximum likelihood criteria, is now only available for use by its authors. This program needs to be documented and modified in such a way that it can be used as a subroutine by others who need to do the test.
- Test for time-space clustering adjusting for several extraneous factors simultaneously, using an approach similar to that of Symons,

Grimson and Yuan, can be very useful in many practical situations. This problem needs to be investigated further, especially the problem of adjusting for categorical variables such as race and sex.

```

ISN 0002      INTEGER IER
ISN 0003      REAL*8 X,A,B,P
ISN 0004      DIMENSION X(6),B(16)
ISN 0005      READ(1,100) (B(I),I=1,16) B = n = number of cases
ISN 0006      READ(1,100) (X(K),K=1,6) X = m = number of cells (time unit)
ISN 0007      100 FORMAT (20F3.0)           = 5,6,7,8,9,10
ISN 0008      DO 10 I=1,16
ISN 0009      DO 8 K=1,6
ISN 0010      XX=1./X(K)           XX = p = 1/m
ISN 0011      J=(B(I)*XX)+0.501    J= [n/2 - 0.5] = r
ISN 0012      BB=B(I)-J+1 BB = n-r+1
ISN 0013      AA=J
ISN 0014      CALL MDBeta (XX,AA,BB,P,IER) → I_p(r,n-r+1) = α = I_p^1(r,n)
ISN 0015      8 PRINT 200,XX,AA,BB,P,IER
ISN 0016      10 CONTINUE
ISN 0017      200 FORMAT (F15.4,2F10.0,F15.8,I5)
ISN 0018      STOP
ISN 0019      END

```

$$XX=p=\frac{1}{m}$$

$$AA=r=\lfloor \frac{n}{m}-0.5 \rfloor$$

$$n-r+1$$

$$I_p(r, n-r+1) = \text{Incomplete Beta} = I_p^1(r, n) = \alpha$$

0.2000	1.	5.	0.67231999
0.1667	1.	5.	0.59812234
0.1429	1.	5.	0.53733562
0.1250	1.	5.	0.48700105
0.1111	1.	5.	0.44507102
0.1000	1.	5.	0.40950938
0.2000	1.	6.	0.73785601
0.1667	1.	6.	0.66510193
0.1429	1.	6.	0.60343055
0.1250	1.	6.	0.55120466
0.1111	1.	6.	0.50672982
0.1000	1.	6.	0.46855889
0.2000	1.	7.	0.79028480
0.1667	1.	7.	0.72091828
0.1429	1.	7.	0.66008334
0.1250	1.	7.	0.60730414
0.1111	1.	7.	0.56153761
0.1000	1.	7.	0.52170299
0.2000	2.	7.	0.49668352
0.1667	1.	8.	0.76743190
0.1429	1.	8.	0.70864282
0.1250	1.	8.	0.65639107
0.1111	1.	8.	0.61025564
0.1000	1.	8.	0.56953268
0.2000	2.	8.	0.56379239
0.1667	2.	8.	0.45734112
0.1429	1.	9.	0.75026528
0.1250	1.	9.	0.69934223
0.1111	1.	9.	0.65356056
0.1000	1.	9.	0.61257939
0.2000	2.	9.	0.62419031
0.1667	2.	9.	0.51548312
0.1429	1.	10.	0.78594170
0.1250	1.	10.	0.73692445
0.1111	1.	10.	0.69205384
0.1000	1.	10.	0.65132145
0.2000	2.	11.	0.72512208
0.1667	2.	11.	0.61866723
0.1429	2.	11.	0.52819799
0.1250	2.	11.	0.45329611
0.1111	1.	12.	0.75668452
0.1000	1.	12.	0.71757035
0.2000	3.	13.	0.60197673
0.1667	3.	13.	0.46777497
0.1429	2.	14.	0.65336995
0.1250	2.	14.	0.57592229
0.1111	2.	14.	0.50869630
0.1000	2.	14.	0.45095680
0.2000	4.	15.	0.49897448
0.1667	3.	16.	0.59734551
0.1429	3.	16.	0.48546873
0.1250	2.	17.	0.67716031
0.1111	2.	17.	0.60993407
0.1000	2.	17.	0.54971592
0.2000	4.	17.	0.58855109
0.1667	3.	18.	0.67134075
0.1429	3.	18.	0.55960963
0.1250	3.	18.	0.46469144
0.1111	2.	19.	0.66809205
0.1000	2.	19.	0.60825281
0.2000	5.	21.	0.57932566
0.1667	4.	22.	0.61843334

Appendix 1.1

(continued)

Appendix 1.1

(continued)

0.1429	4.	22.	0.43806582
0.1250	3.	23.	0.62039099
0.1111	3.	23.	0.53624765
0.1000	3.	23.	0.46290575
0.2000	6.	25.	0.57243752
0.1667	5.	26.	0.57566003
0.1429	4.	27.	0.63827137
0.1250	4.	27.	0.52661953
0.1111	3.	28.	0.66279904
0.1000	3.	28.	0.58864854
0.2000	7.	29.	0.56715826
0.1667	6.	30.	0.53981207
0.1429	5.	31.	0.57314293
0.1250	4.	32.	0.65237532
0.1111	4.	32.	0.55507742
0.1000	4.	32.	0.46901487
0.2000	8.	33.	0.56285403
0.1667	7.	34.	0.50898270
0.1429	6.	35.	0.51665710
0.1250	5.	36.	0.57130526
0.1111	4.	37.	0.66290723
0.1000	4.	37.	0.57686905
0.2000	9.	37.	0.55928354
0.1667	8.	38.	0.48195700
0.1429	6.	40.	0.63691412
0.1250	6.	40.	0.49943124
0.1111	5.	41.	0.56390455
0.1000	5.	41.	0.47286181
0.2000	10.	41.	0.55625950
0.1667	8.	43.	0.60894070
0.1429	7.	44.	0.58433728
0.1250	6.	45.	0.60651295
0.1111	6.	45.	0.48613266
0.1000	5.	46.	0.56390127

EVEL 21.0 (JUN 74)

