A TEST FOR CONCORDANT NONRANDOM PATTERNS AMONG SERIES WITH EPIDEMIOLOGIC APPLICATIONS

by

Deborah D. Ingram

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1430

February 1983

A TEST FOR CONCORDANT NONRANDOM PATTERNS AMONG SERIES WITH EPIDEMIOLOGIC APPLICATIONS

by

Deborah D. Ingram

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

Chapel Hill
1983

Approved by:

Advisor

Reader

Reader

DEBORAH D. INGRAM. A Test for Concordant Nonrandom Patterns Among Series with Epidemiologic Applications. (Under the direction of ROGER C. GRIMSON)

A number of tests are available in the literature for use in detecting clusters of disease in space, in time, or in space and time. These tests are reviewed briefly. A new disease clustering test, the test for concordant, nonrandom patterns between two series is proposed. The test is sensitive simultaneously to temporal clustering in series of incidence data from different locations and to concordance between the series.

IThe exact first five moments of the test statistic are derived. The first three moments are identical to those of a binomial variable, the fourth and fifth moments are not. An expression for the exact p-value is derived. Since it is infeasible to evaluate this expression, even for small sample sizes, an appropriate approximate test of significance is sought. Comparison of the normal approximation, Gram-Charlier and Edgeworth series approximations, and Pearson curve approximation with simulated distributions of the test statistic indicated that the Pearson curves approximated the upper tail of the distribution most closely.

Two epidemiologic examples are presented to illustrate in which situations the test is useful and how the test works.

Two possible extensions of the test to r series are presented along with some distributional properties. In addition, some initial attempts to modify the test so that it is sensitive to time order are presented.

ACKNOWLEDGEMENTS

I wish to thank my advisor, Dr. Roger C. Grimson, for his guidance and encouragemnt during the course of this dissertation reearch. Thanks is also expressed to the other members of my committee, Drs. P. K. Sen, Dana Quade, Michael D. Hogan, and Gerardo Heiss for their suggestions and ideas. Discussions with Dr. Norman L. Johnson, during the early days of the research, were very helpful.

I also wish to thank my friends and family for their encouragement and moral support throughout the years of my graduate studies.

The financial assistance of the Department of Biostatistics during the course of this graduate program is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

Table | Page

Table | Page

LIST OF FIGURES

Chapter I

INTRODUCTION AND REVIEW OF THE LITERATURE

1.0 Introduction

During the past twenty years a number of statistical tests have been developed for use in detecting clusters of events in space, in time, and in space and time simultaneously. Mantel (1967) and Klauber (1974) reviewed many of these tests. Events are said to cluster in space, or in time, if they occur more closely together geographically, or temporally, than would be expected if they were randomly distributed. Events are said to cluster in space and time if those that occur more closely together in space also occur more closely together in time than would be expected. Knox (1964a) referred to space-time clustering as space-time interaction. The clustering tests reviewed here differ from classical clustering measures, such as factor analysis, because they attempt to determine whether or not clustering occurs, not to define clusters.

Generally, these clustering tests have been applied by epidemiologists to the study of diseases in defined geographic areas and time periods. The tests often have been used for detecting clustering among cases of rare diseases such as cancer, birth defects and malformations, and diseases with long latent periods. Epidemiologists have been interested in detecting clustering because its presence, while not proving anything about the etiology of a disease, may provide support or clues for

hypotheses about the etiology. Researchers feel that the presence of clustering may indicate that a disease is contagious or that its etiology involves some environmental factor. For example, since cat leukemia is caused by a virus, scientists speculate that childhood leukemia also may be caused by a virus and hence be contagious. Numerous studies have applied clustering tests to leukemia data in the hopes of detecting significant clustering and thus providing support for the viral hypothesis. Detection of disease clusters also is useful in identifying areas requiring monitoring or intervention.

Clustering tests have not been used only to detect disease clustering. Some examples of other uses include: Klauber (1971) has studied the tendency of earthquakes to occur more often following nuclear explosions; Mantel (1967) suggested using his test to study interpersonal relationships; and Barton and David (1962) have studied chromosome patterns for various types of clustering.

The tests reviewed here do not require knowledge of the distribution of the population at risk or of the disease rate. Generally, all that is required is knowledge of the times and locations of occurrence of the cases of disease and some assumptions about the distribution of the population in space and/or in time. Some of the tests assume that the population at risk is uniformly distributed over the geographic area being studied and some assume that population shifts over the time period being studied are minimal. All of the tests assume a constant reporting rate. The null hypothesis of most of the time(space) clustering tests is that the cases are distributed randomly among the time (space) units. Likewise, the null hypothesis of most of the space-time clustering tests is that the n observed times are matched randomly to the n observed locations.

For most of the tests it is infeasible to calculate the exact p-value of the test statistic involved and assumptions have been made about its asymptotic distribution so that an approximate p-value can be calculated. For example, some investigators have assumed that their test statistics are asymptotically distributed like Poisson variables, others have assumed that their test statistics are asymptotically distributed like chi-squared variables, and many have assumed that their test statistics are asymptotically normally distributed. For some of the tests the appropriateness of the assumed asymptotic distribution has been examined empirically or theoretically, but for most it has not.

It is important to realize that because of the different types of nonrandomness, the different situations to which tests are applied, and the different questions that are of interest, there is no omnibus clustering test, nor can there be one. Departures from randomness can take several forms, namely, irregular or sudden clusters, trends, oscillations, and too regular spacing. The different types of nonrandomness generally reflect different underlying processes. Thus, tests tend to be appropriate for one type of departure and not for the others. The tests reviewed here test for irregular clustering, specifically, excessive numbers of cases. A few of them also can be used to test for oscillations or seasonality.

Tests often are appropriate only for specific categories of disease. For example, tests that are appropriate for diseases with short latent periods may not be appropriate for diseases with long latent periods.

Another factor that must be taken into account is that departures from randomness can occur in the spatial dimension, in the temporal dimension, in both the spatial and temporal dimensions separately, or in the spatial and temporal dimensions simultaneously. Preferably, a

clustering test is sensitive to only one of these possibilities, which one depending on the hypothesis of interest. Most of the space-time tests reviewed here are sensitive to nonrandomness only when it occurs in both dimensions simultaneously.

Another complicating factor in the study of disease clustering is that diseases occur in space and time continua and thus there may be optimal units of both for displaying a pattern; that is, significant clustering may be found when one scale is used and not when another scale is used. The role of scale in clustering tests is complex because scale enters the picture at three different levels. First, scale is involved in the definition of the geographic area and/or the time period being studied since the size of the area and the length of the time period must be specified. Clearly, if the geographic area is too small, clustering may not be detected. One example of this effect involves the cardiovascular disease belt that runs along the east coast of the United States. When looking at county data, the belt is clearly defined. However, when looking at data within counties in the belt, significant clustering is not found. Conversely, if the area is too large, clusters may be missed, especially when data are pooled. For example, when looking at town data within a county, significant clustering may be detected, while when looking at county data for the whole state no clustering may be found. If the time period is too short, nonrandom patterns, such as trends or seasonal fluctations may not be detected, whereas if it is too large such underlying assumptions as minimal population shift effects may be violated. Once the spatial and temporal boundaries have been decided upon, scale becomes involved again in the choice of the unit of observation, i.e. individuals, neighborhoods,

towns, and so forth. Finally, scale may become involved a third time because many of the clustering tests require either specification of "critical" distances or times or division of the geographic area or time period into subareas or subintervals, respectively.

Development of good clustering tests has been difficult because there are so many factors that must be considered. To date, there are no completely satisfactory tests; all of them have their disadvantages and shortcomings. In addition, little is known about the sensitivity and power of the tests, about the appropriateness of the approximations often used to obtain p-values, or about the situations for which a test works well and those for which it does not.

1.1 Literature review: Description of specific methods

1.1.2 Pinkel and Nefzger cell occupancy approach

Pinkel and Nefzger (1959) after examining data from a study of childhood leukemia adapted a classical combinatorial result to test for space-time clustering. They proposed an exact test for determining the significance of observing, out of $r(r-1)/2$ possible pairs of cases, k close pairs. A pair is defined to be close if the cases occur within a specified distance and length of time of each other. According to a cell occupancy result, if r cases are randomly allocated to n cells, the probability of placing k cases in cells each of which contains at least one other case is

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

where $P(k)$ is conditional on n and r. Thus, to obtain the significance

of an observed k_0, we calculate

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

The authors determined n by dividing the study area and time period into 3-dimensional time-space cells based on the time and space distances used to define closeness. The test has been criticized since it is based on the assumption that the population at risk is uniformly distributed both in space and in time, an assumption that clearly does not hold in urban areas. Also, Mantel (1967) has pointed out that the test is sensitive not only to space-time clustering, but also to spatial clustering alone and to temporal clustering alone, a property which is not desirable.

There are two other problems with this test, neither of which has been mentioned in the literature. First, k represents the number of cases in cells containing at least one other case. However, k_0 was not calculated by determining which of the cells each case was in and then counting the number of pairs within the n cells. Secondly, the probability formula only counts the number of different configurations of cases in the cells, and not the different ways of arranging the individual cases, given a certain configuration. Since Pinkel and Nefzger look at all possible pairs of cases when calculating k, this should be done when evaluating the distribution. To clarify, given three cases and two time-space cells, with two cases occurring in the first cell and one in the second, the probability formula would count this as one arrangement, namely (X X ,X), and not as three arrangements, namely (X_1 X_2; X_3), (X_1 X_3; X_2), and (X_2 X_3: X_1). Thus, the exact probability formula given by Pinkel and Nefzger is not appropriate given their definition of the test statistic.

1.1.2. Pinkel, Dowd, and Bross ridit approach

Pinkel, Dowd, and Bross (1963) employed a ridit approach to study space-time clustering in Pinkel and Nefzger's leukemia data and in solid tumor and traffic accident data. They computed the distances and lengths of time between all 4465 possible pairs from among 95 leukemia cases and classified them into a space x time contingency table. They divided the spatial dimension into nine categories, eight for the distances 0 to 1 mile, in 1/8-th mile steps, and one category for all distances greater than one mile. The time dimension is divided into one year categories. Pinkel et. al. treated the 4070 pairs falling into the >1 mile category as a reference distribution against which to compare the temporal distribution of the eight other spatial categories, under the assumption that pairs of cases occurring more than one mile apart were not related. As they demonstrated, however, any suitable group of controls can be used as the reference distribution. The average ridit of the reference distribution was calculated so that it equaled .5. The average ridit of each of the eight temporal distributions was calculated by referring back to the temporal distribution of the reference population. If the average ridit is significantly less than .5 for a category, then spatial-temporal clustering exists in that category. The authors did not suggest a test of significance but calculated confidence intervals for the average ridit of each spatial category.

This test has been criticized because Pinkel et al. did not take into account the dependence of the $n(n-1)/2$ pairs of cases.

1.1.3 Knox's contingency table tests

Knox (1963) suggested a space-time interaction test that involved dividing the spatial dimension into n categories and the time period into

m categories. The lengths of time and distances between all n(n-1)/2 possible pairs of cases are computed and the pairs assigned to the appropriate space-time cell in the nxm contingency table. Knox used a χ^2 test with (n-1)(m-1) degrees of freedom to analyze the table, even though he pointed out that the dependence of the pairs might render the test inappropriate. Abe(1963) derived the correct test of significance.

Knox (1964a,1964b) examined the idea of space-time interaction and modified his earlier test. For this second test, now known as Knox's test, he dichotomized the spatial and temporal dimensions. Using a critical length of time and a critical distance, Knox classified each of the n(n-1)/2 pairs of cases as close in time or not and as close in space or not, forming a 2x2 contingency table. The test statistic, X, is the number of pairs close both in time and in space. Because the pairs are dependent, assessment of the statistical significance of an observed X is difficult. Knox conjectured that if X were small compared to the total number of pairs, i.e. a rare event, then its null distribution would be approximately Poisson. Therefore, he treated X as a Poisson variable calculating its expectation from the marginal totals in the usual way.

Barton and David (1966) confirmed, for some situations, the appropriateness of the Poisson approximation for the null distribution of X. They conceptualized X as the intersection of a space graph in which pairs close in space are joined and a time graph in which pairs close in time are joined. In the intersection of the two graphs, pairs close in both space and time are connected. Using this graph theoretic approach, Barton and David derived a general expression for the rth factorial moment of X and explicit expressions for the mean and variance of X. The expression they derived for the mean was identical to that reported

by Knox. The variance expression contained several terms, the first term being the mean. In situations where the mean and the coefficient of variation are moderately large, the effect of the additional terms on the variance is small and X will be approximately Poisson. Barton and David concluded that, in general, use of the Poisson approximation could yield misleading results. Mantel(1967) has outlined the methodology for obtaining the exact permutational distribution of X.

Using Knox's data, Barton and David found the mean and variance to be .833 and .802, respectively, which are adequately close to .833, the value obtained using the Poisson approximation. Barton and David cited Pike's work which also confirmed the appropriateness of the Poisson approximation for Knox's data. Pike approximated the null distribution of X by randomly allocating the 96 space coordinates to the 96 time measures 2000 times. The frequency distribution he obtained was remarkably close to that obtained for a Poisson variable with a mean of .833.

Clearly, both of these tests developed by Knox involve subjectivity in defining the boundaries of the time and space categories, i.e., the critical distances. Both also involve a loss of information, and hence of power, since the actual time and space distances are not used. Although the dichotomization in the second test means a greater loss of information, it also means that the close distances are accentuated while the large distances are compressed and thus less likely to mask clustering.

1.1.4 Pike and Smith's extension of Knox's test

Pike and Smith (1968) extended Knox's test to diseases with long latent periods using Barton and David's graph theoretic approach. For each case they defined a period of time and an area of infectivity and a

period of time and an area of susceptibility. Clearly, some arbitrariness is involved in defining the time periods and areas of infectivity and susceptibility. Pairs of cases are considered close in time if their periods of infectivity and susceptibility overlap and close in space if their areas of infectivity and susceptibility overlap. The test statistic, X, is the number of pairs close in both time and space.

Pike and Smith suggested a randomization procedure for determining the exact null distribution of X and hence an exact test of significance. The exact null distribution can be evaluated by calculating X for all possible allocations of the n space coordinates to the n time coordinates. An exact p-value is obtained by comparing an observed X to the distribution. Since in most cases enumeration of this distribution is impractical, Pike and Smith suggested either using Monte Carlo simulation to approximate the null distribution or assuming asymptotic normality and using the normal deviate $[X-E(X)]/V(X)^{1/2}$. $E(X)$ and $V(X)$ in the normal approximation are the randomization expectation and variance of X, both of which Pike and Smith derived.

1.1.5 Barton and David's points-on-a-line approach

Barton, David, and Merrington (1965) and David and Barton (1966) adapted a test originally used to study the randomness of points on a line (Barton and David, 1962b) for use in the detection of space-time interaction. The test, analogous to analysis of variance techniques, involves a ratio of within group variance to overall variance. This is accomplished by determining a critical length of time according to some predetermined criteria or on the basis of knowledge of the etiology of the disease being studied. Pairs of successive cases separated in time by less than the critical distance are formed into time clusters. For

example, a cluster of three is formed if the first and second cases are close and the second and third cases are close. The test statistic is

$$Q = \left(\frac{n-1}{n-h}\right)\left[1 - \sum_{t=1}^{h} \frac{r_t((\bar{X}_t - \bar{X})^2 + (\bar{Y}_t - \bar{Y})^2)}{n(\mathrm{Var}X + \mathrm{Var}Y)}\right],$$

where n is the number of cases, h is the number of time clusters, and r_t is the number of cases in the t-th cluster, $\bar{X}_t$ and $\bar{Y}_t$ are the means of the distance coordinates for the t-th cluster, and $\bar{X}$ and $\bar{Y}$ are the overall means of the distance coordinates. Thus, Q is the ratio of the average squared distances within clusters to the overall average squared distance between pairs of cases.

Under the null hypothesis of no space-time interaction, Q has an expected value of 1. When clustering is present, Q is smaller than 1. To assess significance, David and Barton suggested using a randomization test to determine the exact distribution of Q. Since calculation of the exact distribution often is not feasible, Barton and David (1962b) suggested using a β approximation when n is small and a normal approximation when n is large. When the number of clusters is large, Q is approximately normally distributed, i.e. $(Q-1)/\sigma \sim N(0,1)$, where σ^2 is the variance of Q. When the number of clusters is small, they suggested the rule of thumb, $h-1<3\sigma n$, a better approximation can be obtained with an F approximation,

$$\frac{(n-1) - (n-h)Q}{(h-1)Q} \sim F((h-1)\Theta,(n-h)\Theta),$$

where

$$\Theta = \frac{2}{(n-1)}\left[\frac{(n-1)}{(n-h)\sigma^2} - 1\right].$$

David and Barton applied this test to Knox's data. Unlike Knox, they found no evidence of clustering.

As pointed out earlier, one of the drawbacks of Knox's test is the arbitrariness involved in choosing the temporal and spatial critical points. One advantage of the test proposed by Barton and David is that the actual distances are used and the only arbitrariness is in selection of the critical time point. One disadvantage of Barton and David's test is that the small distances, which are of most interest, have less influence on the statistic than do the large distances. In fact, the large distances may so dominate the statistic that they mask any clustering.

1.1.6 Mantel's generalized regression approach

Mantel (1967) developed a "generalized regression" approach to the detection of disease clustering in space and time. One advantage of Mantel's approach is that the actual time and space distances are used, thus avoiding the use of arbitrary cutpoints and the loss of information. Mantel proposed using reciprocal transformations of the distances in order to spread out the close distances and compress the long distances.

The test statistic is

$$Z = \sum_{i=1}^{n} \sum_{j=1}^{n} g(X_{ij} + a)h(Y_{ij} + b) ,$$

where X_{ij} is the length of time and Y_{ij} is the spatial distance between the ith and jth observations, $X_{ii}=Y_{jj}=0$, g and h are some functions such as the reciprocal transformation, and a and b are arbitrary constants. Note that Knox's test is a special case of Mantel's test in which a=b=0 and $g(X_{ij})=I(X_{ij}<t_0)$ and $h(Y_{ij})=I(Y_{ij}<s_0)$, where I is the identity function

and t_0 and s_0 are the critical distances in time and space, respectively.

Mantel suggested that his test be used to detect disease clustering or even to study interpersonal relationships. Myers, who studied 70 cases of childhood leukemia using both no transform and the reciprocal transform found significant results only after he augmented the data by duplicating each case and then only for the transformed data. Mantel concluded that his test has low power when no transformation is used. Siemiatycki (1971) also came to this conclusion after analyzing some typhoid and salmonellosis data.

A constant must be added to the distances before making the reciprocal transformation because of the possibility of very small or zero time and/or space distances. Unfortunately, choice of these constants influences the value of Z and hence the outcome of the test of significance if the normal approximation is used. Mantel suggested that, for best results, the constants be close to the expected distances between close pairs. He admitted that the researcher might have to resort to trial and error to select optimal constants; clearly, such a practice will affect the validity of the test. Glass (1971) in a study of childhood leukemia and Siemiatycki (1971) in a study of Burkitt's lymphoma found that as the size of the constants increased, Z tended to decrease. However, Siemiatycki reported that the significance levels obtained by fitting Pearson curves to the data using the first four moments of Z are affected only slightly by changes in the constants.