OS/360 FORTRAN H

```

      COMPILER OPTIONS - NAME= MAIN,OPTAB2,LINECNT=58,SIZE=9900K,
                        SOURCE,EBCDIC,NOLIST,DECK,LOAD,MAP,QUEDIT,IO
ISN 0002      INTEGER R,M,K,F1,MK,I,J,F3,F5,F6,N
ISN 0003      REAL*8 P,SUM1,SUM2,MFAC,MKFAC,IFAC,JFAC,A,B,AAA
ISN 0004      DIMENSION XX(6),AA(10),BS(16)
ISN 0005      READ(1,100) (BB(I),I=1,16)   Read in n
ISN 0006      READ(1,100) (XX(K),K=1,6)   Read in m
ISN 0007      100 FORMAT (20F3,0)
ISN 0008      DO 10 II=1,16
ISN 0009      DO 8 KK=1,6
ISN 0010      P=1./XX(KK)   p = 1/m
ISN 0011      N=BB(II)
ISN 0012      AAA=(N*P)+0.501   AAA = r = [n/m - 0.5]
ISN 0013      R=AAA           r
ISN 0014      M=N-(R*2)      m = (n-2r)
ISN 0015      SUM2=0
ISN 0016      MFAC=1
ISN 0017      DO 1 F1=1,M,1
ISN 0018      MFAC=MFAC*F1   m! = (n-2r)!
ISN 0019      1 CONTINUE
ISN 0020      DO 2 K=0,M,1
ISN 0021      MK=M-K        + n - 2r - k
ISN 0022      MKFAC=1
ISN 0023      DO 3 F3=1,MK,1
ISN 0024      MKFAC=MKFAC*F3   (n-2r-k)!
ISN 0025      3 CONTINUE
ISN 0026      SUM1=0
ISN 0027      DO 4 I=0,K,1
ISN 0028      J=K-I         j = k-i
ISN 0029      IFAC=1
ISN 0030      DO 5 F5=1,I,1
ISN 0031      IFAC=IFAC*F5     i!
ISN 0032      5 CONTINUE
ISN 0033      JFAC=1
ISN 0034      DO 6 F6=1,J,1
ISN 0035      JFAC=JFAC*F6     (k-i)!
ISN 0036      6 CONTINUE
ISN 0037      A=MFAC/(IFAC*JFAC*MKFAC*(I+R)*(J+R))
ISN 0038      SUM1=SUM1+A
ISN 0039      4 CONTINUE
ISN 0040      B=((( -1)**K)*(P**K))*SUM1
ISN 0041      SUM2=SUM2+B
ISN 0042      2 CONTINUE
ISN 0043      8 PRINT 200,P,R,M,SUM2
ISN 0044      10 CONTINUE
ISN 0045      200 FORMAT (F15.4,2(I10,1X),F15.10)
ISN 0046      STOP
ISN 0047      END

```

$$I_p^2(r,n) = \frac{n!p^{2r}}{(n-2r)![(r-1)!]^2} \text{ [Sum-2]}$$

$$\text{where Sum-2} = \sum_{k=0}^{n-2r} -p^k \sum_{i=0}^k \frac{(n-2r)!}{i!(k-i)!(n-2r-k)!(r+i)(r+k-i)}$$

Appendix 1.2

(continued)

0.2000	1	3	0.5207000212
0.1667	1	3	0.5282779556
0.1429	1	3	0.5304836828
0.1250	1	3	0.6767573125
0.1111	1	3	0.7072189451
0.1000	1	3	0.7335000339
0.2000	1	4	0.4355066880
0.1667	1	4	0.5015947348
0.1429	1	4	0.5547966525
0.1250	1	4	0.5981608073
0.1111	1	4	0.6340598662
0.1000	1	4	0.6642067658
0.2000	1	5	0.3622400218
0.1667	1	5	0.4288838315
0.1429	1	5	0.4842030317
0.1250	1	5	0.5304260254
0.1111	1	5	0.5694254117
0.1000	1	5	0.6026701100
0.2000	2	4	0.0759866726
0.1667	1	6	0.3689244624
0.1429	1	6	0.4244153221
0.1250	1	6	0.4718800933
0.1111	1	6	0.5126590666
0.1000	1	6	0.5479246532
0.2000	2	5	0.0576120690
0.1667	2	5	0.0749166368
0.1429	1	7	0.3735779643
0.1250	1	7	0.4211280743
0.1111	1	7	0.4626910119
0.1000	1	7	0.4991344971
0.2000	2	6	0.0440481572
0.1667	2	6	0.0599016242
0.1429	1	8	0.3301810664
0.1250	1	8	0.3770043466
0.1111	1	8	0.4186087544
0.1000	1	8	0.4555749048
0.2000	2	8	0.0263973820
0.1667	2	8	0.0387133032
0.1429	2	8	0.0508962416
0.1250	2	8	0.0624469405
0.1111	1	10	0.3450949885
0.1000	1	10	0.3817125519
0.2000	3	9	0.0056566610
0.1667	3	9	0.0096004484
0.1429	2	11	0.0297363400
0.1250	2	11	0.0387568231
0.1111	2	11	0.0476846969
0.1000	2	11	0.0563151860
0.2000	4	10	0.0016561661
0.1667	3	12	0.0045745385
0.1429	3	12	0.0073255441
0.1250	2	14	0.0246354056
0.1111	2	14	0.0316495077
0.1000	2	14	0.0387585661
0.2000	4	12	0.0003561174
0.1667	3	14	0.0028525668
0.1429	3	14	0.0048447449
0.1250	3	14	0.0072215496
0.1111	2	16	0.0243271046
0.1000	2	16	0.0304554752
0.2000	5	15	0.0001412471
0.1667	4	17	0.0004915251
0.1429	4	17	0.0010055813
0.1250	3	19	0.0030018755

Appendix 1.2

(continued)

0.1111	3	19	0.0044622666
0.1000	3	19	0.0061441235
0.2000	6	18	0.0000245727
0.1667	5	20	0.0000951020
0.1429	4	22	0.0003357039
0.1250	4	22	0.0005513420
0.1111	3	24	0.0021069905
0.1000	3	24	0.0030382001
0.2000	7	21	0.0000044308
0.1667	6	23	0.0000203709
0.1429	5	25	0.0000740511
0.1250	4	27	0.0002591692
0.1111	4	27	0.0004714494
0.1000	4	27	0.0007654013
0.2000	8	24	0.0000008198
0.1667	7	26	0.0000045532
0.1429	6	28	0.0000178864
0.1250	5	30	0.0000612577
0.1111	4	32	0.0002116058
0.1000	4	32	0.0003669105
0.2000	9	27	0.0000001547
0.1667	8	29	0.0000010582
0.1429	6	33	0.0000055446
0.1250	6	33	0.0000162560
0.1111	5	35	0.0000530452
0.1000	5	35	0.0001022453
0.2000	10	30	0.0000000296
0.1667	8	34	0.0000002380
0.1429	7	36	0.0000013613
0.1250	6	38	0.0000059106
0.1111	6	38	0.0000151109
0.1000	5	40	0.0000473766

p r n-2r Sum-2

Appendix 1.3

SAS Program to calculate $I_p^2(r,n)$ from Sum-2

$$I_p^2(r,n) = \frac{n!p}{(n-r)![(r-1)!]^2} = \frac{n!}{(n-2r)![(r-1)!]^2} m^2$$

```

101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

NOTE: THE PROCEDURE PRINT USED 0.46 SECONDS AND 194K AND PRINTED PAGES 1 TO 2.

NOTE: THE PROCEDURE PRINT USED 0.49 SECONDS AND 202K AND PRINTED PAGES 3 TO 4.

NOTE: THE PROCEDURE PRINT USED 0.24 SECONDS AND 178K.

NOTE: THE PROCEDURE PRINT USED 0.22 SECONDS AND 178K.

NOTE: THE PROCEDURE PRINT USED 0.41 SECONDS AND 194K AND PRINTED PAGES 5 TO 6.

Appendix 1.3 (continued)

OHS	P	R	II	4	N	SUM2	IZ
1	0.2000	1	0.677320	5	5	0.526900	0.422400
2	0.1667	1	0.596122	6	5	0.500270	0.327932
3	0.1429	1	0.557336	7	5	0.638484	0.260606
4	0.1250	1	0.487091	8	5	0.676758	0.211487
5	0.1111	1	0.445071	9	5	0.707819	0.174770
6	0.1000	1	0.409510	10	5	0.733500	0.146700
7	0.2000	1	0.737856	5	6	0.435307	0.522368
8	0.1667	1	0.665102	6	6	0.531595	0.417926
9	0.1429	1	0.603431	7	6	0.554797	0.336671
10	0.1250	1	0.551235	8	6	0.598161	0.280368
11	0.1111	1	0.506730	9	6	0.634060	0.234837
12	0.1000	1	0.468559	10	6	0.664207	0.199262
13	0.2000	1	0.790285	5	7	0.362240	0.608563
14	0.1667	1	0.728918	6	7	0.428884	0.500364
15	0.1429	1	0.667083	7	7	0.484203	0.415031
16	0.1250	1	0.607304	8	7	0.530426	0.348032
17	0.1111	1	0.561538	9	7	0.569425	0.295258
18	0.1000	1	0.521703	10	7	0.602670	0.253121
19	0.2000	2	0.496684	5	8	0.675987	0.204252
20	0.1667	1	0.767432	6	8	0.368974	0.573802
21	0.1429	1	0.708643	7	8	0.424415	0.485046
22	0.1250	1	0.656391	8	8	0.471860	0.412895
23	0.1111	1	0.610256	9	8	0.512659	0.354431
24	0.1000	1	0.569533	10	8	0.547925	0.306638
25	0.2000	2	0.563792	5	9	0.057612	0.278750
26	0.1667	2	0.457341	6	9	0.074917	0.174805
27	0.1429	1	0.759265	7	9	0.373578	0.548931
28	0.1250	1	0.699342	8	9	0.421128	0.471769
29	0.1111	1	0.653561	9	9	0.462691	0.411281
30	0.1000	1	0.612579	10	9	0.499134	0.359377
31	0.2000	2	0.624190	5	10	0.044048	0.355204
32	0.1667	2	0.515483	6	10	0.059802	0.252562
33	0.1429	1	0.785942	7	10	0.330181	0.606455
34	0.1250	1	0.736924	8	10	0.377004	0.530162
35	0.1111	1	0.692054	9	10	0.418609	0.465121
36	0.1000	1	0.651321	10	10	0.455575	0.410017
37	0.2000	2	0.725122	5	12	0.026397	0.501761
38	0.1667	2	0.618667	6	12	0.038713	0.354372
39	0.1429	2	0.528198	7	12	0.050496	0.251331
40	0.1250	2	0.453296	8	12	0.062447	0.181121
41	0.1111	1	0.756685	9	12	0.5095	0.562377
42	0.1000	1	0.717370	10	12	0.381713	0.503961
43	0.2000	3	0.601977	5	15	0.005657	0.326149
44	0.1667	3	0.467775	6	15	0.009600	0.185379
45	0.1429	2	0.653370	7	15	0.029736	0.475732
46	0.1250	2	0.575722	8	15	0.039757	0.105979
47	0.1111	2	0.504696	9	15	0.047685	0.238096
48	0.1000	2	0.459397	10	15	0.056315	0.184489
49	0.2000	4	0.498974	5	18	0.001656	0.287787
50	0.1667	3	0.597346	6	18	0.004575	0.327630
51	0.1429	3	0.485469	7	18	0.007326	0.209363
52	0.1250	2	0.677160	8	18	0.024635	0.441705

Appendix 1.3 (continued)

CRS	P	R	II	M	N	SU42	I2
57	0.1429	3	0.559610	7	20	0.0048447	0.297302
58	0.1250	3	0.464691	8	20	0.0072215	0.192197
59	0.1111	2	0.666792	9	20	0.0243272	0.431169
60	0.1000	2	0.600253	10	20	0.0304555	0.354136
61	0.2000	5	0.579326	5	25	0.0001412	0.297854
62	0.1667	4	0.618433	6	25	0.0004915	0.354495
63	0.1429	4	0.488066	7	25	0.0010335	0.210867
64	0.1250	3	0.620391	8	25	0.0030019	0.365042
65	0.1111	3	0.536248	9	25	0.0044623	0.267665
66	0.1000	3	0.462906	10	25	0.0061443	0.195869
67	0.2000	6	0.572488	5	30	0.0001246	0.289581
68	0.1667	5	0.575661	6	30	0.0000962	0.301089
69	0.1429	4	0.638271	7	30	0.0003387	0.385147
70	0.1250	4	0.526615	8	30	0.0006512	0.254456
71	0.1111	3	0.662799	9	30	0.0021070	0.423742
72	0.1000	3	0.588649	10	30	0.0030883	0.330072
73	0.2000	7	0.567158	5	35	0.0000944	0.283221
74	0.1667	6	0.535812	6	35	0.0001204	0.259759
75	0.1429	5	0.573143	7	35	0.0001741	0.303191
76	0.1250	4	0.582375	8	35	0.0002592	0.407203
77	0.1111	4	0.550077	9	35	0.0004714	0.288697
78	0.1000	4	0.469015	10	35	0.0007654	0.201761
79	0.2000	8	0.562954	5	40	0.0000008	0.278142
80	0.1667	7	0.508983	6	40	0.0000946	0.226757
81	0.1429	6	0.516657	7	40	0.0000179	0.240153
82	0.1250	5	0.571305	8	40	0.0000613	0.304666
83	0.1111	4	0.662907	9	40	0.0002118	0.423807