The effects of five different transformations, namely, no tranform, the reciprocal transform, the reciprocal square transform, a dichotomizing transform, and a truncating transform, on Z and the corresponding normal deviate, and on the significance levels obtained by

fitting Pearson curves has been examined by Siemiatycki (1971) and Siemiatycki and McDonald (1972). They found that the values of Z and the corresponding normal deviates differed but that the Pearson percentage points were almost identical.

Mantel suggested the three usual tests of significance, namely, 1) obtain the exact randomization distribution of Z, 2) use Monte Carlo simulation to obtain an approximation to the distribution of Z, or 3) assume asymptotic normality and calculate the normal deviate utilizing the expressions for the expectation and variance of Z that Mantel derived.

Mantel asserted that while no general statement about asymptotic normality could be made, Z had the classic appearance of a U-statistic and therefore the normality approximation seemed reasonable. Unfortunately, a number of simulation studies have found the distribution of Z to be highly skewed and have shown that while the use of the normal approximation is appropriate when Z is highly significant or nonsignificant, its use is inappropriate when Z has borderline significance. Klauber (1971) and Siemiatycki (1971) found that even sample sizes of several hundred were not large enough for asymptotic normality to be used with assurance for evaluating the significance of borderline cases. Their findings do not indicate that asymptotic normality does not hold, simply that it does not hold in the situations examined and for the sample sizes studied.

1.1.7. Klauber's extension of Mantel's test

Klauber (1971) extended Mantel's "generalized regression" approach to the two-sample case. Given two sets of data, say A and B, it may be of interest to know whether there is a tendency for pairs, where one member of the pair is from A and the other is from B, to be close in

both time and space. Klauber's suggested statistic is:

$$Z = \sum\sum S_{ij}R_{ij}$$

where $S_{ij}=S((X_i - X_j^*)^2 + (Y_i - Y_j^*)^2)$ and $R_{ij} = R(|T_i - T_j^*|)$ are functions of space and time respectively, and (X_i, Y_i, T_i) and (X_j^*, Y_j^*, T_j^*) are the coordinates for the ith case in A and the jth case in B, respectively. An exact p-value is determined by comparing Z with its randomization distribution obtained by considering A random, that is, randomly allocating the n_A time coordinates to the n_A space coordinates, B random, or both A and B random depending on the hypothesis being tested. When enumeration of the randomization distribution is not feasible, Klauber advocated using the normal deviate to obtain an approximate p-value. He derived expressions for both the mean and variance of Z.

Klauber examined the appropriateness of the normal approximation using cat and dog leukemia data, nuclear explosion and earthquake data, and contrived data. For each of these data sets he used a range of sample sizes and additive constants, both the reciprocal transform and indicator functions that produced a Knox-like statistic, and all three randomization models. Klauber found that the empirical distribution of Z tended to be nonnormal. The distribution becomes more normal as sample sizes increased and as the sizes of the additive constants or critical distances increased. Unfortunately, Klauber also found that the test lost power as the constants increased in size.

Klauber (1975) extended his test to three or more sets of data, again employing Mantel's Z statistic approach. The q-sample statistic is formed by summing over all possible 2-sample Z's. There are 2^{q-1}

possible randomization models.

1.1.8 Pike and Smith's case-control approach

Pike and Smith (1974) developed a case-control approach to space-time clustering in which they used a test statistic based on the total number of pairs of cases having "effective" contact. A pair of cases has "effective" contact if both members are in the right place at the right time for one to transmit and the other to catch the disease. The concept of "effective" contact is not new here; Pike and Smith first introduced it in an earlier paper in which they extended Knox"s test. The test statistic proposed here, $Z=\sum\sum X_{ab}$, where $X_{ab}=1$ if the a-th and b-th case have effective contact and 0 otherwise and the summation is over all possible pairs, is the same as the one proposed in the earlier paper. The unique feature of this test is the use of matched controls either obtained by pair matching or by stratification, in determining the null distribution of Z.

Using matched case-control pairs, the null distribution of Z is obtained by calculating Z for each of the 2^n possible random samples of size n selected with the restriction that one member of each pair be included. Under the null hypothesis of no clustering, all of the samples are equally likely. The significance of an observed Z can be obtained by comparing it with the exact distribution.

Alternatively, if one does not wish to match one-to-one, a sample of m controls can be selected and the m controls and n cases divided into k strata on the basis of some criteria such as age or sex. The null distribution is obtained by calculating Z for all possible random samples of size n selected with the restriction that the number of subjects selected from each stratum be equal to the number of cases in that

stratum.

When n is too large to evaluate the exact null distribution of Z, Pike and Smith suggested using a Monte Carlo simulation to approximate it, or employing an approximate 1 degree of freedom χ^2 test. Pike and Smith derived expressions for E(Z) and V(Z).

Pike and Smith extended the case-control approach to three other totals, namely, the total number of cases transmitting the disease, the total number of cases receiving the disease, and the total number of cases involved in the transmission of the disease. The significance of each of these statistics can be evaluated by using any of the three methods described above. Pike and Smith derived the first two moments of each of the three statistics.

1.1.9 Ederer, Myers, and Mantel (EMM) approach

Ederer, Myers, and Mantel (1964) developed a test for temporal clustering based on the maximum number of cases observed in the time units within each subinterval of the time period. The time period under study is divided into disjoint subintervals of k time units each. For example, Ederer et al. divided 15 years of leukemia data into three subintervals each containing five one-year time units. Under the null hypothesis of no clustering, the n_t cases occurring in the t-th subinterval are distributed randomly among the k time units. Thus, this is a classical occupancy problem and the n_t cases have the multinomial distribution. The probability of a given arrangement of cases within a subinterval is

$$P(n_1, n_2, \ldots, n_k) = \frac{n_t!}{n_1!n_2!\ldots n_k!} \cdot \frac{k!}{k_0!k_1!\ldots k_{n_t}!} ,$$

where n_i, i=1,2,...k are the numbers of cases in the ith time unit, $\sum_{i=1}^{k} n_i = n_t$, and k_j is the number of time units with j cases, $\sum_{j=0}^{n_t} k_j = k$. The test statistic is based on the maximum number of cases, m_1, within each subinterval, i.e. $m_1 = \max(n_1, n_2, \ldots, n_k)$ and on n_t. The exact null distribution of m_1, conditional on n_t, and its expectation and variance have been calculated by Ederer et al. and the tables extended by Mantel, Krysoto, and Myers (1976). This test can be performed for one or more spatial locations.

Assuming normality and summing over all spatial locations and time subintervals, a one degree of freedom continuity corrected χ^2 is obtained as the test statistic,

$$\chi^2 = \frac{(|\sum m_1 - \sum E(m_1)| - .5)^2}{\sum Var(m_1)}$$

Ederer, Myers, and Mantel suggested two extensions of the test: 1) base the test statistic on the maximum number of cases in two consecutive time units, and 2) when the disease is not rare, focus on "vacuities", i.e. zero or minimum frequencies.

Stark and Mantel (1967a, 1967b) used the EMM method to explore the possibility of seasonal or temporal spatial clustering in childhood leukemia and in Down's syndrome births. They found no evidence of seasonality or of clustering for either.

1.1.10 Scan test

Naus (1965) proposed a test of temporal clustering that has come to be known as the scan test because the test statistic, the maximum number of cases, n, occurring in an interval of length t, is found by

"scanning" all intervals of length t. For example, Wallenstein (1980) in a study involving 24 months of trisomy data, set t=2 months and scanned the intervals 1-2 months, 2-3 months, ..., 23-24 months. Naus (1966a) calculated some probabilities, and the expectation and variance for the statistic for some values of n. Wallenstein expanded the probability tables to the lower tail of the distribution for certain values of n, N, and L where N is the total number of cases and L is the length of the time period under study divided by t.

Naus (1966b) compared the power of the scan test with that of the EMM test and concluded that if the scan interval, t, is small enough and the data are continuous over the interval, then the scan test is the more powerful of the two.

1.1.11 Bailar, Eisenberg, and Mantel's test of temporal clustering

Bailar, Eisenberg, and Mantel (1970) suggested a test of temporal clustering based on the number of pairs in a given area that occur within a specified length of time, d, of each other. The numbers of close pairs occurring in q areas are summed. The resulting test statistic is

$$t = \sum_{h=1}^{q} \sum_{i=1}^{N-d} n_{hi} n_{h,i+d}$$

where N is the number of years in the time period under study, n_{hi} is the number of cases occurring in the ith year and the hth area, and $n_{h,i+d}$ is the number of cases occurring in the i+d-th year and the hth area. Bailar et al. claim that t is approximately normally distributed; however, the appropriateness of the approximation has not yet been investigated.

1.1.12 LLoyd and Roberts' test

Lloyd and Roberts (1973) outlined a test for either spatial or temporal clustering which as Pike and Smith (1974) noted can be viewed as a special case of Knox's test. Lloyd and Roberts suggested using the number of pairs among all possible pairs of cases that are close in time, or in space, as the test statistic. A test of significance is obtained by calculating the mean number of close pairs for sets of randomly selected controls and assuming a Poisson distribution with this mean. Pike and Smith pointed out that the randomization distribution of X could be obtained and suggested further the use of matched controls.

1.2.13. Goldstein and Cuzick's temporal-spatial test

Goldstein and Cuzick (1978) proposed a method for identifying temporal-spatial patterns so that environmental events could be linked to health. They developed the method for use with daily counts of emergency room visits to several New York hospitals for asthma and respiratory illness. The method involves two tests, performed sequentially. The first test determines whether or not variation in the number of visits occurring daily at one hospital can be explained as random fluctuation. The number of visits occurring on a given day is assumed to be a Poisson variable and Cox and Lewis' d-statistic is used to assess significant nonrandomness,

$$d = \frac{1}{\bar{N}(k-1)} \sum_{i=1}^{k} (N_i - \bar{N})$$

where k=number of days, N_i=number of visits on the ith day, and $\bar{N}$ is the average daily count. A modification of the d statistic was used to

ascertain whether there was any nonrandom variation not explained by a linear or quadratic trend. The second test, a 2xn χ^2 test of homogeneity, was applied to pairs of hospitals to determine whether or not observed nonrandom patterns were global or local in nature.

Chapter II

A NEW TEST FOR CONCORDANT NONRANDOM PATTERNS

2.0 Introduction

A test is developed in this chapter that can be used to determine whether or not two series of incidence data from different places display concordant nonrandom patterns. In epidemiologic studies, we sometimes have series of data from more than one location and want to know whether or not these series exhibit significant clustering in the same time unit(s), e.g. on the same days or in the same years. For example, Goldstein and Cuzick (1978) looked at series of counts of daily visits to hospitals for asthma or respiratory disorders to determine whether or not the counts were random and whether or not the patterns were similar between hospitals. The method they proposed involved sequential tests for nonrandomness and concordance. The method we present, unlike Goldstein and Cuzicks', is simultaneously sensitive to concordance between the series and to nonrandomness within each of the series. That is, the test will be statistically significant only if there is some degree of concordance between the series and some degree of nonrandomness. The more concordant the series are and the more nonrandom each series is, the smaller the p-value of the test will be.

The test is most sensitive to nonrandomness in the form of temporal clustering, so these terms will be used interchangeably. Temporal clustering generally has been considered to be present in a series if one

of the time cells contains a significantly larger than expected number of cases. As a result, both the EMM test and the scan test focus on maximum cell frequencies. Clearly, this is a fairly narrow definition of temporal clustering. The proposed test involves all of the time cells and hence is sensitive to a broader range of nonrandom patterns than either the EMM or scan test. For example, it will be sensitive to the presence of say, two unusually large cell frequencies neither of which is quite large enough to be considered significant by either the EMM test or the scan test. A more detailed description of the forms of temporal clustering the test is sensitive to, as well as the amount of clustering and concordance that must be present to obtain a significant result will be given later.

The test developed here has several advantages. First, it does not require a control or comparison group. The test uses actual counts, and therefore, may be applied to existing morbidity and mortality incidence data. Second, the test makes use of all of the data. Thus, it has more power than tests based on comparisons of ranks or on only the maxima. Finally, it is simple and easy to apply.

2.1 Description of the test

Consider two series of incidence data, A and B, occurring in two locations over t time cells (e.g. years), and containing n_1 and n_2 cases, respectively. Under the null hypothesis of no concordance between the two series, and no temporal clustering, the n_1 and n_2 cases are mutually independent and are distributed randomly among the t time cells with probability of a case being in a given time cell being 1/t.

Let

$$X_{ij} = \begin{cases} 1 \text{ if the i-th case in Series A and the j-th case in Series B occur in the same time cell} \\ 0 \text{ otherwise .} \end{cases}$$

The test statistic is

$$X = \sum_{i}^{n_1} \sum_{j}^{n_2} X_{ij} \quad . \tag{2.1}$$

Thus, X is the number of pairs of cases from different series occurring in the same time cell, and assumes a value in the range of 0 to $N=n_1n_2$. Values of X larger than expected indicate some degree of concordance between the two series and some degree of temporal clustering. Values of X much smaller than expected indicate a lack of concordance, but some degree of clustering. We are interested only in large positive values of X, hence, the test of significance will be one-sided.

Note that X is simply an unstandardized correlation coefficient. Note also that while the cases are mutually independent, the X_{ij} are not necessarily so. To see this, suppose that $X_{11}=1$, $X_{12}=1$, and $X_{21}=1$. Then, X_{22} must also be equal to 1. Thus, X is the sum of dependent Bernoulli random variables and is not binomial.

A second formula for X, preferable for computational purposes, is

$$X = \sum_{h=1}^{t} a_h b_h \quad , \tag{2.2}$$

where a_h and b_h are the number of cases from series A and B, respectively, falling into the h-th time cell.

The validity of the test depends on the implicit assumption of minimal population shifts over time. More specifically, in order to assume that the cases are uniformly distributed across the t time cells we implicitly assume that the underlying population remains constant over time. This assumption is reasonable for small t, but not for large t. Obviously, how large t can be depends on the time unit being used, e.g. days, weeks, years, and on the population dynamics of the geographic area being studied. The time unit that generally will be used with the test is one year, since this is the unit that most of the data are recorded in. Ederer, Myers, and Mantel (1967) found that when the time unit is one year, a five year time period generally works well.

As mentioned previouly, the appropriate type of data for use with this test is incidence, or count, data. This is an advantage because incidence data is usually readily available. Also, it is much more common to have information about the number of cases of the disease than about the size of the population at risk. The use of count data is justifiable because under the asumption of minimal population shifts over the time period being studied, it is valid to compare the counts within a series.

2.2. Illustrative examples

To clarify how the test works and what it is sensitive to, several contrived examples are presented. In the first example, (a), note that there appears to be a cluster in the first time cell of series A and in the last time cell of series B. Thus, clustering occurs in both series, but the patterns of the two series are not concordant. In the second example, (b), the two series are concordant, however, neither series contains a cluster. In the third example, (c), there appears to be

clustering in only one of the series. Finally, in the fourth example, (d), there appears to be clustering in both of the two series and the two series are concordant.

(a) A: 11 1 1 1 1
B: 1 1 1 1 11
X = 11 + 1 + 1 + 1+ 11 = 25

(b) A: 4 3 3 3 2
B: 4 3 3 3 2
X = 16 + 9 + 9 + 9 + 4 = 47

(c) A: 11 1 1 1 1
B: 4 3 3 3 2
44 + 3 + 3 + 3 + 2 = 55

(d) A: 11 1 1 1 1
B: 11 1 1 1 1
X =121 + 1 + 1 + 1 + 1 = 125

As can be seen, the value of X is much larger for (d) than for (a), (b), or (c), and, in fact, is significantly large. The values of X for (a), (b), and (c) are close to the expected value, which for these sample sizes is 45, and are not significant. This is as expected since X should be significant only when clustering and concordance both occur.

2.3. Specification of the forms of nonrandomness and concordance

In order to know to which situations to apply the test and how to interpret the results, we must have a clear understanding about the forms of nonrandomness and concordance the test is sensitive to and how nonrandom and concordant the series must be to achieve a significant test result. We shall first describe the forms of nonrandomness and concordance the test is sensitive to, and then consider how much clustering must be present in each series and how concordant the two

8eries must be to obtain significance. As we have not yet presented formulas for, or tables of, significance levels, this discussion will necssarily be heuristic in nature.

As stated previously, the test is most sensitive to nonrandomness in the form of temporal clustering, that is, in the form of an unusually large number of cases in one or more of the time cells in the series. For shorter series, there usually will be only one cluster in a series, though for longer series, say $t \geq 7$, or when n_1 or n_2 is large, there could be more than one cluster. It is unlikely that more than two clusters would be detected in short series. When there are two clusters, both of the clusters could be large as in (e), or one could be large and one moderately large as in (f).

(e)	A:	15	3	2	2	2	2	14
(f)	A:	16	3	2	2	4	2	11

Specifying the forms of concordance that, in conjunction with clustering, result in significance is difficult. We shall begin by defining perfect concordance. Perfect concordance occurs when the patterns of the two series are parallel, that is, the ratio of the counts in the two locations in each time cell equals some constant. Obviously, with real data, perfect concordance between two series will occur rarely. With this definition in mind, we shall now consider concordance as it occurs in conjunction with each of the forms of temporal clustering described above.

When each series contains one cluster, the two series will be said to be concordant if the cluster in the first series and the cluster in the second series occur in the same time cell, as illustrated previously in (d). If both series have two clusters, a number of configurations may

occur. First, if both series contain two large clusters, their patterns would be considered concordant if the clusters in one series are matched with the clusters in the second series, as in example (g).

(g)	A:	15	3	2	2	2	2	14
	B:	15	2	1	2	2	2	16

If one of the series contains two large clusters and the other contains one large and one moderately large cluster, the two series would be considered concordant if the clusters in the first series occur in the same time cell as the clusters in the second series, as in (h).

(h)	A:	15	3	2	2	2	2	14
	B:	16	3	2	2	4	2	11

If both series contain one large and one moderately large cluster, concordance can assume two possible forms. The first is illustrated in example (i). Here, the largest frequency in the first series is matched with the largest frequency in the second, and the next largest frequency in the first series is matched with the next largest frequency in the second series.

(i)	A:	2	2	5	3	3	10	15
	B:	3	2	2	4	2	11	16

In the second form, shown in (j), the largest frequency in one series occurs in the same time cell as the next largest frequency in the other, and vice versa.

(j)	A:	2	2	5	3	3	10	15
	B:	3	2	2	4	2	16	11

Note that X will be larger when the series have the form shown in (i) than when they have the form shown in (j). Finally, two series may be considered concordant if one series has one large and one moderately large cluster, the other series has one large cluster, and the two large

clusters occur in the same time cell, as shown in (k).

(k) A:	2	2	5	3	3	10	15
B:	3	3	4	3	5	5	17

We have been focusing attention on the clusters in the series, that is, on the largest and next largest counts. Since the test statistic uses all of the counts, we should examine the contribution of the smaller counts. Since the test statistic is based on fixed sums, the size and importance of the contribution of the smaller counts depends on the magnitude of the clusters. The larger the clusters are, the smaller the other counts will be, and the smaller their contribution to the sum will be. In fact, if the clusters are large enough, the test statistic may be significant even before the contribution of the smaller counts is considered. Although the same concordance properties described for clusters pertain to the smaller counts, there is more flexibility in the way the small counts in one series can be matched with the small counts in the other series and yet have concordance. Clearly though, the more perfectly the smaller counts are matched, the larger X will be.

Since we use the information in all of the time cells we must contend with many possible arrangements of the data. Therefore, it is not easy to specify exactly how much clustering must be present in each series and how concordant the series must be in order to obtain a significant test result. Examination of examples for selected sample sizes and values of t indicates that for the one cluster case, the size of the cluster in each series does not have to be as large as that specified by the EMM or scan test in order to achieve signifcance provided that the other cells contribute to X. If the cluster in one of the series is about the same size as that specified by the EMM or scan tst, then the cluster in the other series must also be about that size or larger, in order to obtain a

significant result. If, however, the cluster in one of the series is much larger than that specified by the EMM or scan test, then the cluster in the second series can be smaller, and a significant test result will still be obtained. How concordant the two series must be depends on the distribution of the remaining cases to the other time cells, the size of the cluster in the second series and the distribution of cases in that series, and the amount of concordance.

If most of the cases in each series occur in just a few of the time cells, extremely large clusters will occur which will dominate the test statistic. When both series have an extreme cluster, and these clusters occur in the same time cell, as in the first two examples shown below, we expect and obtain a significant result.

(1)	A:	40	0	0	0	0	0	0	(m)	A:	40	0	0	0	0	0	0
	B:	40	0	0	0	0	0	0		B:	33	7	0	0	0	0	0

When, however, both series have extreme clusters which do not occur in the same time cell, as in (n), we would not expect, but may obtain, a significant result.

(n)	A:	40	0	0	0	0	0	0
	B:	7	33	0	0	0	0	0

Observe that the cluster in series A is matched with the next-largest count in series B. Thus, while obtaining a significant result in this example may not be desirable, it is explicable. Observe further that if the next-largest count in series B is too small, the test is nonsignificant, as in the following example.

(o)	A:	40	0	0	0	0	0	0
	B:	6	34	0	0	0	0	0

Thus, when a significant result is obtained, the investigator should scrutinize the data to ascertain that significance was not obtained solely because of extreme clustering.

2.4 The significance level of the test

In the next chapter, the distributional properties of the test will be presented and the possible approximate tests of significance will be compared. However, in addition to determining the appropriate test of significance, it also is necessary to determine the appropriate significance level of the test. Recall that a significant test result indicates some degree of concordance and some degree of nonrandomness. Actually, we are interested only in test results that indicate the presence of a significant amount of concordance and a significant amount of temporal clustering. Thus, we can expect that at nominal significance levels such as $\alpha=.05$ or $\alpha=.01$, the amount of concordance and clustering present will not be sufficient for our purposes. Indeed, examination of numerous examples shows this to be the case. In order to ensure the presence of sufficient concordance and sufficient clustering, we must evaluate the test at some smaller p-value. We have chosen this smaller p-value to be $p=.0025$.

The rationale behind this choice is as follows: Consider concordance and temporal clustering to be two independent events. If the probability of each occurring is p_1 and p_2, respectively, then the probability of the two events occurring simultaneously, $p=p_1p_2$. Thus, if the probability of each event occurring is $p_1=p_2=.05$, then the joint probability will be $p=(.05)^2 = .0025$. According to this argument the p-value of the test for concordance and temporal clustering should be at least as small as $p=.0025$ to allow for concordance and clustering at

reasonable levels of significance. Systematic examination of pairs of series shows that at this significance level, the amount of concordance and clustering present are sufficient, except in such cases as depicted in examples (n) and (o).

As stated previously, we are willing to obtain a significant test result when the amount of clustering is somewhat less than the amount required for significance by other tests. We are willing to do this because we are testing simultaneously for concordance and clustering and are using the information in all of the cells, and therefore, wish to allow for more flexibility in the patterns the test is sensitive to.

2.5 Effects of Permutations of the Time Order

Before concluding this chapter, one other characteristic of the test for concordant clustering, namely, its lack of sensitivity to permutations of the time dimension, will be discussed.

Consider the following example:

(q)

A:	16	3	2	5	3	4	7
B:	6	4	4	5	3	3	15
X =	96	+12	+12	+25	+ 9	+12	+105 = 271

Permuting the time dimension, i.e. the columns, we obtain the following arrangement:

(r)

A':	16	7	3	3	5	2	4
B':	6	15	4	4	5	3	3
X =	96	+105	+12	+12	+25	+ 9	+ 9 = 271

There appears to be a temporal cluster in both series. However, the patterns of A and B are not parallel since the cluster in A occurs in the first time cell while the cluster in B occurs in the sixth and last time cell. However, the patterns of A' and B' are more similar since the

clusters occur in adjacent time cells. In some situations, it might be desirable if the value of X for A' and B' were larger than the value of X for A and B. Attempts to modify the test so that it is not invariant with respect to time order will be presented in Chapter 5.

If one had reason to believe that the same disease process was operating in two locations, but that its onset was slightly delayed in the second location, it would be possible to shift the second series by that amount of time so as to align it with the first series. For example, if we had some apriori reason to believe that onset of the disease in location B began one year later than it did in location A, we could shift the two series as shown in (u).

(t)									(u)								
	A:	3	4	2	3	8	16	4		A:	-	3	4	2	3	8	16	4
	B:	4	3	4	2	3	8	16		B:	4	3	4	2	3	8	16	-

The test result for (ts) is nonsignificant, whereas, for (ut), it is significant.

Chapter III

DISTRIBUTIONAL PROPERTIES OF THE TEST

3.0 Introduction

In this chapter, the first five moments of X are derived and the distribution of X and the binomial distribution are compared. In addition, an expression for an exact p-value is derived and the appropriateness of the normal, Gram-Charlier series, Edgeworth series, and Pearson curve approximations are examined.

3.1. The moments of X

The first three moments of X have been derived by Grimson and Ingram(1982).

3.1.1. The expected value of X

The expected value of X is

$$E(X)=n_1n_2/t.$$

To see this, observe that $E(X_{ij}) = \frac{1}{t}$, and thus,

$$E(X) = E\left(\sum_{i}^{n_1} \sum_{j}^{n_2} X_{ij} \right) = \sum_{i}^{n_1} \sum_{j}^{n_2} E(X_{ij})$$

$$= \sum_{i}^{n_1} \sum_{j}^{n_2} 1/t$$

$$= n_1n_2/t \ .$$

Let $N=n_1n_2$, then

$$E(X) = N/t \ . \tag{3.1}$$

3.1.2. The variance of X

The expression for the second moment of X about zero can be partitioned into four special cases:

$$E(X_{ij}X_{ij}) = \frac{1}{t}, \quad E(X_{ij}X_{kj}) = \frac{1}{t^2}, \quad E(X_{ij}X_{il}) = \frac{1}{t^2}. \quad E(X_{ij}X_{kl}) = \frac{1}{t^2} \ .$$

$$\text{Thus, } E(X^2) = \sum_i^{n_1} \sum_j^{n_2} \sum_k^{n_1} \sum_l^{n_2} E(X_{ij}X_{kl})$$

$$= \sum_i^{n_1} \sum_j^{n_2} \frac{1}{t} + \sum_i^{n_1} \sum_j^{n_2} \sum_k^{n_1} \frac{1}{t^2} + \sum_i^{n_1} \sum_j^{n_2} \sum_l^{n_2} \frac{1}{t^2} + \sum_i^{n_1} \sum_j^{n_2} \sum_k^{n_1} \sum_l^{n_2} \frac{1}{t^2}$$

$$i \neq k \qquad j \neq l \qquad i \neq k, j \neq l$$

$$= n_1n_2/t + n_1(n_1-1)n_2/t^2 + n_1n_2(n_2-1)/t^2 + n_1(n_1-1)n_2(n_2-1)/t^2$$

$$= \frac{n_1n_2}{t}\left[1 + \frac{n_1n_2-1}{t}\right]. \tag{3.2}$$

This expression simplifies to

$$E(X^2) = \frac{N}{t} + \frac{N^{(2)}}{t^2} \ , \tag{3.3}$$

where $N^{(r)} = N(N-1)\ldots(N-r+1)$. The variance of X is

$$V(X) = \frac{N(t-1)}{t \ t} \ . \tag{3.4}$$

Observe that the mean and variance of X are identical to the mean and variance of a binomial variable with parameters N and p=1/t.

3.1.3. The third moment of X

The third moment of X about zero ,

$$E(X^3) = E(\sum_{i}^{n_1}\sum_{j}^{n_2}\sum_{k}^{n_1}\sum_{l}^{n_2}\sum_{m}^{n_1}\sum_{n}^{n_2} X_{ij}X_{kl}X_{mn}) \ .$$

can be partitioned into 25 terms as shown in Table 3.1.

Table 3.1
Terms of the third moment of X, $E(_{XijXklXmn})$

Combinations of Indices	Frequency	Power of t	Evaluation of $\sum\sum\sum\sum\sum\sum E$
i j k l m n	1	t^3	$n_1^{(3)}n_2^{(3)}$
i j i l m n	3	t^3	$n_1^{(2)}n_2^{(3)}$
i j k j m n	3	t^3	$n_1^{(3)}n_2^{(2)}$
i j i l i n	1	t^3	$n_1^{(1)}n_2^{(3)}$
i j k j m j	1	t^3	$n_1^{(3)}n_2^{(1)}$
i j i j m n	3	t^2	$n_1^{(2)}n_2^{(2)}$
i j i j i j	1	t	$n_1^{(1)}n_2^{(1)}$
i j k j i n	6	t^3	$n_1^{(2)}n_2^{(2)}$
i j i j i n	3	t^2	$n_1^{(1)}n_2^{(2)}$
i j i j m j	3	t^2	$n_1^{(2)}n_2^{(1)}$

Thus,

$$\begin{aligned} E(X^3) = \frac{1}{t^3}(&n_1^{(3)}n_2^{(3)} + 3n_1^{(2)}n_2^{(3)} + 3n_1^{(3)}n_2^{(2)} + n_1^{(1)}n_2^{(3)} \\ &+ n_1^{(3)}n_2^{(1)} + 6n_1^{(2)}n_2^{(2)}) \\ &+ \frac{1}{t^2}(3n_1^{(2)}n_2^{(2)} + 3n_1^{(1)}n_2^{(2)} + 3n_1^{(2)}n_2^{(1)}) \\ &+ \frac{1}{t}n_1n_2 \ . \end{aligned} \qquad (3.5)$$

Using the Stirling numbers of the first kind, s(q,k), which are defined

$$n^{(q)} = \sum_{k=1}^{q} s(q,k)n^k ,$$

(3.5) can be simplified as follows:

$$E(X^3) = \frac{1}{t^3}(n_1\ n_1^2\ n_1^3)\begin{vmatrix} s(1,1) & s(2,1) & s(3.1) \\ 0 & s(2,2) & s(3,2) \\ 0 & 0 & s(3,3) \end{vmatrix}\begin{vmatrix} 0 & 0 & 1 \\ 0 & 6 & 3 \\ 1 & 3 & 1 \end{vmatrix}\begin{vmatrix} s(1,1) & 0 & 0 \\ s(2,1) & s(2,2) & 0 \\ s(3,1) & s(3,2) & s(3,3) \end{vmatrix}$$

$$x \begin{vmatrix} n_2 \\ n_2^2 \\ n_2^3 \end{vmatrix} \qquad (3.6)$$

$$+ \frac{1}{t^2}(n_1\ n_1^2) \begin{vmatrix} s(1,1) & s(2,1) \\ 0 & s(2,2) \end{vmatrix}\begin{vmatrix} 0 & 3 \\ 3 & 3 \end{vmatrix}\begin{vmatrix} s(1,1) & 0 \\ s(2,1) & s(2,2) \end{vmatrix}\begin{vmatrix} n_2 \\ n_2^2 \end{vmatrix}$$

$$+ \frac{1}{t}(n_1 n_2)$$

$$= \frac{N^{(3)}}{t^3} + \frac{3N^{(2)}}{t^2} + \frac{N}{t} . \qquad (3.7)$$

This is just the third moment about zero of a binomial variable with parameters N and 1/t.

The third moment about the mean is

$$E(X-\mu)^3 = E(X^3) - 3E(X^2)E(X) + 2(E(X))^3$$

$$= \frac{N(2-3t+t^2)}{t\quad t^2} . \qquad (3.8)$$

3.1.4. The fourth moment of X

The fourth moment of X about zero is:

$$E(X^4) = E\Big(\sum_{i}^{n_1}\sum_{j}^{n_2}\sum_{k}^{n_1}\sum_{l}^{n_2}\sum_{m}^{n_1}\sum_{n}^{n_2}\sum_{o}^{n_1}\sum_{p}^{n_2} X_{ij}X_{kl}X_{mn}X_{op}\Big).$$

This sum can be partitioned into 324 terms as shown in Table 3.2. Thus,

$$E(X^4) = \frac{1}{t^4}(n_1^{(4)}n_2^{(4)} + 6n_1^{(4)}n_2^{(3)} + 6n_1^{(3)}n_2^{(4)} + 7n_1^{(4)}n_2^{(2)}$$
$$+ 7n_1^{(2)}n_2^{(4)} + n_1^{(4)}n_2^{(1)} + n_1^{(1)}n_2^{(4)} + 30n_1^{(3)}n_2^{(3)}$$
$$+ 24n_1^{(3)}n_2^{(2)} + 24n_1^{(2)}n_2^{(3)})$$
$$+ \frac{1}{t^3}(6n_1^{(3)}n_2^{(3)} + 18n_1^{(3)}n_2^{(2)} + 18n_1^{(2)}n_2^{(3)} + 42n_1^{(2)}n_2^{(2)}$$
$$+ 6n_1^{(3)}n_2^{(1)} + 6n_1^{(1)}n_2^{(3)})$$
$$+ \frac{1}{t^2}(7n_1^{(2)}n_2^{(2)} + 7n_1^{(2)}n_2^{(1)} + 7n_1^{(1)}n_2^{(2)})$$
$$+ \frac{1}{t}(n_1n_2) .$$

Using the Stirling numbers of the first kind, this simplifies to

$$E(X^4) = \frac{N^{(4)}}{t^4} + \frac{6N^{(3)}}{t^3} + \frac{7N^{(2)}}{t^2} + \frac{N}{t} + 6(\frac{1}{t^3} - \frac{1}{t^4})(N^2 + N - n_1N - n_2N). \quad (3.9)$$

The last term on the righthand side is the increment of $E(X^4)$ over the binomial fourth moment. Because of the dependencies among the X_{ij}, $E(X_{ij}X_{in}X_{mj}X_{mn})=1/t^3$ and not $1/t^4$ as it would if the X_{ij} were independent. Thus, the fourth moment of X is larger than that of a binomial variable. For this reason, the symmetric form of the third moment given in 3.6 cannot be generalized.

3.1.5. The fifth moment of X

The fifth moment of X about zero is

$$E(X^5) = E(\sum_{i}^{n_1}\sum_{j}^{n_2}\sum_{k}^{n_1}\sum_{l}^{n_2}\sum_{m}^{n_1}\sum_{n}^{n_2}\sum_{o}^{n_1}\sum_{p}^{n_2}\sum_{q}^{n_1}\sum_{r}^{n_2} X_{ij}X_{kl}X_{mn}X_{op}X_{qr}).$$

As for the first four moments, the above sum can be partitioned into numerous terms. The fifth moment of X is identical to that of a

TABLE 3.2

Terms of the fourth moment of X, $E(x_{ij}X_{kl}X_{mn}X_{op})$

Combinations of Indices	Frequency	Power of t	Evaluation of $\Sigma\Sigma\Sigma\Sigma\Sigma\Sigma\Sigma\Sigma E$
i j k l m n o p	1	t^4	$n_1^{(4)}n_2^{(4)}$
i j i l m n o p	6	t^4	$n_1^{(3)}n_2^{(4)}$
i j k j m n o p	6	t^4	$n_1^{(4)}n_2^{(3)}$
i j i l i n o p	4	t^4	$n_1^{(2)}n_2^{(4)}$
i j k j m j o p	4	t^4	$n_1^{(4)}n_2^{(2)}$
i j i l i n i p	1	t^4	$n_1^{(1)}n_2^{(4)}$
i j k j m j o j	1	t^4	$n_1^{(4)}n_2^{(1)}$
i j i j m n o p	6	t^3	$n_1^{(3)}n_2^{(3)}$
i j i j i n o p	12	t^3	$n1^{(2)}n_2^{(3)}$
i j i j m j o p	12	t^3	$n_1^{(3)}n_2^{(2)}$
i j i j i n i p	6	t^3	$n_1^{(1)}n_2^{(3)}$
i j i j m j o j	6	t^3	$n_1^{(3)}n_2^{(1)}$
i j i j i j i p	4	t^2	$n_1^{(1)}n_2^{(2)}$
i j i j i j o j	4	t^2	$n_1^{(2)}n_2^{(1)}$
i j i j i j o p	4	t^2	$n_1^{(2)}n_2^{(2)}$
i j i l m j o p	30	t^4	$n_1^{(3)}n_2^{(3)}$
*i j i n m j m n	6	t^3	$n_1^{(2)}n_2^{(2)}$
i j i l m j m p	12	t^4	$n_1^{(2)}n_2^{(3)}$
i j i j i n o n	12	t^4	$n_1^{(3)}n_2^{(2)}$
i j i j m n m n	3	t^2	$n_1^{(2)}n_2^{(2)}$
i j i j m n m p	6	t^3	$n_1^{(2)}n_2^{(3)}$
i j i j m n o n	6	t^3	$n_1^{(3)}n_2^{(2)}$
i j i l i n o j	12	t^4	$n_1^{(2)}n_2^{(3)}$
i j k j m j i p	12	t^4	$n_1^{(3)}n_2^{(2)}$
i j i j i n o j	12	t^3	$n_1^{(2)}n_2^{(2)}$
i j i j i n i n	3	t^2	$n_1^{(1)}n_2^{(2)}$
i j i j m j m j	3	t^2	$n_1^{(2)}n_2^{(1)}$
i j i j i n o n	12	t^3	$n_1^{(2)}n_2^{(2)}$
i j i j m j m p	12	t3	$n_1^{(2)}n_2^{(2)}$
i j i l m n m p	3	t4	$n_1^{(2)}n_2^{(4)}$
i j k j m n o n	3	t^4	$n_1^{(4)}n_2^{(2)}$

*X and the binomial variable with parameters N and 1/t have different powers of t for this term.

Table 3.3
Terms in the fifth moment of X that are affected by the dependencies among the X_{ij}

Combinations of Indices	Frequency	Power of t	Evaluation of $\sum\sum\sum\sum\sum E$
i j m j i n m n q r	30	t^4	$n_1^{(3)}n_2^{(3)}$
i j m j i n m n m r	60	t^4	$n_1^{(2)}n_2^{(3)}$
i j m j i n m n q n	60	t^4	$n_1^{(3)}n_2^{(2)}$
i j m j i n m n m n	120	t^3	$n_1^{(2)}n_2^{(2)}$

binomial variable with parameters N and 1/t except for those terms affected by dependencies among the X_{ij}. Table 3.3 presents the terms affected by these dependencies.