84	0.1000	4	0.576869	10	40	0.0003669	0.316032
85	0.2000	9	0.559284	5	45	0.0000002	0.274043
86	0.1667	8	0.481957	6	45	0.0000011	0.199784
87	0.1429	6	0.626914	7	45	0.0000055	0.383228
88	0.1250	6	0.499431	8	45	0.0000163	0.226318
89	0.1111	5	0.569905	9	45	0.0000530	0.305758
90	0.1000	5	0.472862	10	45	0.0001022	0.205494
91	0.2000	10	0.556259	5	50	0.0000000	0.270259
92	0.1667	8	0.608941	6	50	0.0000002	0.342142
93	0.1429	7	0.584337	7	50	0.0000014	0.316560
94	0.1250	6	0.606513	8	50	0.0000059	0.347331
95	0.1111	6	0.466133	9	50	0.0000151	0.216046
96	0.1000	5	0.568801	10	50	0.0000474	0.306599

Appendix

Standardized residuals from the regression estimate of the expected value and variance of A.

OBS	H	H	R	I1	I2	ERA	EVA	DECIMAL	PEA	PFA	FEI	REZ
1	5	1	1	0.672320	0.424400	3.16160	0.50925	0.00000	3.0	0.4656	-10.757	-8.689
2	6	5	1	0.588122	0.327932	3.50873	0.54769	-0.16667	3.3	0.5115	-6.046	-6.607
3	7	5	1	0.537336	0.260606	3.76135	0.55904	-0.26571	3.6	0.5580	-4.290	-0.186
4	8	5	1	0.487091	0.211487	3.89673	0.55550	-0.37500	3.9	0.6045	0.088	8.821
5	9	5	1	0.445071	0.174770	4.00564	0.54394	-0.44444	4.2	0.6510	4.152	19.682
6	10	5	1	0.409510	0.146700	4.09510	0.52857	-0.50000	4.5	0.6975	4.087	32.036
7	5	6	1	0.37856	0.122168	4.0926	0.52585	-0.20000	3.3	0.5115	10.552	2.730
8	6	6	1	0.665102	0.417996	3.99061	0.60750	0.00000	3.6	0.5580	-9.788	-7.845
9	7	6	1	0.339671	0.22401	4.22401	0.64792	-0.14286	3.9	0.6045	-7.671	-6.701
10	8	6	1	0.551205	0.280388	4.40964	0.66646	-0.25000	4.2	0.6510	-4.754	-2.319
11	9	6	1	0.506730	0.234837	4.56057	0.67005	-0.33333	4.5	0.6975	-1.328	4.097
12	10	6	1	0.468559	0.199262	4.68559	0.66443	-0.40000	4.8	0.7440	2.442	11.876
13	5	7	1	0.790285	0.608563	3.95142	0.50894	-0.30000	5.1	0.7905	-2.243	-3.515
14	6	7	1	0.720918	0.500364	4.32551	0.62641	-0.40000	4.4	0.7440	2.442	11.876
15	7	7	1	0.660083	0.415031	4.52056	0.70210	0.16667	3.6	0.5580	-8.894	9.640
16	8	7	1	0.607304	0.348092	4.85843	0.74722	0.00000	3.9	0.6045	-9.837	-3.498
17	9	7	1	0.561538	0.292258	5.05384	0.77116	-0.12500	4.5	0.6975	-7.378	-7.278
18	10	7	1	0.521703	0.253121	5.21703	0.78056	-0.22222	4.8	0.7440	-5.023	-3.515
19	5	8	2	0.496683	0.204252	2.48342	0.40110	-0.30000	5.1	0.7905	-2.243	-3.515
20	6	8	1	0.767432	0.573682	4.60459	0.61880	0.40000	2.4	0.3720	-3.359	-7.255
21	7	8	1	0.708643	0.485046	4.96050	0.72586	0.14286	4.5	0.6975	-8.787	-5.203
22	8	8	1	0.656391	0.412895	5.25113	0.79890	0.00000	4.8	0.7440	-8.591	-3.909
23	9	8	1	0.610256	0.354431	5.49230	0.84596	-0.11111	5.1	0.7905	-7.143	-6.872
24	10	8	1	0.569533	0.306838	5.69533	0.87398	-0.20000	5.4	0.8370	-5.185	-6.556
25	5	9	2	0.563792	0.278750	2.81896	0.47422	-0.50000	2.7	0.4185	-4.220	-6.404
26	6	9	2	0.457341	0.174805	2.74405	0.45842	-0.28571	4.8	0.7440	-1.605	-8.708
27	7	9	1	0.750265	0.546931	5.25184	0.82471	0.12500	5.1	0.7905	-8.787	-5.203
28	8	9	1	0.699342	0.473769	5.59474	0.84871	0.00000	5.4	0.8370	-2.627	-6.277
29	9	9	1	0.653561	0.411281	5.88204	0.89582	0.14286	5.1	0.7905	-8.843	-4.149
30	10	9	1	0.612579	0.359377	6.12579	0.94436	-0.10000	5.7	0.8835	-6.951	-6.566
31	5	10	2	0.624190	0.352204	3.12095	0.48470	-0.10000	5.4	0.8370	-8.195	-6.444
32	6	10	2	0.515483	0.232562	3.09296	0.50373	0.00000	3.0	0.4650	-3.875	-4.064
33	7	10	1	0.785942	0.608455	5.50159	0.82890	-0.33333	3.0	0.4650	-3.004	-7.689
34	8	10	1	0.736924	0.530162	5.89540	0.87059	0.42857	5.1	0.7905	-7.300	12.098
35	9	10	1	0.692054	0.465121	6.24844	0.92316	0.25000	5.4	0.8370	-9.403	0.390
36	10	10	1	0.651321	0.410317	6.51321	0.99282	0.11111	5.7	0.8835	-8.485	-4.297
37	5	12	2	0.725122	0.501761	3.62561	0.51579	0.00000	0.0	0.5400	-7.880	-6.327
38	6	12	2	0.618667	0.354872	3.71200	0.57919	0.40000	3.6	0.5580	-0.706	8.184
39	7	12	2	0.528198	0.251831	3.69733	0.60364	-0.50000	3.6	0.5580	-3.017	-3.659
40	8	12	2	0.453296	0.181120	3.62637	0.61857	-0.23571	3.6	0.5580	-2.634	-7.562
41	9	12	1	0.756684	0.562377	6.81016	0.67057	0.33333	6.1	0.9765	-0.727	-9.791
42	10	12	1	0.717570	0.503861	7.17570	0.73826	0.20000	6.6	1.0230	-8.023	-0.914
43	5	15	3	0.601977	0.326149	3.00980	0.47347	0.00000	3.0	0.4650	-0.328	-1.790
44	6	15	3	0.467775	0.165379	2.80665	0.49074	-0.50000	2.7	0.4185	-3.600	-14.720
45	7	15	2	0.653370	0.405732	4.57359	0.59361	0.14286	4.5	0.6975	-1.609	0.128
46	8	15	2	0.575922	0.309979	4.60788	0.73826	-0.12500	4.5	0.6975	-4.231	-5.521
47	9	15	2	0.508696	0.210096	4.50957	0.77733	-0.33333	4.5	0.6975	-1.710	-8.106
48	10	15	2	0.450957	0.184448	4.57827	0.76088	-0.40000	4.5	0.6975	-0.212	-10.270
49	5	18	4	0.498974	0.207737	2.49487	0.42623	-0.40000	2.4	0.3720	-3.003	-12.734
50	6	18	4	0.593795	0.327630	3.56437	0.46740	0.00000	3.6	0.5580	0.944	-1.656
51	7	18	4	0.465409	0.208663	3.50863	0.46740	-0.00000	3.6	0.5580	0.944	-1.656
52	8	18	4	0.617160	0.448705	5.41724	0.80863	-0.42857	5.4	0.8370	-2.242	-13.103
53	9	18	2	0.609334	0.434206	5.41724	0.80863	-0.20000	5.4	0.8370	-1.669	1.669
54	10	18	2	0.549716	0.344641	5.49716	0.80863	0.00000	5.4	0.8370	-1.669	1.669
55	5	20	4	0.588551	0.303213	2.94275	0.46722	-0.00000	1.0	0.4650	1.945	-0.474

Appendix 2 (continued)

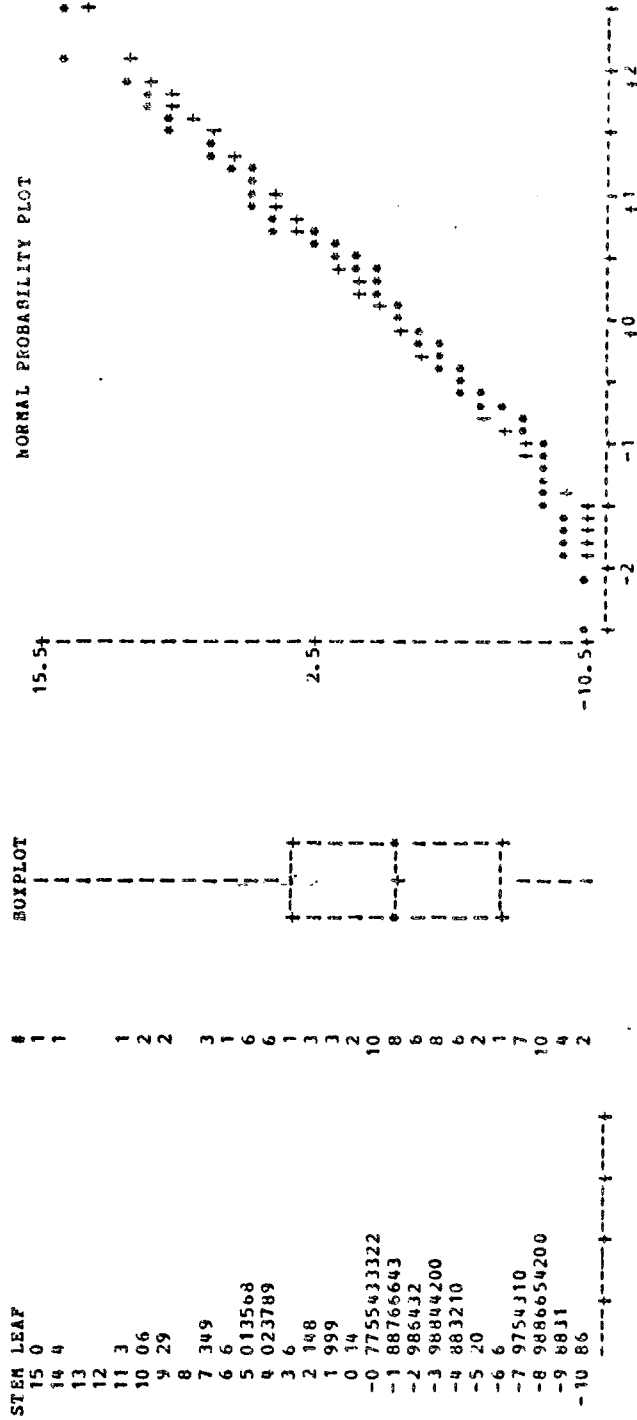
OBS	M	N	R	LE	L2	EEA	EVA	DECIMAL	PEA	PVA	RE1	RE2
56	6	20	3	0.671341	0.426564	4.62804	0.599834	0.33333	4.2	0.6510	4.2690	8.530
57	7	20	3	0.559610	0.287302	3.91727	0.633976	-0.14236	3.9	0.6045	-0.4403	-5.195
58	8	20	3	0.464691	0.192397	3.71753	0.660526	-0.50690	3.6	0.5580	-3.1616	-15.522
59	9	20	2	0.668092	0.431149	6.01203	0.931424	0.22222	6.0	0.9300	-0.2733	3.170
60	10	20	2	0.608253	0.354136	6.08253	0.957644	0.00000	6.0	0.9300	-1.3568	-2.887
61	5	25	5	0.579326	0.297854	4.89063	0.463248	0.00000	3.0	0.4650	3.5087	0.378
62	6	25	4	0.618443	0.354495	3.71060	0.576903	0.16667	3.9	0.6045	5.3041	4.784
63	7	25	4	0.488066	0.210807	3.41646	0.600668	-0.42857	3.3	0.5115	-3.4088	-14.845
64	8	25	3	0.620391	0.365042	4.96313	0.772852	0.12500	5.1	0.7905	2.7578	2.281
65	9	25	3	0.536248	0.267665	4.82623	0.805621	-0.22222	4.8	0.7440	-0.5435	-7.649
66	10	25	3	0.462906	0.195809	4.62906	0.839071	-0.50000	4.5	0.6975	-2.7880	-15.870
67	5	30	6	0.572407	0.289561	2.86244	0.460500	0.00000	3.0	0.4650	4.0058	0.977
68	6	30	5	0.575661	0.301089	3.45379	0.56745	0.00000	3.6	0.5580	4.2280	0.225
69	7	30	4	0.638271	0.385147	4.46790	0.681959	0.20571	4.6	0.7440	7.4330	9.097
70	8	30	4	0.526614	0.254456	4.21292	0.713815	-0.25000	4.2	0.6510	-0.3066	-8.800
71	9	30	3	0.662799	0.423742	5.26519	0.891137	0.33333	6.3	0.9765	5.0127	9.579
72	10	30	3	0.588648	0.330072	5.88648	0.942228	0.00000	6.0	0.9300	1.9284	-1.298
73	5	35	7	0.567158	0.283221	2.83579	0.498504	0.00000	3.0	0.4650	5.7906	1.417
74	6	35	6	0.539812	0.259759	3.23807	0.541354	-0.16667	3.3	0.5115	1.8873	-5.515
75	7	35	5	0.573143	0.303190	4.01200	0.649054	0.00000	4.2	0.6510	4.6859	0.176
76	8	35	4	0.652375	0.407203	5.21900	0.744376	0.37500	5.7	0.8835	9.2163	12.637
77	9	35	4	0.555077	0.208697	4.99570	0.824399	-0.11111	5.1	0.7905	2.0879	-4.170
78	10	35	4	0.469015	0.201761	4.69015	0.851116	-0.50000	4.5	0.6975	-4.0542	-18.049
79	5	40	8	0.562854	0.276142	4.31427	0.456988	0.00000	3.0	0.4650	6.5996	1.753
80	6	40	7	0.508983	0.226757	3.05390	0.530317	-0.33333	3.0	0.4650	-1.7648	-12.317
81	7	40	6	0.516657	0.240153	3.61660	0.623237	-0.28571	3.6	0.5580	-0.4590	-10.467
82	8	40	5	0.571305	0.304665	4.57044	0.742771	0.00000	4.8	0.7440	5.0227	0.165
83	9	40	4	0.662907	0.423807	5.96616	0.865176	0.44444	6.6	1.0280	10.6238	15.570
84	10	40	4	0.576869	0.316032	5.76869	0.933774	0.00000	6.0	0.9300	4.0097	-0.404
85	5	45	9	0.559283	0.274043	2.79642	0.457320	0.00000	3.0	0.4650	7.2301	1.679
86	6	45	8	0.481957	0.199783	2.89174	0.523077	-0.50000	2.7	0.4185	-6.6307	-19.993
87	7	45	6	0.636914	0.383228	4.45840	0.76662	0.42657	5.1	0.7905	14.3908	16.824
88	8	45	6	0.499431	0.223318	3.99545	0.705632	-0.37500	3.9	0.6045	-2.3890	-14.332
89	9	45	5	0.569904	0.305758	5.12914	0.635625	0.00000	5.4	0.8370	5.2806	0.165
90	10	45	5	0.472842	0.205494	4.72962	0.863289	-0.50000	4.5	0.6975	-4.8348	-19.204
91	5	50	10	0.556259	0.270259	2.78130	0.450864	0.00000	3.0	0.4650	7.8633	3.135
92	6	50	8	0.608941	0.342142	3.55364	0.568782	0.33333	4.2	0.6510	14.9537	14.455
93	7	50	7	0.584337	0.316550	4.09036	0.654827	0.14286	4.5	0.6975	10.0147	6.517
94	8	50	6	0.606513	0.347331	4.85210	0.759734	0.25000	5.4	0.6370	11.2919	10.170
95	9	50	6	0.486133	0.215046	4.37519	0.780185	-0.44444	4.2	0.6510	-4.0043	-17.405