The expression for the fifth moment of X simplifies to

$$E(X^5) = \frac{N^{(5)}}{t^5} + \frac{10N^{(4)}}{t^4} + \frac{25N^{(3)}}{t^3} + \frac{15N^{(2)}}{t^2} + \frac{N}{t}$$
$$+ 30\left(\frac{1}{t^4} - \frac{1}{t^5}\right)(N-4)(N^2 - n_1N - n_2N + N) \qquad (3.10)$$
$$+ 120\left(\frac{1}{t^3} - \frac{1}{t^4}\right)(N^2 - n_1N - n_2N + N) .$$

The first line of 3.10 is just the binomial fifth moment about zero. The last two terms in the above equation are the increment of $E(X^5)$ over the binomial fifth moment.

3.2. Comparison of X with the binomial

Since the first three moments of X are identical to those of a binomial variable with parameters $N=n_1n_2$ and $p=1/t$, and since the fourth and fifth moments of X are similar to the binomial fourth and fifth moments, we expect that the distribution of X and the binomial distribution should bear some resemblance to each other. In this section,

we shall examine this resemblance by comparing the fourth and fifth moments of these two distributions and by comparing the coefficients of skewness and kurtosis for X, the binonmial, and the normal distribution.

As $N\to\infty$, the ratio of the fourth moment of X about zero to the fourth binomial moment about zero approaches 1 as shown below.

$$\lim_{N\to\infty}(E(X^4)/E(B(N,1/t)^4) = \lim_{N\to\infty} \frac{\frac{1}{6}(t^3-t^4)(N^2-n_1N-n_2N+N)}{\frac{N^{(4)}}{t^4} + \frac{6N^{(3)}}{t^3} + \frac{7N^{(2)}}{t^2} + \frac{N}{t}} \to 0 .$$

The same relationship applies to the fifth moments. Empirical investigation has shown how close to this limiting value the ratios for selected n_1, n_2, and t come. See Table 3.4.

An expression for the limit, as $N\to\infty$, of the ratio of the fourth central moment of X to the fourth central moment of the binomial is shown below:

$$\lim_{N\to\infty} E(X-\mu)^4/E(B(N,1/t)^4) = \lim_{N\to\infty} \frac{2(N-n_1-n_2+1)}{(N-2)(t-1) - t^2/3} \qquad (3.11)$$

Evaluating this expression for t=5, 7, and 9, we find that when t=5, (3.11) approaches 1.5, when t=7, (3.11) approaches 1.333, and when t=9, (3.11) approaches 1.25. Note that as t also increases, the ratio approaches 1.

An expression for the limit, as $N\to\infty$, of the ratio of the fifth central moments of X and the binomial is shown below:

$$\lim_{N\to\infty} E(X-\mu)^5/E(B(N,1/t)^5) = \lim_{N\to\infty} \frac{120(t-1)^2(N-n_1-n_2+1)}{(10t-4)(5N-6)-10(4N-5)+5t^3(2N-3)} \qquad (3.12)$$

Evaluating this expression for t=5, 7, and 9, we find that when t=5, (3.12) approaches 5.0, when t=7, (3.12) approaches 3.4, and when t=9,

(3.12) approaches 12/7=2.7143. See Table 3.5.

Two other measures of distributional shape, namely, the coefficients of skewness, $\beta_1 = \frac{(\mu_3^2)}{(\mu_2^3)}$, and kurtosis, $\beta_2 = \frac{\mu_4}{\mu_2^2}$, were examined. Keeping in mind that these coefficients do not always measure what they are supposed to, we nevertheless wish to compare them for X and the binomial, and to observe how they behave as N increases and as t varies. See Table 3.6. X and the binomial distribution have the same coefficient of skewness since their first three moments are identical. Thus, as $N \to \infty$, $\beta_1 \to 0$, which is the value for the normal distribution. Examination of β_2 for X, denoted by $\beta_2(X)$, reveals that as N increases, $\beta_2(X)$ increases towards the limit shown below,

$$\lim_{N\to\infty} \beta_2(X) = \lim_{N\to\infty} 3 + \frac{6}{t-1} + \frac{t^2-6t+12-6(n_1+n_2)}{N(t-1)} \to 3 + \frac{6}{t-1} ,$$

whereas β_2(binomial) decreases towards 3, which is the value of β_2 for the normal distribution. See Table 3.5. Thus, the distribution of X is leptokurtic. Note that for small values of t, say t=2, the coefficient of kurtosis for X is much bigger than the value for the normal distribution, but that as t increases, the coefficient approaches the normal value.

3.3. Evaluation of the significance of X

In order to use X as a test statistic for evaluating epidemicity, we must be able to calculate the exact significance level or estimate the approximate significance level of a given value of X. Recall that values of X, larger than the expected value, indicate some degree of concordance between the two series and some degree of temporal clustering. Thus, the test of significance will be one-sided.

Table 3.4

Comparison of the fourth(fifth) moment of X about zero and the fourth fifth) moment about zero of the binomial variable with parameters N and 1/t

t	n_1	n_2	N	$E(X^4)/E(B(N,1/t)^4)$	$E(X^5)/E(B(N,1/t)^5)$
5	20	20	400	1.0001	1.0006
5	30	30	900	1.00003	1.0001
5	20	50	1000	1.00002	1.0001
5	40	40	1600	1.000009	1.00004
5	50	50	2500	1.000004	1.00002
5	20	100	2000	1.000006	1.00003
5	50	100	5000	1.0000009	1.000005
5	100	100	10000	1.0000002	1.000001
5	100	200	20000	1.00000006	1.0000003
5	200	200	40000	1.00000001	1.00000007
5	200	500	100000	1.000000002	1.00000001
5	500	500	250000	1.0000000004	1.000000002
7	20	20	400	1.0002	1.0009
7	30	30	900	1.00004	1.0002
7	20	50	1000	1.00003	1.0002
7	40	40	1600	1.00001	1.00007
7	20	100	2000	1.000008	1.00004
7	50	50	2500	1.000005	1.00003
7	50	100	5000	1.000001	1.000007
7	100	100	10000	1.0000004	1.000002
7	100	200	20000	1.00000009	1.0000004
7	200	200	40000	1.00000002	1.0000001
7	200	500	100000	1.000000004	1.00000002
7	500	500	250000	1.0000000006	!.000000003
9	20	20	400	1.0002	1.001
9	30	30	900	1.00005	1.0003
9	20	50	1000	1.00004	1.0002
9	40	40	1600	1.00002	1.00009
9	20	100	2000	1.00001	1.00006
9	50	50	2500	1.000007	1.00004
9	50	100	5000	1.000002	1.000009
9	100	100	10000	1.0000005	1.000002
9	100	200	20000		
9	200	200	40000	1.00000003	1.0000001
9	200	500	100000	1.000000005	1.00000002
9	500	500	250000	1.0000000008	1.000000004

Table 3.5
Comparison of the fourth(fifth) central moments of X and the fourth(fifth) central moments of the binomial variable with parameters N and 1/t

t	n_1	n_2	N	$E(X-\mu)^4/E(B(N,1/t)^4)$	$E(X-\mu)^5/E(B(N,1/t)^5)$
5	20	20	400	1.451	4.615
5	30	30	900	1.467	4.740
5	20	50	1000	1.465	4.726
5	40	40	1600	1.475	4.804
5	20	100	2000	1.470	4.763
5	50	50	2500	1.480	4.842
5	50	100	5000	1.485	4.881
5	200	100	10000	1.490	4.921
5	100	200	20000	1.493	4.940
5	200	200	40000	1.495	4.960
5	200	500	100000	1.497	4.970
5	500	500	250000	1.498	4.990
7	20	20	400	1.300	3.168
7	30	30	900	1.311	3.244
7	20	50	1000	1.310	3.235
7	40	40	1600	1.317	3.282
7	20	100	2000	1.313	3.258
7	50	50	2500	1.320	3.305
7	50	100	5000	1.323	3.329
7	100	100	10000	1.327	3.352
7	100	200	20000	1.328	3.364
7	200	200	40000	1.330	3.376
7	200	500	100000	1.331	3.386
7	500	500	250000	1.332	3.397
9	20	20	400	1.225	2.548
9	30	30	900	1.233	2.602
9	20	50	1000	1.232	2.596
9	40	40	1600	1.237	2.630
9	20	100	2000	1.235	2.612
9	50	50	2500	1.240	2.647
9	50	100	5000	1.242	2.663
9	100	100	10000	1.245	2.680
9	100	200	20000	1.246	2.689
9	200	200	40000	1.247	2.697
9	200	500	100000	1.248	2.702
9	500	500	250000	1.249	2.700

TABLE 3.6
Comparison of β_1 and β_2 for X and the binomial distribution with parameters N and 1/t

t	n_1	n_2	β_1	β_2	
			X & binomial	binomial	X
5	20	20	.0036	3.0006	4.3544
5	20	30	.0038	3.0004	4.3779
5	20	40	.0028	3.0003	4.3897
5	20	50	.0025	3.0003	4.3968
5	20	60	.0019	3.0002	4.4015
5	20	70	.0016	3.0002	4.4048
5	20	80	.0014	3.0002	4.4073
5	20	90	.0013	3.0001	4.4093
5	20	100	.0011	3.0001	4.4108
5	70	70	.0005	3.0001	4.4375
5	70	80	.0004	3.0000	4.4601
5	70	90	.0004	3.0000	4.4622
5	70	100	.0003	3.0000	4.4638
5	80	80	.0004	3.0000	4.4628
5	80	90	.0003	3.0000	4.4648
5	80	100	.0003	3.0000	4.4665
5	90	90	.0003	3.0000	4.4669
5	90	100	.0003	3.0000	4.4685
7	20	20	.0104	3.0054	3.9079
7	20	30	.0064	3.0035	3.9219
7	20	40	.0052	3.0027	3.9290
7	20	50	.0042	3.0022	3.9332
7	20	60	.0035	3.0018	3.9360
7	20	70	.0030	3.0015	3.9380
7	20	80	.0026	3.0014	3.9395
7	20	90	.0023	3.0012	3.9407
7	20	100	.0021	3.0011	3.9416
9	20	20	.0153	3.0103	3.6956
9	20	30	.0102	3.0069	3.6956
9	20	40	.0097	3.0052	3.6998
9	20	50	.0061	3.0041	3.7024
9	20	60	.0051	3.0034	3.7041
9	20	70	.0044	3.0029	3.7053
9	20	80	.0038	3.0026	3.7062
9	20	90	.0034	3.0023	3.7069
9	20	100	.0031	3.0021	3.7074

3.3.1. Exact significance level of X

Attempts to derive an explicit expression for the distribution function of X have been fruitless. This is unfortunate because even if the expression could not be exactly evaluated, it is possible that it would have some known distribution as its limiting distribution. In either case, we would have a test of significance. In principle, the exact distribution of X could be calculated by enumerating and tabulating the t^{n1+n2} possible arrangements of the cases, which, under the null hypothesis, are all equally likely. However, even for small values of n_1, n_2, and t, t^{n1+n2} is so large that this is infeasible.

An expression for the exact p-value of X has been derived by Grimson and Ingram(1982),

$$P(X \geq x \text{ concordant pairs of cases}) = \sum_{k \geq X} P(k \text{ concordant pairs of cases})$$

where it is understood that the probabilities are conditional on n_1, n_2, and t. The probabilities can be expressed in terms of a classical occupancy problem where n_1+n_2 objects are arranged randomly in t cells. There are t^{n1+n2} possible arrangements of the cases, which under the null hypothesis, are all equally likely. The number of these arrangements yielding k concordances is the number of ways of arranging the cases into two sets of occupancy numbers $a_1, a_2, \ldots, a_t$ and $b_1, b_2, \ldots b_t$ such that

$$a_1b_1 + a_2b_2 + \ldots + a_tb_t = k, \text{ where } \sum_{h=1}^{t} a_h = n_1 \text{ and } \sum_{h=1}^{t} b_h = n_2.$$

Thus,

$$P(X \geq x \text{ concordant pairs}) = \frac{1}{\binom{n_1+n_2}{t}} \sum_{k \geq x} \sum_{\substack{a_1 b_1 = k_1 \\ \vdots \\ a_t b_t = k_t \\ k_1+\ldots+k_t=k \\ a_1+\ldots+a_t=n_1,\ b_1+\ldots+b_t=n_2}} \frac{n_1!}{a_1!\ldots a_t!} \frac{n_2!}{b_1!\ldots b_t!} . \tag{3.13}$$

This expression is difficult to evaluate even for small values of n_1, n_2, and t.

Given the difficulty of obtaining an exact test of significance, we attempt to find a good, yet simple, approximate test of significance.

3.3.2. Approximation of the significance level of X

Many of the disease clustering tests obtain a test of significance by assuming that the standardized test statistic, $Z=(X-E(X))/V(X)^{1/2}$, is asymptotically normally distributed. Use of the normal approximation is appealing because it is simple to calculate, tables are readily available, and epidemiologists and others who use the tests are familiar with the normal distribution.

In our case assuming X to be asymptotically normal may not be plausible because as we saw in Section 3.2, the coefficient of kurtosis does not approach the normal value as $N \to \infty$. However, the normal curve may still provide a reasonable approximation of the distribution of X since we can fit the first two moments exactly and the third moment will be asymptotically correct. Since as t increases the coefficient of kurtosis approaches 3, we might expect that the normal curve may provide a better fit of the distribution of X for large values of t.

The indication of anormality provided by the coefficient of kurtosis of X was further supported by other findings. First, because of the

dependencies that exist among the X_{ij}, the usual central limit theorems do not apply. The Frechet-Shohat central limit theorem (Puri and Sen, 1971) cannot be applied because a general expression for the moments of X has not been derived. An attempt was made to establish asymptotic normality through U-statistic theory since X can be expressed as a 2-sample U-statistic. When this attempt failed, because X is stationary of order 1, Monte Carlo simulation procedures were performed to determine how well the upper tail probabilities of the distribution of X could be approximated by the normal distribution. Finally, the Gram-Charlier Type A and Edgeworth series expansions and Pearson curves were examined as possible estimates of the distribution. These approaches are described in detail below.

(i) U-Statistics Approach

The U-statistic approach (Hoeffding, 1948; Lehmann, 1951; Fraser, 1957; Sen, 1960; Bhapkar, 1961; Puri and Sen, 1971) is motivated as follows:

Consider that we have two random samples of size n_1 and n_2 which are drawn from populations 1 and 2, respectively. Let a_i and b_j denote the time cells in which the ith case in population 1 and the jth case in population 2 occur. Thus, a_i and b_j assume integer values from 1 to t. The kernel of the 2-sample U-statistic is

$$w(a_i,b_j) = X_{ij} = \begin{cases} 1 & \text{if } a_i=b_j, \text{ that is if the ith case in population 1 and the jth case in population 2 occur in the same time cell} \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$X = \sum_{i}^{n_1} \sum_{j}^{n_2} X_{ij} = n_1 n_2 U.$$

M_1 and M_2 are defined to be the degree of ω, that they are the smallest sample sizes for which ω can be constructed. In this case, $m_1=m_2=1$. Define,

$$\zeta_{10} = \text{cov}[w(a_i,b_j),w(a_i,b_m)]$$
$$= \text{cov}(X_{ij},X_{im}) ,$$

and

$$\zeta_{01} = \text{cov}[w(a_i,b_j),w(a_k,b_j)]$$
$$= \text{cov}(X_{ij},X_{kj}) .$$

The following theorem applies,

Theorem(Lehmann and Sen): If U is a two-sample U-statistic for some estimable parameter, ω, with degree (M_1, M_2), and if $0<\max(\zeta_{10},\zeta_{01})<\infty$, then, as n_1 and n_2 tend to ∞ in fixed ratios (i.e. $n_1/n \to p_1>0$ and $n_2/n \to p_2>0$, as $n \to \infty$, where $n_1+n_2=n$),

$$\frac{U-\omega}{\frac{M_1^2\zeta_{10}}{n_1} + \frac{M_2^2\zeta_{01}}{n_2}}$$

has asymptotically a standard normal distribution.

Calculating ζ_{10},

$$\zeta_{10} = \text{cov}(X_{ij},X_{im}) = E(X_{ij} - E(X_{ij}))E(X_{im} - E(X_{im}))$$
$$= E(X_{ij}X_{im}) - E(X_{ij})E(X_{im})$$

$$= \frac{1}{t^2} - \frac{1}{t}\cdot\frac{1}{t}$$

$$= 0 .$$

Likewise, $\zeta_{01}=0$. Thus, X/n_1n_2 is stationary of order 1. Work by Hall (1979), Gregory (1977), and Neuhaus (1977) extend U-statistic theory to

cover this case. We presently are trying to determine the asymptotic distribution of X using their work.

(ii) Monte Carlo Simulation

Monte Carlo simulation techniques have been known for decades, but were not used extensively until the advent of high-speed electronic computers. The Monte Carlo method has numerous applications. The one we are interested in has been called distribution sampling or model sampling by some authors. The purpose is to estimate the empirical distribution of a statistic whose exact null distribution is unknown or too complicated to evaluate. The statistic is a known function of one or more random variables whose distribution is known. To estimate the distribution of the statistic, a value for each of the variables is randomly selected from their distributions and the corresponding value of the statistic computed. This procedure is repeated many times and the error in the estimate is a function of the number of replications.

The use of Monte Carlo procedures to obtain significance tests involves estimating the empirical distribution as described above and then ranking the observed value of the test statistic relative to the values of the test statistic obtained from the simulation. Specifically, for a one-sided test of significance, the null hypothesis is rejected at an approximate significance level of $\alpha=(N-M)/N$ if the observed value is larger than at least M of the N simulated values. A simplified Monte Carlo procedure for significance testing has been proposed by Barnard (1963) and elaborated by Hope (1968) , Besag and Diggle (1977), and Margott (1979) among others. It requires fewer replications, and hence, the computations are simpler. The procedure, a reinterpretation of the

usual Monte Carlo method, is based on both the observed value and the N simulated values. The null hypothesis is rejected at the exact significance level of (N-M+1)/(N+1), if the observed value of the statistic is larger than at least M of the N simulated values.

The random variates used in Monte Carlo simulation are generated from variates uniformly distributed in the unit interval. When performing simulations on the computer, one works with pseudo-random numbers, not random numbers. Pseudo-random numbers are generated by a deterministic algebraic formula, and thus, are not truly random since each number is dependent on the previous one. Pseudo-random numbers appear to be random and they behave like random numbers in that they pass standard statistical tests of randomness.

The discrete uniform random number generator in the International Mathematical Statistical Library (IMSL) was used in the simulations in this study. The generator produces pseudo-random numbers in the range [1,K] by multiplying the real equivalent of K by a pseudo-random number in the range of (0,1), adding 1 to the product, and truncating the result to an integer. This procedure is repeated until the number of random variables requested has been generated. The random variables from the unit interval are generated by a linear congruential generator with a multipier of 16807 and a modulus of $2^{31}-1$. For a given input seed, the sequences of variables produced on computer systems on which IMSL is installed are identical in the first 23 bits of the mantissa. The double precision input seed is automatically replaced by a new seed after a sequence of numbers has been generated, so that in making a series of calls a series of independent sequences is produced.

Since the exact null distribution of X is unknown, the exact p-value

is too complicated to evaluate, and complete enumeration of the distribution is infeasible, we resorted to the Monte Carlo method to simulate the distribution of X and to determine the best approximate test of significance. We would like the estimate of the distribution of X to be good enough to use as a reference distribution against which to compare the fit of the normal distribution, the Gram-Charlier Type A and Edgeworth series, and the Pearson curves. Note that a researcher could perform a Monte Carlo simulation, or a simplified Monte Carlo simulation, to evaluate the significance of a given X (Besag and Diggle, 1977; Foutz, 1980). However, since the test will generally be applied by nonstatisticians who may have limited knowledge of, or access to, computers, it would be preferable to offer a simpler significance test based on concrete formulas. With this goal in mind, simulations, each with 5000 replications, were carried out for specified sample sizes.

Simulations were performed for t=5 and all combinations of the sample sizes n_1=10(10)100 and n_2=10(10)100 and for t=7 and t=9 and all combinations of the sample sizes n_1=20(10)100 and n_2=20(10)100. Simulations also were performed for t=5,7, and 9 and the four sample size combinations (100,200), (200,200), (200,500), and (500,500). Remember that since $N=n_1n_2$, N will be large when both n_1 and n_2 are large, or when n_1 is large and n_2 small, or vice versa. Thus, it is necessary to consider what happens as $n_1 \to \infty$, as $n_2 \to \infty$, and as both n_1 and n_2 tend to infinity.

For each simulation, 5000 samples of integers in the range of [1,t] and of sizes n_1 and n_2 were generated independently using the IMSL discrete uniform random number generator. The frequency of numbers 1 to t in each sample was counted and X calculated for each pair of samples.

Finally, the empirical distribution was tabulated. The simulation programs are reproduced in Appendix 1. Plots of the frequency distribution of X when $n_1=n_2=100$ and t=5, 7, and 9, are given in Figures 3.7-3.9.

The goodness of fit of the simulated distribution to the exact null distribution of X was judged by comparing the mean and variance of the simulated distribution to the exact mean and variance. For all of the simulations, the sample mean was within ± 1% of the exact mean and usually within rounding error. The sample variance were within ± 3% of the exact variance for all sample size combinations and usually within ±2%. Comparison of the median, mode, and mean of the empirical distribution showed, that in many cases, the mode=mean=median, and that in all cases the mode is within ±1% of the median. In all cases, the median was within rounding error of the mean. The median, mode, and mean, for simulations with the same N but different n_1 and n_2, were within ± 1.0 of each other. Tables of the mean, mode, median, and variance of the simulated distributions and of the exact mean and variance can be found in Appendix 2.

Critical values for the 90th, 95th, 97.5th, 99th, 99.5th, and 99.75th percentiles of the distribution of X were determined from the simulation for combinations of n_1, n_2, and t. Since X assumes only integer values, it was not always possible to obtain critical values for the exact percentiles, in which case the value of X for the next larger percentile was used. The critical values for simulations with the same N, but different n_1 and n_2 were in close agreement. The critical values obtained from the simulations were treated as reference values against which to compare critical values obtained using the normal approximation,

FIGURE 3.1
FREQUENCY DISTRIBUTION OF X WITH N1=N2=100 AND T=5

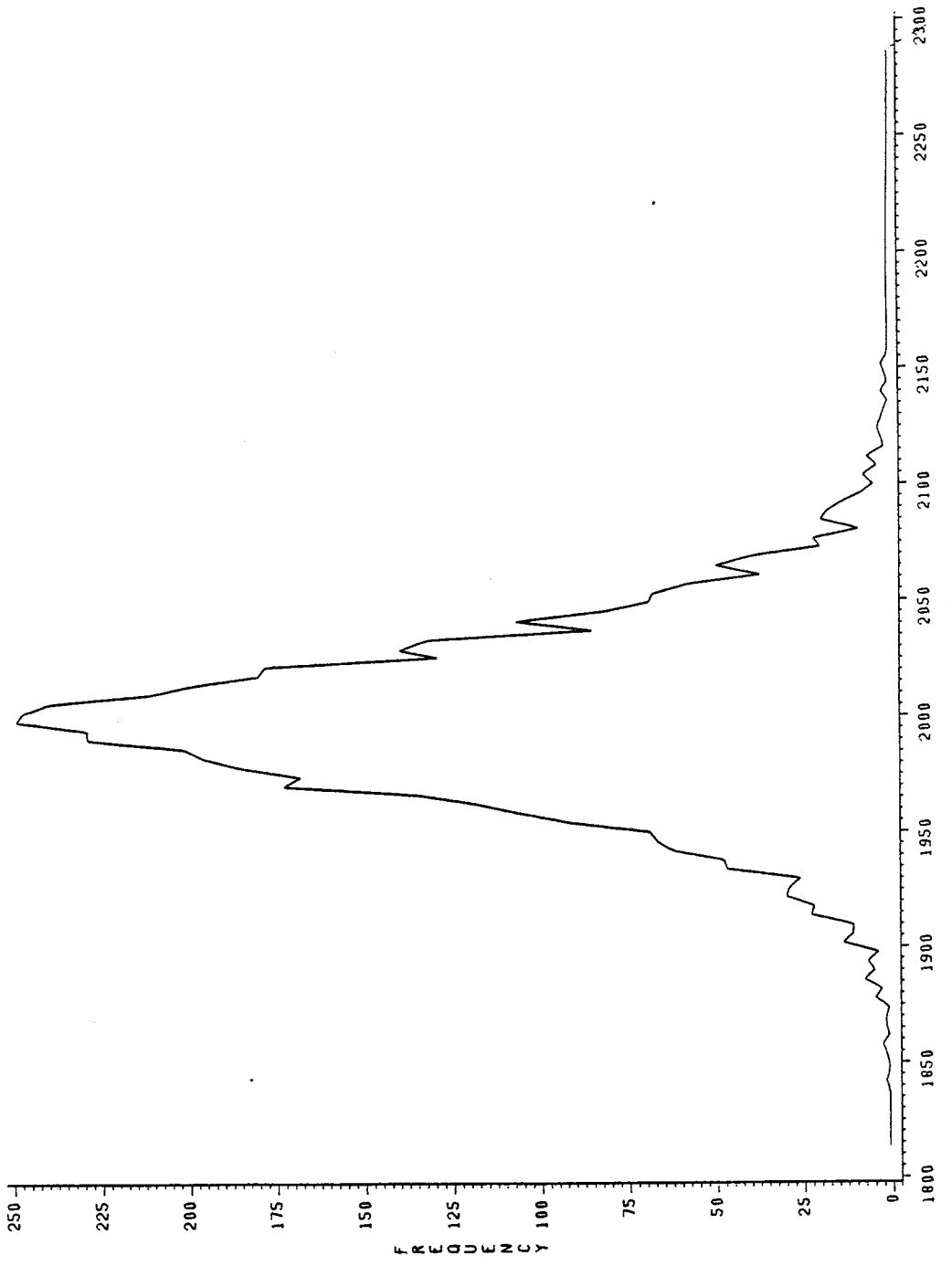


FIGURE 3.2
FREQUENCY DISTRIBUTION OF X WITH N1=N2=100 AND T=7

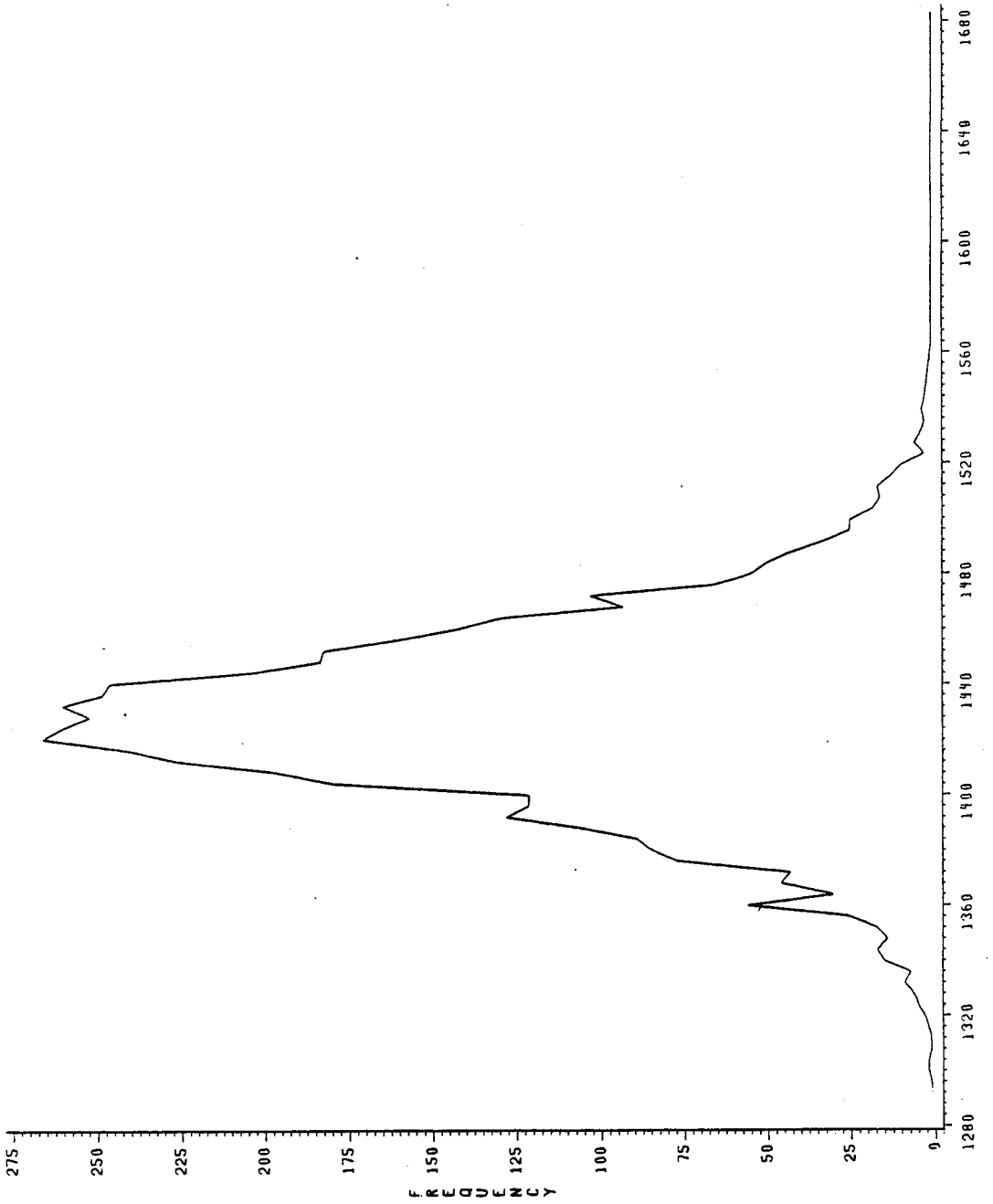
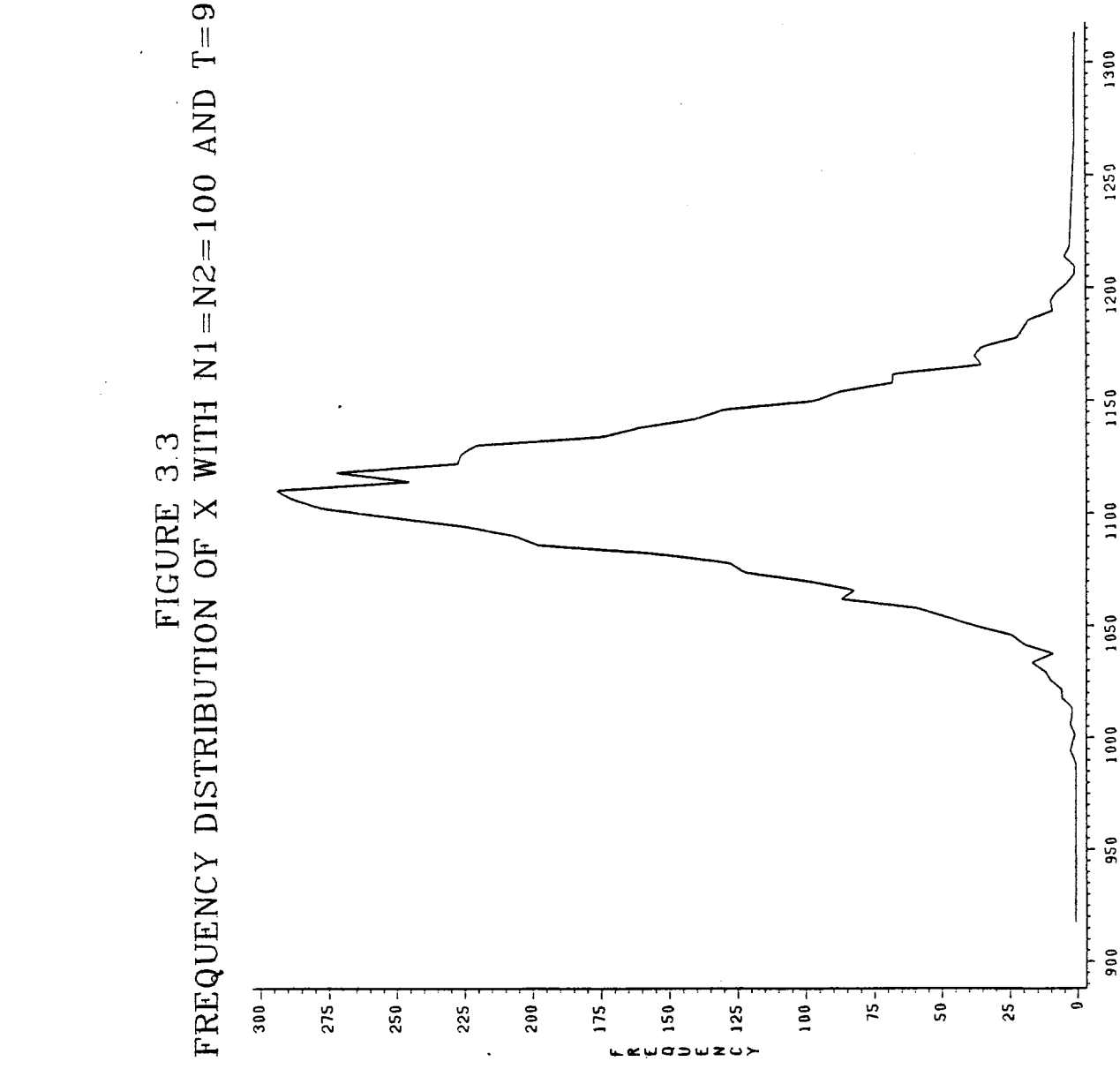


FIGURE 3.3
FREQUENCY DISTRIBUTION OF X WITH N1=N2=100 AND T=9



the Gram-Charlier and Edgeworth approximations and the Pearson curve approximations.

(iii) Gram-Charlier Type A series and Edgeworth series

Several systems of distributions have been developed that can be used to represent a probability density function as an expansion in series (Kendall and Stuart, 1964; Cramer, 1946). A curve that has the same first four moments as the density function is constructed and used to represent the density funtion. Such representations are useful when an explicit expression for the density function of a random variable does not exist, but at least the first four moments of the density function are known. The Gram-Charlier Type A series and the Edgeworth series, the two most commonly used series expansions, represent the expansion of the density funtion as a series of derivatives of the standard normal density funtion.

A probability density function can be expanded in a Gram-Charlier Type A series

$$f(X) = \sum_{j=0}^{\infty} c_j H_j(X)\phi(X),$$

where the c_j are constant expressions involving the moments of the distribution and the $H_j(X)$, known as the Chebyshev-Hermite polynomials, are coefficients of the jth order derivatives of the standard normal density function, $\phi(x)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$.

Thus,

$$f(X) = \phi(X)\left[1 + \frac{1}{2}(\mu_2-1)(X^2-1) + \frac{1}{6}\mu_3(X^3-3X) + \frac{1}{24}(\mu_4-6\mu_3+3)(X^4-6X^2+3)\right.$$

$$\left. + \frac{1}{120}(\mu_5 - 10\mu_3)(X^5 - 10X^3 + 15X) + \ldots\right]$$

where μ_i, $i>1$, is the ith moment about the mean. When f(X) is in standard measure, that is $Y=[X-E(X)]/V(X)^{1/2}$, then $\mu_1=0$. $\mu_2=1$, and in general, the rth moment is divided by σ^r. The equation above reduces to

$$f(Y) = \phi(Y)[1 + \frac{1}{6}\frac{\mu_3}{\sigma^3}(Y^3-3Y) + \frac{1}{24}(\frac{\mu_4}{\sigma^4}-3)(Y^4-6Y^2+3) + \frac{1}{120}(\frac{\mu_5}{\sigma^5} - \frac{10\mu_3}{\sigma^3})(Y^5 + 10Y^3 - 15Y) + \ldots]$$

The distribution function of f(X) is

$$F(Y) = \int_{-\infty}^{y} f(Y)dy = \Phi(Y) - \sum c_j H_{j-1}(Y)\phi(Y).$$

$$= \Phi(Y) - \phi(Y)[\frac{1}{6}\frac{\mu_3}{\sigma^3}(Y^2-1) + \frac{1}{24}(\frac{\mu_4}{\sigma^4}-3)(Y^3-4Y) + \frac{1}{120}(\frac{\mu_5}{\sigma^5}-\frac{\mu_3}{\sigma^3})(Y^4-6Y^2+3) + \ldots]$$

where, $\Phi(Y) = \int_{-\infty}^{y} \phi(Y)dy$.