96	10	50	5	0.568801	0.306599	5.68801	0.928464	0.00000	6.0	0.9300	5.4850	0.165

Appendix 2 (continued)

VARIABLE=RE1 = Residual of the expected value

UNIVARIATE

		MOMENTS				QUANTILES (DEP=4)				EXTREMES	
N	96	SUM	96	100% MAX	14.9537	99%	14.9537	LOREST	10.0147	HIGHEST	10.0147
MEAN	-1.17585	SUM	-112.882	75% Q3	3.36597	95%	10.1061	-10.7568	10.6238	10.6238	
STD DEV	6.01188	VARIANCE	36.1427	50% MED	-1.73718	90%	7.32597	-10.5517	11.2919	11.2919	
SKEWNESS	0.533001	KURTOSIS	-0.270248	25% Q1	-6.26936	10%	-8.65864	-9.83722	14.3908	14.3908	
USS	3566.29	CSS	3433.56	0% MIN	-10.7568	5%	-9.35908	-9.78827	18.9537	18.9537	
CV	-511.28	STD MEAN	0.613585	RANGE	25.7105	1%	-10.7568	-9.28334			
T:MEAN=0	-1.91636	PROB>T	0.0583254	Q3-Q1	9.63533						
D:NORMAL	0.102988	PROB>D	0.013	MODE	-10.7568						



Appendix 2 (continued)

VARIABLE=BET

UNIVARIATE

FREQUENCY TABLE

VALUE	COUNT	PERCENTS CELL	CUM	VALUE	COUNT	PERCENTS CELL	CUM	VALUE	COUNT	PERCENTS CELL	CUM	VALUE	COUNT	PERCENTS CELL	CUM	VALUE	COUNT	PERCENTS CELL	CUM
-10.7568	1	1.0	1.0	-5.18543	1	1.0	26.0	-1.70953	1	1.0	51.0	3.5667	1	1.0	76.0				
-10.5517	1	1.0	2.1	-5.02268	1	1.0	27.1	-1.62872	1	1.0	52.1	4.03975	1	1.0	77.1				
-9.83722	1	1.0	3.1	-4.83477	1	1.0	28.1	-1.60902	1	1.0	53.1	4.22801	1	1.0	78.1				
-9.78827	1	1.0	4.2	-4.75406	1	1.0	29.2	-1.60518	1	1.0	54.2	4.26837	1	1.0	79.2				
-9.28334	1	1.0	5.2	-4.28966	1	1.0	30.2	-1.3568	1	1.0	55.2	4.64592	1	1.0	80.2				
-9.10238	1	1.0	6.3	-4.22006	1	1.0	31.3	-1.32808	1	1.0	56.3	4.80576	1	1.0	81.3				
-8.8936	1	1.0	7.3	-4.05422	1	1.0	32.3	-1.27146	1	1.0	57.3	4.85218	1	1.0	82.3				
-8.84292	1	1.0	8.3	-4.00426	1	1.0	33.3	-1.206364	1	1.0	58.3	5.02257	1	1.0	83.3				
-8.78669	1	1.0	9.4	-3.87548	1	1.0	34.4	-1.14143	1	1.0	59.4	5.1043	1	1.0	84.3				
-8.60376	1	1.0	10.4	-3.80268	1	1.0	35.4	-1.07695	1	1.0	60.4	5.23079	1	1.0	85.4				
-8.59109	1	1.0	11.5	-3.7999	1	1.0	36.5	-1.01242	1	1.0	61.5	5.44439	1	1.0	86.5				
-8.48497	1	1.0	12.5	-3.40882	1	1.0	37.5	-1.00792	1	1.0	62.5	5.61271	1	1.0	87.5				
-8.4031	1	1.0	13.5	-3.359	1	1.0	38.5	-1.00345	1	1.0	63.5	5.79054	1	1.0	88.5				
-8.19519	1	1.0	14.6	-3.16156	1	1.0	39.6	-1.0016	1	1.0	64.6	6.59958	1	1.0	89.6				
-8.04557	1	1.0	15.6	-3.01732	1	1.0	40.6	-1.00171	1	1.0	65.6	7.43302	1	1.0	90.6				
-8.02295	1	1.0	16.7	-3.00362	1	1.0	41.7	-1.00171	1	1.0	66.7	7.86335	1	1.0	91.7				
-7.87958	1	1.0	17.7	-2.89208	1	1.0	42.7	0.444383	1	1.0	67.7	9.21626	1	1.0	92.7				
-7.67076	1	1.0	18.8	-2.78798	1	1.0	43.8	1.88732	1	1.0	68.8	9.88745	1	1.0	93.8				
-7.49117	1	1.0	19.8	-2.63391	1	1.0	44.8	1.9284	1	1.0	69.8	10.0147	1	1.0	94.8				
-7.37754	1	1.0	20.8	-2.38897	1	1.0	45.8	1.94529	1	1.0	70.8	10.6238	1	1.0	95.8				
-7.29956	1	1.0	21.9	-2.33057	1	1.0	46.9	2.08786	1	1.0	71.9	11.2919	1	1.0	96.9				
-7.14274	1	1.0	22.9	-2.24323	1	1.0	47.9	2.40176	1	1.0	72.9	14.3908	1	1.0	97.9				
-6.95084	1	1.0	24.0	-1.76744	1	1.0	49.0	2.75778	1	1.0	74.0	14.9537	1	1.0	99.0				
-6.63067	1	1.0	25.0	-1.76483	1	1.0	50.0				75.0				100.0				

Appendix 2 (continued)

STATISTICAL ANALYSIS SYSTEM

UNIVARIATE

VARIABLE=REZ

FREQUENCY TABLE

VALUE	COUNT	PERCENTS	VALUE	COUNT	PERCENTS	VALUE	COUNT	PERCENTS	VALUE	COUNT	PERCENTS
-19.9926	1	1.0	-7.25461	1	1.0	-2.88666	1	1.0	2.62739	1	1.0
-19.2043	1	1.0	-6.87219	1	1.0	-2.72952	1	1.0	3.13529	1	1.0
-18.0488	1	1.0	-6.70144	1	1.0	-2.31935	1	1.0	3.17608	1	1.0
-17.4052	1	1.0	-6.65365	1	1.0	-1.78975	1	1.0	3.86936	1	1.0
-15.8697	1	1.0	-6.61211	1	1.0	-1.65626	1	1.0	4.09771	1	1.0
-15.5219	1	1.0	-6.60712	1	1.0	-1.29773	1	1.0	4.78157	1	1.0
-14.8448	1	1.0	-6.56562	1	1.0	-.913761	1	1.0	5.20287	1	1.0
-14.7202	1	1.0	-6.55611	1	1.0	-.474235	1	1.0	5.79433	1	1.0
-14.3321	1	1.0	-6.46377	1	1.0	-.404156	1	1.0	6.5166	1	1.0
-13.1033	1	1.0	-6.44442	1	1.0	-.186069	1	1.0	8.1839	1	1.0
-12.7239	1	1.0	-6.32731	1	1.0	0.127647	1	1.0	8.5301	1	1.0
-12.3167	1	1.0	-5.52123	1	1.0	0.144559	1	1.0	8.82129	1	1.0
-10.4674	1	1.0	-5.51467	1	1.0	0.165381	1	1.0	9.09739	1	1.0
-10.2702	1	1.0	-5.39546	1	1.0	0.165434	1	1.0	9.5791	1	1.0
-9.79147	1	1.0	-4.29663	1	1.0	0.176394	1	1.0	9.84027	1	1.0
-8.70795	1	1.0	-4.23144	1	1.0	0.224465	1	1.0	10.1701	1	1.0
-8.68855	1	1.0	-4.1701	1	1.0	0.378221	1	1.0	11.976	1	1.0