The Edgeworth series is similar to the Gram-Charlier Type A series and, in fact, is identical to it when only the first three moments are used. When f(X) is in standard measure, the distribution function has the form,

$$F(Y) = \Phi(Y) - \phi(Y)[\frac{1}{6}\frac{\mu_3}{\sigma}(X^2-1) + \frac{1}{24}(\frac{\mu_4}{\sigma}-3)(X^3-3X) + \frac{1}{72}(\frac{\mu_3}{\sigma})^2(X^5-10X^3-15X) + \frac{1}{120}(\frac{\mu_5}{\sigma^5}-10\frac{\mu_3}{\sigma^3})(X^4-6X^2+3) + \frac{1}{144}\frac{\mu_3}{\sigma^3}(\frac{\mu_4}{\sigma^4}-3)(X^6-15X^4+45X^2-15) + \frac{1}{1296}(\frac{\mu_3}{\sigma^3})^3(X^8-28X^6+210X^4-420X^2+105)] \quad .$$

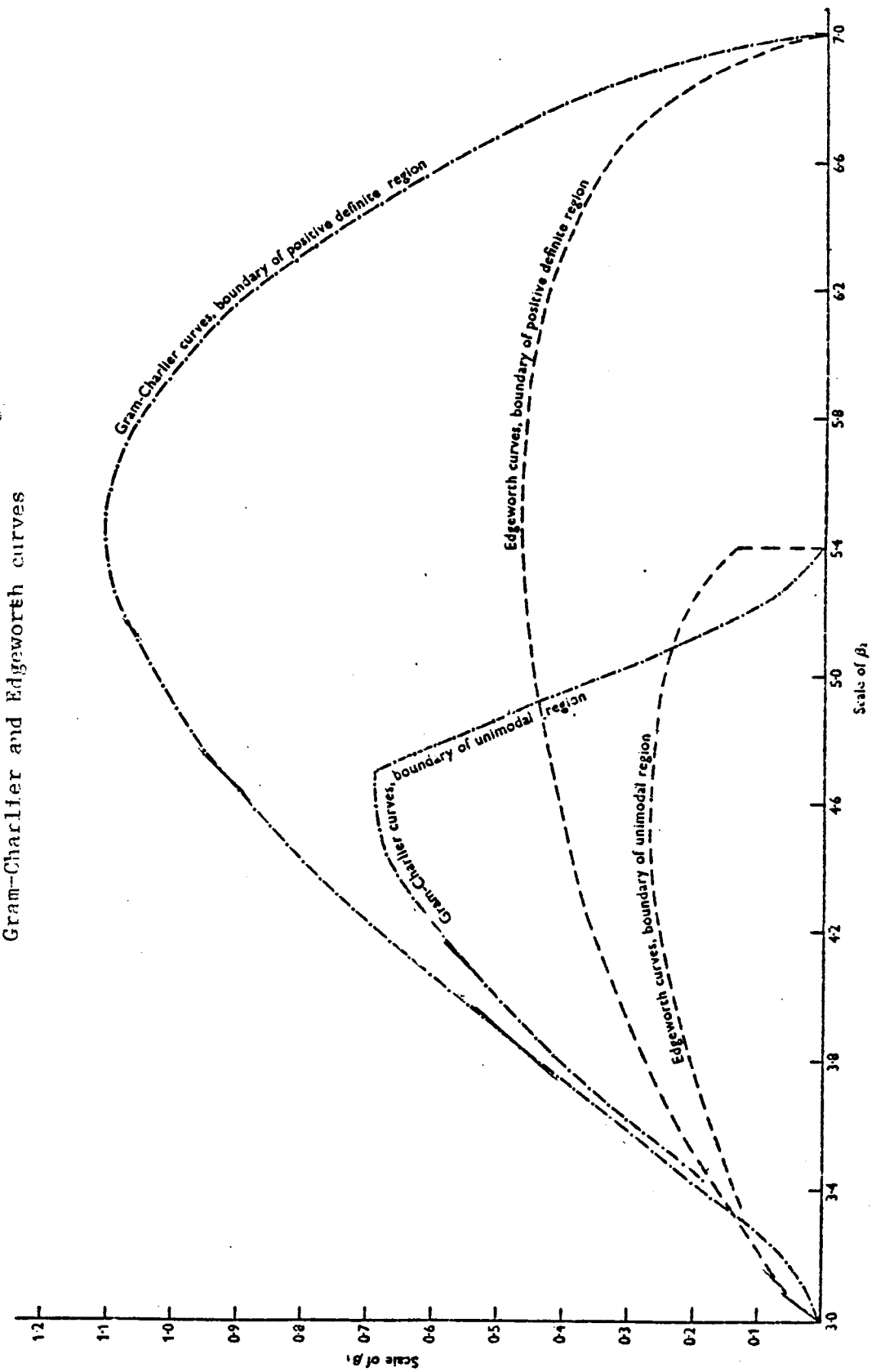
For practical reasons, when representing a frequency function by these series, only a finite number of terms is taken into account, and

one hopes that the remainder is negligible. Usually only those terms involving the first four moments are used because sampling effects render higher moments calculated from the data unreliable. When the exact moments are used, sampling effects are not a problem and the series expansion can be extended, though usually not past the term involving $H_6(X)$. Johnson and Kotz (1969) recommend using only the first three moments and including a continuity correction when standardizing X, that is, $Z = [X + 1/2 - E(X)]/V(X)^{1/2}$.

The Gram-Charlier Type A and Edgeworth series have several shortcomings when only a finite number of the terms in the series is used. First, negative frequencies, particularly in the tails may result. Secondly, the curves may not be unimodal. Thirdly, for the Gram-Charlier series, a lower order series may fit better than a higher order series, e.g. one with k-1 terms versus k terms. Barton and Dennis (1952) mapped out the regions in the (β_1,β_2) plane where f(X) is positive and unimodal. Recall that β_1 and β_2 were defined in Section 3.2. They found that when only the first four moments are used, the approximation is adequate only for distributions with moderate skewness, $|\sqrt{\beta_1}|<.7$ and with $3.0<|\beta_2|<5.4$ (see Figure 3.1). Berndt (1957) mapped the regions of unimodality and positiveness for the Gram-Charlier and Edgeworth series when only the first three moments are used.

Calculation of $\beta 1$ and $\beta 2$ for selected n_1, n_2, and t showed that both were within the ranges specified by Barton and Dennis (See Figure 3.1) and thus, the Gram-Charlier and Edgeworth approximations should be adequate. To evaluate the goodness of fit of these approximations, we used them to calculate cumulative percentiles for values of the test statistic obtained from the simulation and compared these percentiles

Figure 3.4
The β_1, β_2 plane showing unimodal and positive definite regions for the Gram-Charlier and Edgeworth curves



with those specified by the simulation. The expansions involving the first three moments, the first four moments, or the first five moments were compared to determine which was best. We calculated the Gram-Charlier expansions with and without a continuity correction.

(iv) Pearson curves

The curves in the Pearson system were obtained as solutions of the differential equation,

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1x + b_2x^2} ,$$

where $y=f(x)$ is a density function. In practice, terms higher than x^2 are not used (Kendall and Stuart, 1964; Elderton and Johnson, 1969). Pearson defined three main types of curves, Type I,IV, and VI, and eight transition types, on the basis of the nature of the roots of the equation, $b_0+b_1x+b_2x^2$. The criterion,

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} ,$$

which is used to classify the curves, can be calculated from the population or sample moments.

For the various sample sizes we tried, κ was positive and <1, β_1 was also positive and close to zero, and β_2 was greater than 3. Thus, the distribution of X would be classified as a Type IV curve, though it could be either transition Type VII, which has $\kappa=0$, $\beta_1=0$, and $\beta_2>3$. or possibly Type II, the normal curve which has $\kappa=0$, $\beta_1=0$, and $\beta 2=3$. For Types II and VII, the mode equals the mean, which is not always true for X.

It is not necessary to determine the precise form of the function in order to evaluate significance. Johnson, Nixon, and Amos (1963) published

tables of selected upper and lower percentiles for the Pearson curves with $0.0 \leq \sqrt{\beta_1} \leq 2.0$ and $1.6 \leq \beta_2 \leq 14.4$. To use these tables, one need only compute $\sqrt{\beta_1}$ and β_2 and interpolate on them to obtain the value of the standardized test statistic at a given percentile. Johnson et al. recommended using second differences for the interpolation, however, we found that linear interpolation yielded almost identical results.

(v). Comparison of the goodness of fit of the Pearson curve and normal approximations

Using Johnson et al.'s tables and the exact values of β_1 and β_2 for selected n_1, n_2, and t, we calculated the critical values of X for the 90th, 95th, 97.5th, 99th, 99.5th, and 99.75th percentiles. Comparison of these values with those obtained from the simulations, showed them to be in close agreement. The Pearson values usually were identical to the sample values, and in all cases were within ±2 units.

The critical values specified by the normal approximation were obtained by solving the equation

$$Z = [V(X)^{1/2}]\Phi^{-1}(X) + E(X)$$

for Z, and rounding to the nearest integer, where E(X) and V(X) are the exact mean and variance of X, and $\Phi^{-1}(X)$ is the value of the standard normal distribution at a given percentile. The sample and normal approximation 95th percentiles were in close agreement for all sample sizes, usually being identical to each other or within ±1 unit of each other, and with none differing by more than ±2 units. The 90th and 97.5th percentile critical values for the simulated distribution and the mormal approximation tended to be within ±2 units of each other. The critical values for the simulated and normal approximation 99th percentiles, however, are not close; the 99th percentile value specified

by the normal approximation is smaller than that specified by the simulation. In fact, the critical value specified for the 99.5th percentile by the normal approximation is close to the simulated 99th percentile critical value. Likewise, the critical value specified for the 99.5th percentile by the normal approximation is closer to the simulated 99.75th percentile than to the simulated 99.5th percentile. Thus, if the true significance level of the test is in the 99th percentile or higher, the normal approximation is anti-conservative, that is, use of the approximation will yield p-values that are too small.

A sample of the simulated, Pearson, and normal approximation values are given in Table 5.7; more extensive tables can be found in Appendix 2. As can be seen, the Pearson values are in much closer agreement with the simulated values than are the normal values, especially for percentiles above the 99th.

(vi) Comparison of the Gram-Charlier series, Edgeworth series, Pearson curve, and normal approximations

Critical values for the 90th, 95th, 97.5th, 99th, 99.5th, and 99.75th percentiles of the distribution of X were determined from the simulation for combinations of n_1, n_2, and t. These values were standardized and the normal, Gram-Charlier, and Edgeworth cumulative percentiles calculated. In addition, the cumulative percentiles for the Pearson values were obtained from the simulated distribution. All of these percentiles are displayed in Tables 5.8 through 5.28 for purposes of comparison. As can be seen, of all the approximations, the Pearson curves fit the upper-tail of the distribution of X most closely. The Gram-Charlier series with four moments fits better than the other Gram-Charlier or Edgeworth series approximations, and somewhat better than the normal approximation. It is unclear whether the Gram-Charlier

TABLE 3.7

Approximate upper percentile critical values for the distribution of X with t=5 and selected sample sizes, obtained from the simulations (S), the Pearson curve tables(P), and the normal approximation (N)

n_1	n_2	90.00			95.00			97.50			99.00			99.50			99.75		
		S	P	N	S	P	N	S	P	N	S	P	N	S	P	N	S	P	N
20	20	90	90	90	94	94	94	96	96	96	101	101	99	105	104	101	109	108	103
20	100	421	422	423	430	429	429	436	436	435	445	445	442	452	452	446	462	460	451
100	100	2048	2048	2051	2065	2065	2066	2080	2080	2078	2100	2101	2093	2117	2116	2103	2141	2133	2112
100	200	4069	4091	4093	4094	4091	4093	4118	4113	4111	4150	4142	4132	4168	4164	4146	4193	4187	4159
200	200	8097	8129	8103	8132	8129	8132	8168	8160	8157	8215	8201	8186	8244	8233	8206	8265	8266	8225
200	500	20154	20204	20162	20206	20204	20208	20255	20253	20248	20312	20317	20294	20371	20368	20326	20411	20420	20355
500	500	50233	50322	50256	50322	50322	50329	50404	50399	50392	50502	50502	50465	50567	50581	50515	50602	50664	50562

Table 3.8
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes $n_1=20$, $n3_2=20$, and for t=5

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/CC	Gram-Charlier with 4 moments w/0 CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
91.26	91.26	89.44	88.31	90.38	91.16	91.92	89.31	92.78	93.63	90.64
95.10	95.10	94.68	94.46	95.20	94.81	95.34	94.47	96.22	96.65	95.75
97.68	97.68	97.72	97.52	97.85	96.90	97.39	97.53	97.52	97.63	98.17
99.12	99.12	99.55	99.46	99.56	98.12	98.86	99.46	98.45	98.53	99.27
99.56	99.46	99.91	99.87	99.89	99.45	99.50	99.87	99.14	99.18	99.56
99.76	99.72	99.99	99.98	99.98	99.86	99.90	99.98	99.73	99.80	99.84

Table 3.9
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes $n_1=20$, $n_2=100$, and fort=5

% from simul	Pearson curve	Mormal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.16	91.02	87.98	87.94	88.94	89.87	90.62	87.54	90.89	91.35	88.51
95.38	94.88	95.32	95.22	95.49	95.40	95.58	95.22	96.02	96.19	95.79
97.64	97.64	97.79	97.70	97.84	97.05	97.13	97.70	97.32	97.39	97.99
99.04	99.04	99.41	99.36	99.40	98.53	98.59	99.36	98.44	98.49	99.31
99.50	99.50	99.82	99.80	99.80	99.25	99.23	99.80	99.10	99.14	99.67
99.78	99.74	99.97	99.96	99.98	99.78	99.98	99.96	99.70	99.87	99.88

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.10
Comparison of the upper percentiles specified by the simulation, the Pearson curves the normal approximation, and the gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=100, and for t=5

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.34	90.34	88.49	88.47	88.73	90.70	90.90	88.47	91.04	91.24	88.74
95.24	95.24	94.79	94.75	94.89	95.13	95.22	94.78	95.43	95.52	95.02
97.62	97.62	97.72	97.68	97.75	97.02	97.07	97.68	97.15	97.20	97.82
99.00	99.06	99.38	99.36	99.38	98.49	98.51	99.36	98.45	98.47	99.34
99.50	99.48	99.83	99.82	99.83	99.27	99.29	99.81	99.21	99.23	99.76
99.76	99.74	99.98	99.98	99.98	99.79	99.81	99.98	99.76	99.78	99.94

Table 3.11
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=200, and for t=5

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.06	90.06	88.88	88.86	89.03	91.02	91.15	88.86	91.26	91.39	89.06
95.18	94.72	95.17	95.14	95.23	95.38	95.36	95.14	95.59	95.65	95.33
97.54	97.10	98.15	98.12	98.16	97.34	976.37	98.12	97.40	97.43	98.20
99.04	98.86	99.60	99.59	99.60	98.80	98.84	99.59	98.76	98.79	99.59
99.52	99.44	99.85	99.84	99.85	99.33	99.35	99.84	99.29	99.30	99.80
99.76	99.76	99.97	99.97	99.97	99.75	99.75	99.97	99.72	99.72	99.94

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.12
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=200, and for t=5

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.18	90.18	88.73	88.71	88.85	90.91	91.02	88.72	91.09	91.19	88.86
95.00	94.78	95.05	95.03	95.09	95.32	95.41	95.03	95.47	95.50	95.16
97.50	97.02	98.21	98.19	98.21	97.39	96.31	98.20	97.43	96.41	98.24
99.00	98.84	99.64	99.63	99.63	98.88	98.88	99.63	98.85	98.85	99.61
99.50	99.38	99.89	99.89	99.89	99.42	99.43	99.89	99.39	99.40	99.86
99.76	99.76	99.95	99.95	99.96	99.67	99.69	99.95	99.65	99.67	99.92

Table 3.13
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=500, and for t=5

& from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.00	89.86	88.84	88.83	88.89	91.02	91.08	88.83	91.14	91.20	88.92
95.00	94.88	94.83	92.81	94.86	95.48	95.21	94.81	95.65	95.31	94.97
97.50	97.42	97.81	97.80	97.82	97.10	97.11	97.80	97.14	97.15	97.78
99.00	99.06	99.32	99.31	99.31	98.41	98.42	99.31	98.40	98.41	99.31
99.50	99.50	99.83	99.83	99.84	99.27	99.28	99.83	99.24	99.26	99.92
99.76	99.80	99.94	99.94	99.94	99.62	99.63	99.94	99.60	99.62	99.92

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.14
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=500, n_2=500, and for t=5

& from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.04	90.76	87.80	87.80	88.46	90.20	89.95	87.80	90.28	90.15	87.85
95.00	95.00	94.63	94.62	94.65	95,08	95.09	94.62	95.13	05.15	94.68
97.52	97.,44	97.83	97.82	97.84	97.12	97.13	97.82	97.14	97.16	97.85
99.00	99.00	99.40	99.40	99.40	98.51	98.51	99.40	98.51	98.50	99.39
99.50	99.68	99.77	99.77	99.77	99.13	99.13	99.77	99.11	99.11	99.76
99.76	99.86	99.87	99.87	99.87	99.38	99.81	99.87	99.37	99.36	99.86

Table 3.15
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the gram-Charlier and Edgeworth expansions for sample sizes c n_1=20, n_2=20, and for t=7

& from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
92.04	92.04	89.73	89.55	90.72	90.75	91.71	89.54	92.08	93.08	90.60
95.90	95.90	95.49	5.19	95.82	95.27	95.45	95.20	96.31	96.68	96.19
97.58	97.58	97.61	97.33	97.71	96.94	97.26	97.35	97.51	97.66	97.89
99.04	99.04	99.46	99.32	99.43	98.79	98.94	99.32	98.61	98.73	99.22
99.58	99.40	99.86	99.80	99.83	99.49	99.55	99.80	99.24	99.31	99.58
99.78	99.72	99.97	99.95	99.95	99.81	99.85	99.95	99.66	99.72	99.79

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.16
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes c n_1=20, n_2=100, and for t=7

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.72	90.72	89.11	89.04	89.61	90.37	90.86	89.03	91.04	91.48	89.52
95.20	95.68	94.69	94.56	94.90	94.82	95.09	94.56	95.37	95.61	95.04
97.54	97.54	97.72	97.60	97.77	97.17	97.31	97.60	97.41	97.52	97.85
99.04	99.20	99.28	99.20	99.22	98.63	98.65	99.21	98.58	98.59	99.20
99.56	99.56	99.77	99.73	99.75	99.33	99.24	99.73	99.21	99.24	99.63
99.82	99.82	99.92	99.91	99:92	99.70	99.70	99.91	99.61	99.60	99.83

Table 3.17
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=100, and for t=7

% from simul	Pearson ccurve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.52	90.96	88.76	88.73	88.99	90.03	90.40	88.73	90.28	90.68	88.93
95.02	95.42	94.35	94.29	94.43	94.65	94.76	94.29	94.92	95.01	94.52
97.58	97.58	97.79	97.74	97.81	97.28	97.34	97.74	97.38	97.43	97.85
99.02	99.02	99.32	99.29	99.31	98.71	98.72	99.29	98.67	98.69	98.91
99.52	99.54	99.75	99.73	99.75	99.30	99.33	99.73	99.24	99.28	99.69
99.80	99.82	99.92	99.91	99.91	99.66	99.67	99.91	99.61	99.62	99.87

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.18
Comparison of the upper percentiles specified by the simulation, the Pearson curves the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=200, and for t=7

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.06	90.56	88.28	88.26	88.46	89.80	89.95	88.26	90.00	90.15	88.41
95.06	95.44	94.44	94.37	94.52	94.91	94.83	94.38	95.19	95.01	94.63
97.50	97.56	97.60	96.61	97.62	96.18	97.19	97.48	96.44	97.27	97.75
99.00	98.96	99.41	99.39	99.41	98.81	98.84	99.39	98.78	98.81	99.38
99.50	99.50	99.78	99.77	99.78	99.35	99.37	99.77	99.31	99.33	99.74
99.76	99.78	99.92	99.91	99.92	99.67	99.68	99.91	99.63	99.65	99.88

Table 3.19
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=200, and for t=7

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.00	90.00	88.98	88.96	89.02	90.39	90.43	88.96	90.53	90.57	89.08
95.04	94.72	95.23	95.20	95.27	95.35	95.40	95.20	95.41	95.52	95.31
97.50	97.22	97.99	97.96	98.01	97.46	97.51	97.96	97.51	97.55	98.01
99.00	98.70	99.63	99.62	99.63	99.11	99.13	99.62	99.09	99.10	99.60
99.50	99.26	99.87	99.86	99.86	99.54	99.54	99.86	99.51	99.51	99.84
99.76	99.68	99.96	99.96	99.96	99.96	99.79	99.96	99.76	99.77	99.94

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.20
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=500, and for t=7

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/C	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.16	90.58	88.89	88.88	88.66	91.07	90.14	88.88	91.31	90.23	89.08
95.00	94.86	95.02	95.00	95.05	95.20	95.19	95.00	95.28	95.32	95.05
97.50	97.38	97.86	97.84	97.86	97.37	97.38	97.84	97.40	97.41	97.88
99.00	99.00	99.34	99.33	99.53	98.74	98.73	99.33	98.73	98.73	99.33
99.50	99.60	99.70	99.69	99.69	99.22	99.22	99.69	99.20	99.21	99.68
99.76	99.84	99.91	99.91	99.92	99.62	99.65	99.91	99.60	99.64	99.89

Table 3.21
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=500, n_2=500, and for t=7

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.04	90.90	88.06	88.05	88.03	89.63	89.64	88.05	89.68	89.70	88.09
95.08	95.32	94.44	94.43	94.46	94.77	94.79	94.43	94.82	94.84	94.47
97.50	97.36	97.86	97.85	97.87	97.37	97.39	97.85	97.39	97.41	97.87
99.00	99.02	99.30	99.29	99.30	98.69	98.70	99.29	98.68	98.70	99.29
99.50	99.62	99.69	99.69	99.70	99.21	99.22	99.69	99.20	99.21	99.68
99.76	99.84	99.87	99.87	99.87	99.54	99.54	99.87	99.53	99.53	99.86

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.22
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=20, n_2=20, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
91.22	91.22	88.53	88.35	89.78	89.39	90.65	88.34	90.50	91.81	89.21
95.02	96.06	93.57	93.23	94.16	93.61	93.39	93.24	94.71	95.41	94.20
97.68	97.68	97.71	97.38	97.79	97.07	97.43	97.40	97.50	97.75	97.86
99.10	98.82	99.58	99.42	99,50	99.06	99.18	99.43	98.87	98.17	99.31
99.54	99.54	99.84	99.76	99.81	99.51	99.59	99.76	99.30	99.38	99.57
99.76	99.76	99.95	99.91	99.93	99.78	99.82	99.91	99.61	99.68	99.75

Table 3.23
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=20, n_2=100, and for t=9

% from simul	Pearson curves	Normal approx	Gram-Charlier with 3 moments w/0 CC	w/cc	Gram-Charlier with 4 moments w/o CC	w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	w/CC	Edgeworth with 5 moments
90.70	91.77	88.37	88.29	88.98	89.38	90.02	88.29	89.89	90.56	88.69
95.02	95.76	93.93	93.77	94.19	94.10	94.45	93.78	94.61	94.94	94.21
97.52	92.88	97.16	97.00	97.23	96.77	96.96	97.01	97.04	97.20	97.28
99.06	99.00	99.33	99.24	97.32	98.82	98.82	99.24	98.76	98.84	99.23
99.50	99.50	99.77	99.72	99.75	99.42	99.45	99.64	99.31	99.34	99.72
99.76	99.76	99.93	99.91	99.92	99.74	99.76	99.91	99.66	99.68	99.83

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.24
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=100, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/C	Edgeworth with 5 moments
90.26	90.26	89.20	89.16	89.45	90.19	90.45	89.16	90.44	90.70	89.05
95.22	95.22	94.72	94.65	94.83	94.85	95.00	94.65	95.07	95.22	94.89
97.60	97.60	97.72	97.65	97.74	97.32	97.40	97.65	97.41	97.49	98.07
99.06	99.12	99.21	99.17	99.21	98.71	98.76	99.15	98.70	98.74	99.17
99.50	99.64	99.65	99.62	99.64	99.25	99.28	99.62	99.21	99.24	99.59
99.76	99.78	99.87	99.86	99.88	99.62	99.79	99.86	99.57	99.78	99.82

Table 3.25
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=100, n_2=200, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.08	89.86	89.53	89.50	89.70	90.50	90.61	89.50	90.68	90.85	89.64
95.22	94.98	95.15	95.10	95.21	95.22	95.31	95.10	95.37	95.46	95.23
97.52	97.36	97.83	97.78	97.83	97.43	97.48	97.78	97.50	97.54	97.85
99.00	99.00	99.24	99.26	99.28	98.86	98.83	99.26	98.80	98.82	99.26
99.50	99.50	99.73	99.71	99.72	99.38	99.40	99.71	99.35	99.36	99.69
99.76	99.82	99.89	99.88	99.88	99.65	99.66	99.88	99.62	99.63	99.85

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.26
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=200, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.04	90.04	89.14	89.12	89.27	90.12	90.30	89.12	90.18	88.75	89.06
95.16	94.92	95.05	95.01	95.17	95.16	95.31	95.01	95.27	95.41	95.11
97.52	97.38	97.79	97.76	97.80	97.41	97.45	97.76	97.46	97.49	97.81
99.00	98.88	99.45	99.43	99.44	99.00	99.02	99.43	08.99	99.00	99.42
99.52	99.36	99.84	99.83	99.84	99.57	99.58	99.83	99.54	99.55	99.81
99.76	99.64	99.95	99.95	99.95	99.80	99.84	99.95	99.79	99.82	99.93

Table 3.27
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=200, n_2=500, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.18	90.80	88.61	88.60	89.70	89.71	89.80	88.60	89.79	89.88	88.60
95.00	95.18	94.51	94.49	94.54	94.74	94.77	94.49	94.81	94.84	94.55
97.52	97.56	97.61	97.59	97.63	97.26	97.30	97.59	97.30	97.34	97.62
99.-2	99.26	98.99	98.97	98.99	98.51	98.53	98.97	98.51	98.53	98.98
99.50	98.70	99.54	99.53	99.55	99.13	99.14	99.53	99.11	99.13	99.52
99.76	99.85	99.80	99.79	99.80	99.50	99.51	99.79	99.48	99.50	99.78

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

Table 3.28
Comparison of the upper percentiles specified by the simulation, the Pearson curves, the normal approximation, and the Gram-Charlier and Edgeworth expansions for sample sizes n_1=500, n_2=500, and for t=9

% from simul	Pearson curve	Normal approx	Gram-Charlier with 3 moments w/0 CC	Gram-Charlier with 3 moments w/cc	Gram-Charlier with 4 moments w/o CC	Gram-Charlier with 4 moments w/CC	Edgeworth with 4 moments	Gram-Charlier with 5 moments w/0 CC	Gram-Charlier with 5 moments w/CC	Edgeworth with 5 moments
90.08	90.16	88.94	88.92	88.98	90.01	90.06	88.93	90.06	90.11	88.97
95.02	95.04	94.78	94.77	94.80	94.96	94.98	94.77	95.00	95.03	94.80
97.50	97.60	97.51	97.50	97.52	97.20	97.25	97.50	97.22	97.26	97.52
99.00	99.06	99.53	99.52	99.52	99.11	99.11	99.52	99.10	99.10	99.52
99.50	99.48	99.84	99.74	99.74	99.40	99.40	99.74	99.39	99.39	99.73
99.76	99.78	99.92	99.92	99.92	99.71	99.72	99.92	99.70	99.71	99.91

Note:
W/CC=with continuity correction
w/o CC=without continuity correction

approximation fits better when a continuity correction is applied or when it is not. When the first five moments were used for the Gram-Charlier and Edgeworth approximations, the adjustment was too extreme. When only the first three moments were used, the adjustment was inadequate, especially for the 99th, 99.5th, and 99.75th percentiles. The percentiles for the Edgeworth series with four moments were almost identical to those for the Gram-Charlier series with three moments, and hence, showed inadequate adjustment of the 99th, 99.5th, and 99.75th percentiles.

(vii) Recommendations for significance testing

Recall that in Chapter II it was recommended that a p-value of .0025 be considered to represent a significance level of α=.05 since we are testing for the occurrence of both concordance and clustering. Hence, the "best" approximate test of significance will be the test that most closely approximates the 99.75th percentile of the simulated distribution of X. The results presented in the preceeding two sections demonstrate that the Pearson curves fit the upper tail of the distribution of X more closely than do any of the other approximations tried. Unfortunately, the Pearson approach has several disadvantages. First, while the calculations are simpler than those involved for the Gram-Charlier or Edgeworth approximations, they are more complicated than for the normal approximation. Second, the (β_1,β_2) tables are not as readily available as the normal tables. Third, using the Pearson curves, we do not obtain a p-value for a specific value of X. The final disadvantage, and the most serious, is that the Pearson curve tables only go up to the 99.75th percentile, which is the smallest percentile we are willing to consider for the test. Thus, using the Pearson curve values,

we will only be able to assess significance at the $\alpha=.05$ level and not at smaller levels such as $\alpha=.01$ or $\alpha=.001$.

In view of the superiority of the Pearson curve values in approximating the upper tail of the distribution of X, the disadvantage of using them, and the advantage of using the normal approximation, we suggest a two-stage approach to assessing significance. First, calculate the standardized test statistic, Z, and use the standard normal tables to determine the cumulative percentile. If the percentile is smaller than 99.75, then we know, even without calculating the Pearson curve critical value, that the test is nonsignificant., Likewise, if the Z score is so large that it exceeds the range of the table, then we know, without calculating the Pearson curve value, that the test is significant. Calculation of the critical value in this case, while unnecessary, might be useful because we could see by how much the observed value of X exceeded the critical value. Finally, if the percentile is between the 99.75th and the 99.99th percentiles, the Pearson curve critical value would have to be calculated in order to determine significance. This two-stage approach would enable an investigator to avoid unnecessary and tiresome calculations.

Chapter IV

EPIDEMIOLOGIC APPLICATIONS

4.0 Introduction

To illustrate the utility of the test for concordant clustering, patterns of shigellosis morbidity in urban North Carolina counties and of cancer mortality in Cherokee county, N.C. and surrounding counties are examined. The use of graphics to aid in visualization of the disease patterns will be demonstrated. Some suggestions for other areas of application will be made.

4.1 Shigellosis morbidity

Patterns of shigellosis morbidity incidence for the six year period beginning in 1975 and ending in 1980 are compared for the eight most urban North Carolina counties. The analysis was restricted to urban counties because morbidity patterns in urban and rural areas usually differ. The data are presented in Table 4.1. Examination of the data revealed that in 1979, six of the eight counties experienced an increase in the number of shigellosis cases. We would like to determine both whether the increase is significant, and whether the disease patterns of the counties are similar. The null hypothesis is that, within each county, cases are randomly distributed over the time period and that each county's morbidity experience is independent of that of each of the other counties. The alternative hypothesis is that the shigellosis cases cluster in time and that the patterns of disease in the counties are

Table 4.1
Six years of shigellosis morbidity incidence data from the eight mos urban North Carolina counties

County	1975	1976	1977	1978	1979	1980	Total
Buncombe	1	1	1	8	342	0	353
Cumberland	7	0	12	7	5	3	34
Durham	1	1	2	4	4	2	14
Forsyth	2	5	2	8	89	8	114
Guilford	2	0	2	5	7	4	20
Mecklenburg	7	8	4	11	35	36	101
New Hanover	0	1	0	6	2	1	10
Wake	5	4	5	14	18	5	51

similar, i.e. there is a general epidemic pattern. These hypotheses will be tested by applying the test for concordant clustering to the data from each pair of counties. An understanding of the general disease pattern can be obtained by looking at the results of all of these pairwise tests. As discussed in Chapter II, clustering and concordance will be said to occur at significance level $\alpha=.05$ if the test has a p-value of 0.0025.

The values of X for each of the pairwise tests of concordant clustering are given in Table 4.2. As a first step in assessing significance, the standardized statistic, Z, was computed for each test, and compared with the standard normal curve. As can be seen in Table 4.3, a number of the Z-scores are so small that we can be certain that the corresponding tests are nonsignificant, and conversely, a number of the Z scores are so large that we can be certain that the corresponding tests are significant. The remaining Z scores are borderline, that is,

Table 5.2
Observed values of X and critical values for the 99.75th percentile of the Pearson curves(in parentheses) for each of the pairwise tests of concordant clustering of shigellosis morbidity in urban North Carolina counties

	Cumberland	Durham	Forsyth	Guilford	Mecklen-burg	New Hanover	Wake
Buncombe	1785 (2135)	1404* (909)	29827* (6951)	2438* (1279)	12077* (6172)	733* (660)	6282* (3164)
Cumberland		85 (106)	508 (722)	120 (145)	457 (644)	55 (79)	298 (340)
Durham			407* (315)	62 (67)	279 (282)	35 (38)	157* (152)
Forsyth				689* (438)	3553* (2051)	235* (231)	1758* (1062)
Guilford					816* (391)	48 (51)	236* (205)
Mecklenburg						180 (207)	1051* (928)
New Hanover							129 113

*p<.0025, according to the Pearson tables, and thus, the significan e level of the test is <.05.

Table 4.3
Values of the standardized test statistic for each of the pairwise tests of concordant clustering of shigellosis in urban Noerth Carolina counties

	Cumberland	Durham	Forsyth	Guilford	Mecklenburg	New Hanover	Wake
Buncombe	-5.27	22.15*	318.40*	40.27*	87.18*	6.53*	65.62*
Cumberland		0.70	-3.58	0.69	-5.27	-0.24	0.58
Durham			10.01*	2.45	3.07*	2.65	3.82*
Forsyth				18.15*	40.86*	3.58*	29.03*
Guilford					7.70*	2.78	5.55*
Mecklenburg						0.99	1.01
New Hanover							5.23*

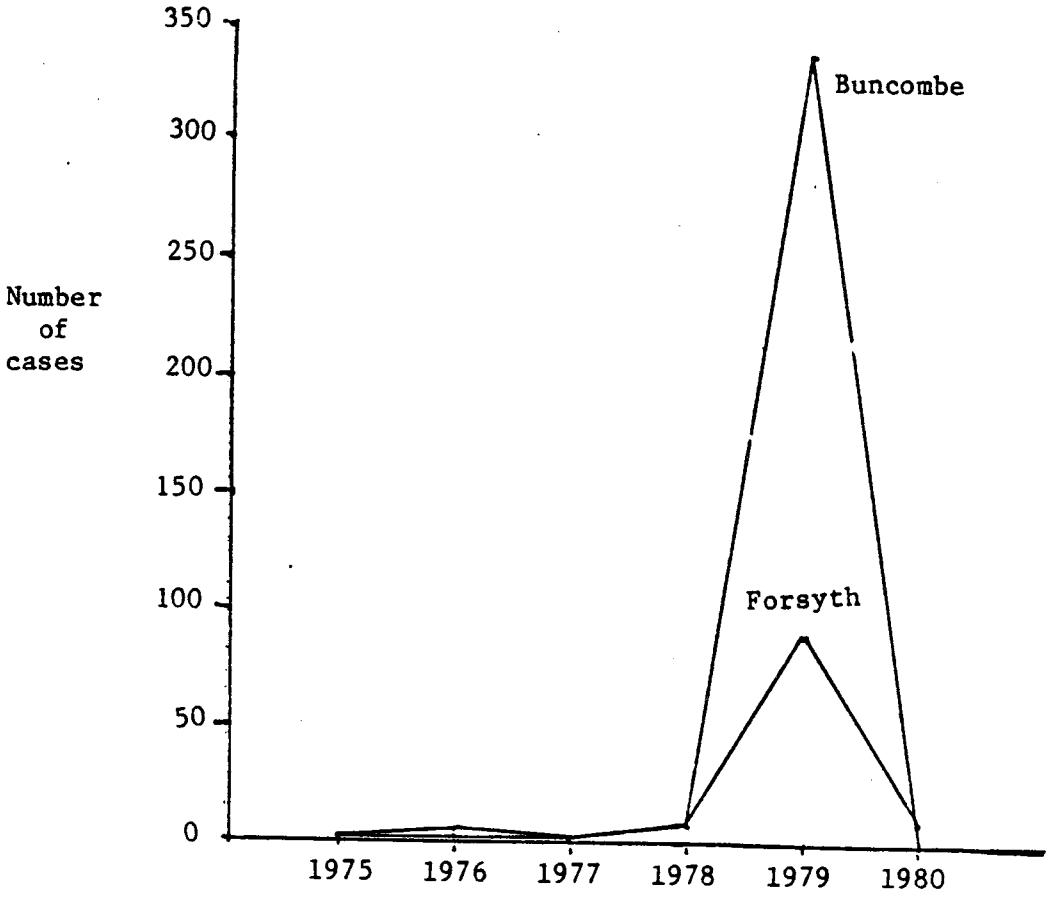
*p<.0025

they have p-values in the neighborhood of 0.0025, and the significance or nonsignificance of the corresponding test must be determined by calculating the critical value for the Pearson curve 99.75th percentile. These critical values are given in Table 4.2.

The most significant test result, as judged by the size of the Z score and by the percent difference bertween the observed value of X and the Pearson value, occurred for the Buncombe-Forsyth test. Both of these counties have extremely large counts in 1979 and relatively similar patterns, so the result is not surprising. A graph of the two series, Figure 4.1, corroborates the test result. Highly significant results as also were obtained for the Buncombe-Mecklenburg, Buncombe-Wake, Buncombe-Guilford, and Forsyth-Mecklenburg tests. Mecklenburg has a large cluster in 1979 that matches with Buncombe's cluster, and another large cluster in 1980 that contributes nothing to the test statistic, as it matches with a zero count in the Buncombe series. Significance is obtained because both series have extremely large clusters in 1979, and because the patterns for the first five years of the series are similar. Wake also has a cluster in 1979, and its second largest count in 1978. The significant result was expected because of the presence of clustering in both series and because of the general similarity of the patterns, especially the correspondence of the clusters and the second-largest counts. The significance of the result, however, is somewhat extreme, probably reflecting partial dominance of the statistic by Buncombe's cluster. Guilford's largest count occurred in 1979 and its second largest count in 1978, as did Buncombe's. Although by the EMM and scan test, Guilford's maximum count would not be considered quite large enough to be a cluster, as mentioned previously, we are willing to obtain

Figure 4.1

Shigellosis morbidity incidence in Buncombe and Forsyth counties during the years 1975 through 1980

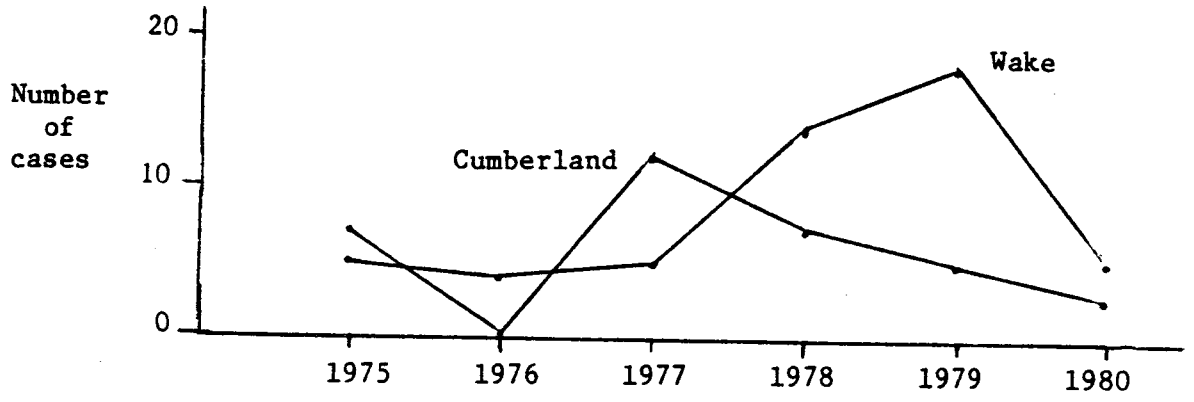


significant results when the cluster in one series is smaller than that specified by either of these two tests. As for the Buncombe-Wake test, the extreme significance of the Buncombe-Guilford test resulted because of dominance by the Buncombe cluster. The significant result obtained for the Forsyth-Mecklenburg test reflects the presence of clustering in both series and similar patterns. Both counties have a cluster of shigellosis cases in 1979. Mecklenburg also has a cluster in 1980 that is matched with Forsyth's second largest count.