-8.30616	1	1.0	-4.14871	1	1.0	0.977286	1	1.0	12.0976	1	1.0
-7.84463	1	1.0	-4.06433	1	1.0	0.937553	1	1.0	12.6372	1	1.0
-7.68908	1	1.0	-3.90941	1	1.0	1.27356	1	1.0	14.455	1	1.0
-7.64889	1	1.0	-3.65901	1	1.0	1.4168	1	1.0	15.5702	1	1.0
-7.56153	1	1.0	-3.51492	1	1.0	1.67935	1	1.0	16.8235	1	1.0
-7.27839	1	1.0	-3.49768	1	1.0	1.75311	1	1.0	19.6819	1	1.0
			-3.01053	1	1.0	2.26348	1	1.0	32.0357	1	1.0

Labels of the variable names in Appendix 2.

M =	}	Same as labels in <u>Table 3.1</u> (Page 36)
N =		
R =		
I1=		
I2=		
EEA =		
EVA =		
Decimal =		

PEA = predicted value of the expected value of A. Using equation 18:

$$PEA = M(0.6 + 0.3 \times \text{Decimal}).$$

PVA = predicted value of the variance of A using PEA and equation 15:

$$PVA = 0.155 \text{ PEA}.$$

$$RE1 = (PEA - EEA) \times 100 / EEA.$$

$$RE2 = (PVA - EVA) \times 100 / EVA.$$

REFERENCES

- Abe, O. (1973) "A note on the methodology of Knox's tests of 'Time and space interaction'", *Biometrics*, 29, 67-77.
- Bailar, J. C., Eisenberg, H. and Mantel, N. (1970) "Time between pairs of leukemia cases", *Cancer*, 25(6), 1301-1303.
- Barton, D. E., David, F. N. and Merrington, M. (1965) "A criterion for testing contagion in time and space", *Annals of Human Genetics*, 29, 97-102.
- Barton, D. E., David, F. N., Fix, E. and Merrington, M. (1967) "Tests for space-time interaction and power functions", Proceedings of the Fifth Berkeley Symposium. Ed. Lucien M. LeCam and Jerzy Neyman, Vol IV, 217-227, University of California Press, Berkeley.
- Bennett, B. M. and Nakamura, E. (1968) "Percentage points of the range from a symmetric multinomial distribution", *Biometrika*, 55, 377-379.
- Besag, J. and Diggle, P. J. (1977) "Simple Monte-Carlo tests for spatial pattern", *Applied Statistics*, 26, 327-333.
- Cox, D. R. and Hinkley, D. V. (1975) Theoretical Statistics, Chapman and Hall, London.
- David, F.N. and Barton, D. E. (1966) "Two space-time interaction tests for epidemicity", *British Journal of Preventive and Social Medicine*, 20, 44-48.
- David, H. A. (1970) Order Statistics, John Wiley & Sons, Inc., New York.
- Darwin, J. H. (1957) "The difference between consecutive members of a series of random variables arranged in order of size", *Biometrika*, 44, 211-218.
- Ederer, F., Myers, M. H. and Mantel, N. (1964) "A statistical problem in space and time: Do leukemia cases come in clusters?" *Biometrics*, 20, 626-638.
- Feller, W. (1968) An Introduction to Probability Theory and Its Applications, John Wiley & Sons, Inc., New York.
- Godwin, H. J. (1949) "Some low moments of order statistics", *Annals of Mathematical Statistics*, 20, 279-285.
- Green, P. E. (1978) Analyzing Multivariate Data, Dryden Press, Hinsdale, Illinois.

- Greenwood, R. E. and Glasgow, M. O. (1950) "Distribution of maximum and minimum frequencies in a sample drawn from a multinomial distribution", *Annals of Mathematical Statistics*, 21, 416-424.
- Grimson, R. C. (1979) "The clustering of diseases", *Mathematical Biosciences*, 46, 257-278.
- Gumbel, E. J. (1958) Statistics of Extremes, Columbia University Press, New York.
- Gumbel, E. J. and Herbach, J. H. (1951) "Exact distribution of the extremal quotient", *Annals of Mathematical Statistics*, 22, 418-426.
- Gumbel, E. J. and Keeney, R. D. (1950) "The extremal quotient", *Annals of Mathematical Statistics*, 21, 523-538.
- Gumbel, E. J. and Pickands III, J. (1967) "Probability tables for the extremal quotient", *Annals of Mathematical Statistics*, 38, 1541-1551.
- Harter, H. L. (1969) Order Statistics and Their Use in Testing and Estimation, Vol. 1, Aerospace Research Laboratories, USAF.
- Harter, H. L. and Owen, D. B. (1975) Editors: Selected Tables in Mathematical Statistics, Vol. 4, Institute of Mathematical Statistics, Dirichlet Distribution, Type 1, Tables by Milton Sobel, V. R. R. Uppuluri and K. Frankowski.
- International Mathematical and Statistical Libraries, Inc., (IMSL) (1979) Reference Manual, Edition 7, Vol. 2, "MDEETA (MDBA) beta probability distribution function".
- Johnson, N. L. (1960) "An approximation to the multinomial distribution; some properties and applications", *Biometrika*, 47, 93-102.
- Johnson, N. L. and Kotz, S. (1977) Urn Models and Their Application, John Wiley & Sons, Inc., New York.
- Johnson, N. L. and Young, D. H. (1960) "Some applications of two approximations to the multinomial distribution", *Biometrika*, 47, 463-469.
- Klauber, M. R. (1971) "Two-sample randomization tests for space-time clustering", *Biometrics*, 27, 129-142.
- Knox, G. (1963) "Detection of low intensity epidemics, application to cleft lip and palate", *British Journal of Preventive and Social Medicine*, 17, 121-127.
- Knox, G. (1964a) "The detection of space-time interaction", *Applied Statistics*, 13, 25-30.
- Knox, G. (1964b) "Epidemiology of childhood leukemia in Northumberland and Durham", *British Journal of Preventive and Social Medicine*, 18, 17-24.

- Lindley, D. V. (1965) Introduction to Probability and Statistics from Bayesian Viewpoint, Part 2, Inference, Cambridge University Press, London.
- Mantel, N. (1967) "The detection of disease clustering and a generalized regression approach", *Cancer Research*, 27, 209-220.
- Mantel, N., Kryscio, R. J. and Myers, M. H. (1976) "Tables and formulas for extended use of the Ederer-Myers-Mantel disease-clustering procedure", *American Journal of Epidemiology*, 104(5), 576-584.
- Naus, J. I. (1965) "The distribution of the size of the maximum cluster of points on a line", *Journal of the American Statistical Association*, 60, 532-538.
- Naus, J. I. (1966a) "A power comparison of two tests for non-random clustering", *Technometrics*, 8, 493-517.
- Naus, J. I. (1966b) "Some probabilities, expectations and variances for the size of the smallest intervals and largest clusters", *Journal of the American Statistical Association*, 61, 1191-1199.
- Naus, J. I. (1982) "Approximations for distributions of scan statistics", *Journal of American Statistical Association*, 77, 177-183.
- North Carolina Vital Statistics (1980) *Leading Causes of Mortality*, 2, North Carolina Center for Health Statistics, Department of Human Resources, Public Health Statistics Branch.
- Pearson, E. S. and Hartley, H. O. (1962) Biometrika Tables for Statisticians, Vol. 1, Cambridge University Press, London.
- Pike, M. C. and Smith, P. G. (1968) "Disease clustering: a generalization of Knox's approach to the detection of space-time interaction", *Biometrics*, 24, 541-546.
- Pinkel, D. and Nefzger, K. (1959) "Some epidemiological features of childhood leukemia in Buffalo, NY, area", *Cancer*, 12, 351-358.
- Pyke, R. (1965) "Spacing", *Journal of the Royal Statistical Society*, B, 27, 395-436. Discussion, 437-449.
- Riordan, J. (1958) An Introduction to Combinatorial Analysis, John Wiley & Sons, Inc., New York.
- Roberson, P. K. (1979) "Distributional and robustness problems in time-space disease clustering", Ph.D. dissertation, University of Washington, Seattle.
- Schneider, K. (1982) "Agent white", *Inquiry*, March 15, 1982, 14-18.
- Sobel, M. and Uppuluri, V. R. R. (1974) "Sparse and crowded cells and dirichlet distribution", *Annals of Statistics*, 2, 977-987.

- Stark, C. R. and Mantel, N. (1967a) "Lack of seasonal or temporal-spatial clustering of Down's syndrome births in Michigan", *American Journal of Epidemiology*, 86, 199-213.
- Stark, C. R. and Mantel, N. (1967b) "Temporal-spatial distribution of birth dates for Michigan children with leukemia", *Cancer Research*, 27, 1749-1755.
- Symons, M. J. (1973) "Bayes modification of some clustering criteria", Institute of Statistics Mimeo Series No. 880, Department of Biostatistics, University of North Carolina at Chapel Hill.
- Symons, M. J., Grimson, R. C. and Yuan, Y. C. (1982) "Clustering of rare events", in press.
- Wallenstein, S. (1980) "A test for detection of clustering over time", *American Journal of Epidemiology*, 111(3), 367-372.
- Wallenstein, S. and Naus, J. I. (1974) "Probabilities for the size of largest clusters and smallest intervals", *Journal of the American Statistical Association*, 69, 690-697.
- Young, D. H. (1962) "Two alternatives to the standard chi-square test of the hypothesis of equal cell frequencies", *Biometrika*, 49, 107-116.