Observe that the maximum number of cases in the Cumberland and New

Hanover series occur in 1977 and 1978, respectively, not in 1979 as do the maximums of the other series. None of the tests involving Cumberland county are significant. Even the Cumberland-Buncombe and Cumberland-Forsyth tests are nonsignificant, which means that despite the extreme size of the Buncombe and Forsyth clusters, the tests were not dominated by them. A graph of the Cumberland and Wake series, Figure 4.2, portrays the lack of concordance. Three of the tests involving New Hanover were significant, specifically, the tests between New Hanover

Figure 4.2
Shigellosis morbidity incidence in Cumberland and Wake counties during the years 1975 through 1980



and Buncombe, Forsyth, and Wake counties. For all three of these pairings, the cluster in the New Hanover series was matched with the second largest count in the other series and vice versa. Note that this is one of the accepted forms of concordance and that given sufficient clustering, a significant result should be obtained. As desired, these three tests were less significant than the tests in which the clusters in both series were matched.

The only other tests we will describe here are those between Durham and Buncombe, Forsyth, and Wake counties. The two largest counts in the Durham series, both 4's, are matched in all three of these tests with the

largest and next-largest counts in the other series. This explains why these tests are significant, whereas, the Durham-Mecklenburg test is not. The smallness of the "cluster" in the Durham series affected the p-values of the three tests, and they were much larger than for some of the other tests.

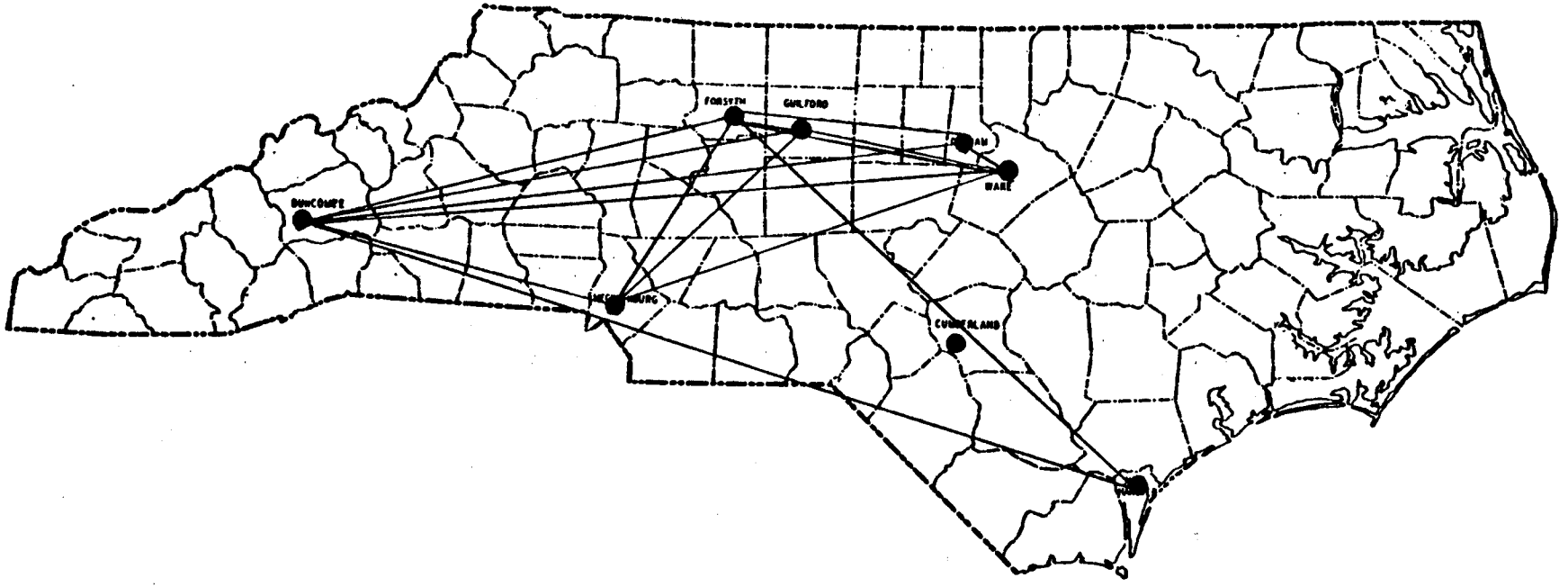
In Figure 4.3, we have linked pairs of counties found to display concordant clustering. Such a map can aid in interpreting and understanding the results. There appears to be a fanning out effect from Buncombe county eastwards. Observe that the two most significant tests are those between Buncombe county and the two counties nearest it, namely, Forsyth and Mecklenburg. No attempt will be made here to interpret the results since this analysis is intended only to illustrate the use of the test and the plausibility of the results.

4.2 Cancer mortality in western North Carolina

Residents of Cherokee county in western North Carolina have become concerned about an apparent increase in the number of deaths in their county that are due to cancer. They attribute the increase to the use of Tordon, an herbicide that, since 1967, has been widely used in the U.S. to defoliate road shoulders, powerline paths, railroad beds, and fence paths. Tordon will kill plants for at least three years after application, and is classified as a "restricted use" pesticide by the Environmental Protection Agency. Also known as Agent White, Tordon was used in Vietnam to kill vegetation that withstood Agent Orange.

Grimson(1972) applied the EMM and scan tests to total cancer mortality incidence data for the five year period 1976 through 1980. He established that Cherokee county and an adjacent county, Macon county, did experience a sharp increase in deaths due to cancer in 1979 and 1980.

Figure 4.3
North Carolina county map displaying the eight most urban counties. Pairs of counties with oncordant clustering are linked.



We wish to compare the total cancer mortality patterns in Cherokee county with those in adjacent counties to see if they are similar or if the increase in cancer deaths is unique to Cherokee. Thus, the null hypothesis is that, within each county, cancer deaths are distributed randomly over the time period, and that the mortality experiences of Cherokee county and each adjacent county are independent. The alternative hypothesis is that Cherokee county and the adjacent counties have parallel patterns of temporal clustering. The hypotheses will be tested by evaluating each of the tests between Cherokee county and the surrounding counties for concordant clustering. The tests will be considered significant at the $\alpha=.05$ level, if the p-value is less than or equal to 0.0025.

The counties adjoining Cherokee county consist of four North Carolina counties (Graham, Swain, Macon, and Clay), two Georgia counties (Union and Fannin), and two Tennessee counties (Polk and Monroe). See Figure 4.4. The total cancer mortality data are presented in Table 4.4.

Figure 4.4
Pairs of the eight most urban North Carolina counties that display concordant nonrandom patterns of shegellosis morbidity incidence are linked by line segments between their respective urban centers

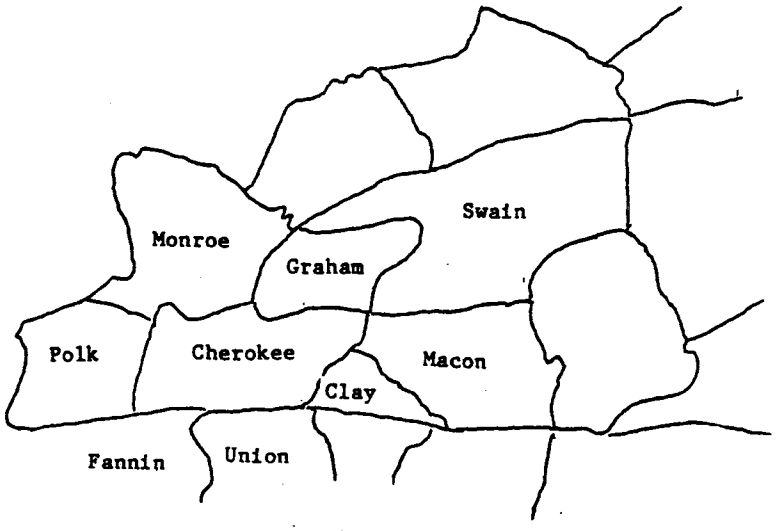


Table 4.4
Five years of total cancer mortality incidence data for Cherokee county, North Carolina and the adjacent counties

County	1976	1977	1978	1979	1980	Total
North Carolina						
Cherokee	27	27	29	45	46	174
Clay	11	12	13	10	8	54
Graham	9	13	14	10	13	59
Macon	36	32	36	52	58	214
Swain	14	19	16	14	16	79
Georgia						
Fannin	16	28	32	19	27	122
Union	14	16	15	24	17	86
Tennessee						
Monroe	31	41	49	36	48	205
Polk	24	22	24	19	35	124

The results of the pairwise tests involving Cherokee county and each of the adjoining counties are shown in Table 4.5. The standardized statistic, Z, was computed for all of the tests. As can be seen in Table 4.5, the Z scores for all of the tests, except the Cherokee-Macon test, were so small that it is not necessary to calculate the Pearson curve critical value to confirm nonsignificance. The Pearson curve critical value for the Cherokee-Macon test is smaller than the observed value of X, and thus, we concluded that the two counties display concordant clustering. Visual examination of the data supports these conclusions. See Figure 4.5. Both Cherokee and Macon counties experienced an increase in the number of cancer deaths in 1979 and 1980. Union county

Table 4.5
Results of pairwise tests of concordant clustering of cancer deaths in Cherokee county and adjoining counties

Cherokee versus	X	Z	Critical value for the Pearson curve 99.75 percentile
North Carolina			
Clay	1816	-1.63	----
Graham	2048	-0.13	----
Macon	7888*	5.71	7705
Swain	2721	-0.60	----
Georgia			
Fannin	4213	-0.56	----
Union	3107	2.33	----
Tennessee			
Monroe	7193	0.78	----
Polk	4403	1.49	----

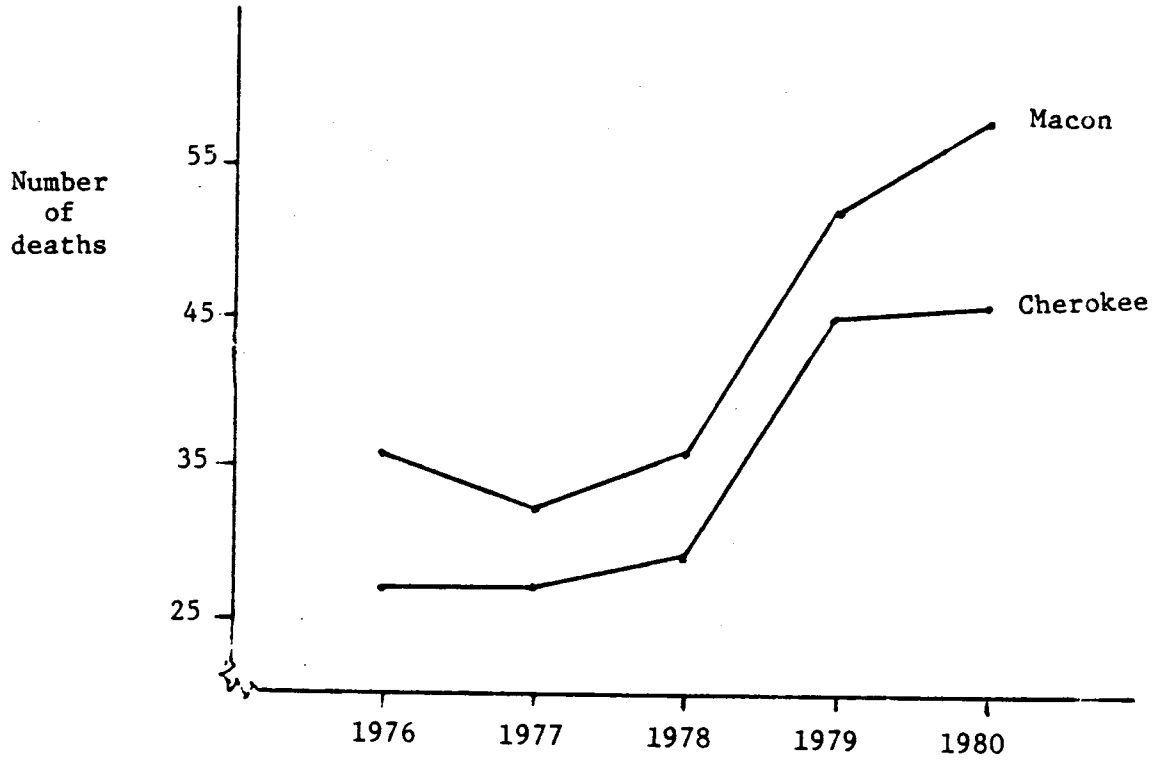
*$p < .0025$ according to the Pearson curve tables. Thus, this test has a significance level of $\alpha < .05$.

experienced a slight, but nonsignificant increase in 1977, while Polk county experienced a nonsignificant increase in 1980. The yearly number of deaths due to cancer in the other counties remained fairly stable over the time period studied. No interpretation of these results will be attempted here. However, we remark that cancer mortality in this area is being studied further. Ten years of data are now available, including data also from South Carolina. Several other counties have been found to have mortality patterns similar to those of Cherokee and Macon counties.

4.3 Other applications

As the two examples illustrate, the test can be readily applied to

Figure 4.5
Total cancer mortality incidence in Cherokee and Macon counties during the years 1976 through 1980



county morbidity and mortality incidence data. The test could also be useful in industrial settings. For example, an investigator might want to compare the health experience of workers with similar jobs at different plants, especially if a new manufacturing process had been introduced at some of the plants and not others. An investigator might also be interested in comparing the health experience of workers in different jobs at the same plant. Another application that comes readily mind is the use of the test to study crime data. The test might also be useful in the study of weather patterns. Doubtless, other areas for application could be found.

Chapter V

EXTENSION TO MORE THAN TWO SERIES AND FURTHER RESEARCH

5.0 Introduction

Several areas invite further research. One area is the extension of the test to the case where there are more than two, say r, series. of incidence data containing n_1, n_2,..., n_r cases, respectively. Such an extension would enable us to determine whether or not the r series share a similar nonrandom pattern. The r populations may be within a defined area, or scattered across a country or around the world. We present two possible extensions and some distributional findings. Note that in extending the test to more than two series it becomes possible to consider the spatial dimension when interpreting the test results. Another area for further research involves modification of the test to make it sensitive to time order. Preliminary work on this area is presented here.

5.1 One possible extension of the test to r series

The first r series statistic, X_r, is a direct extension of the pair-wise statistic; instead of summing over all pairs of cases, we sum over all r-tuples of cases. If a significant result is found we may surmise that the same underlying process is operating in the r populations, i.e. that they are part of a general epidemic or share some environmental factor in time. nonsignificant result indicates either that some of the

series do not share a concordant pattern or that some of the patterns are random or both. A nonsignificant result does not rule out the possibility that subsets of the r series may share a concordant nonrandom pattern. This first extension may be quite sensitive to deviations from a common pattern even if the deviations are slight and even if they occur only in a few of the series. However, no work has been done on this point. The appropriate test of significance has yet to be determined.

5.1.1 Description of the test

Consider r series of incidence data occurring in r locations over t time cells, and containing $n_1, n_2, \ldots, n_r$ cases. As for the two series case, under the null hypothesis of no concordance and no temporal clustering, the cases are mutually independent and are distributed randomly among the t time cells with probability 1/t of a case being in a given cell.

Let

$$X_{i_1 i_2 \ldots i_r} = \begin{cases} 1 & \text{if the } i_1\text{th case in Seies 1, the } i_2\text{th case in Series 2,..., and the } i_r\text{th case in Series r are in the same time cell} \\ 0 & \text{otherwise.} \end{cases}$$

The test statistic is

$$X_r = \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \cdots \sum_{i_r}^{n_r} X_{i_1 i_2 \ldots i_r} . \tag{5.1}$$

A second formula for X_r, preferable for computational purposes is

$$X_r = \sum_{h=1}^{t} a_{1h} a_{2h} \cdots a_{rh} \tag{5.2}$$

where a_{1h}, a_{2h} , ..., a_{rh} are the numbers of cases from series 1,2,...r falling into the hth time cell.

Thus, X_r is the number of r-tuples of cases from different series occurring in the same time cell, and assumes a value in the range of 0 to $N=n_1n_2...n_r$. When X_r is larger than expected, the r series can be said to have concordant patterns and to display some degree of temporal clustering. Here again, X_r is the sum of dependent Bernoulli random variables, and thus, is not a binomial variable.

5.1.2. The moments of X_r

The first three moments of this statistic have been derived using the same principles employed for the moments of X.

(i) The first moment of X_r

The expected value of X_r is,

$$E(X)=n_1n_2...n_r/t^{r-1} .$$

To see this, observe that $E(X_{i_1i_2\cdots i_r})= \frac{1}{t^{r-1}}$, and thus,

$$E(X_r) = E\left(\sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \cdots \sum_{i_r}^{n_r} X_{i_1 i_2 \ldots i_r} \right)$$

$$= \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \cdots \sum_{i_r}^{n_r} E(X_{i_1 i_2 \ldots i_r})$$

$$= \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \cdots \sum_{i_r}^{n_r} 1/t^{r-1}$$

$$= n_1n_2...n_r/t^{r-1} . \qquad (5.3)$$

(ii) The second moment of X_r

The expression for the second moment of X_r about zero can be partitioned into parts just as the second moment of X was. This yields

the following expression,

$$E(X_r^2) = \frac{n_1(2)n_2(2)\ldots n_r(2)}{t^{2(r-1)}} + \sum_{k=1}^{r-1} \frac{1}{t^{2(r-1)}} \sum^* n_{z_1} n_{z_2} \ldots n_{z_k}^{(2)} n_{z_{k+1}}^{(2)} \ldots n_{z_r}$$

$$+ \frac{n_1 n_2 \ldots n_r}{t^{r-1}} \quad , \tag{5.4}$$

where $\sum^*$ denotes summation over all combinations of the r series.

For the case where r=3, the two formulas produce

$$V(X_3) = \frac{n_1(2)n_2(2)n_3(2)}{t^4} + \frac{1}{t^4}(n_1 n_2(2) n_3(2) + n_2 n_1(2) n_3(2) + n_3 n_1(2) n_2(2))$$

$$+ \frac{1}{t^3}(n_1 n_2 n_3(2) + n_1 n_3 n_2(2) + n_2 n_3 n_1(2)) + \frac{1}{t^2}(n_1 n_2 n_3) \tag{5.5}$$

$$- \frac{1}{t^4}(n_1^2 n_2^2 n_3^2) \ .$$

(iii). The third moment of X_r

The third moment of X_r about zero has been derived by partitioning the sum into disjoint components. The resulting expansion is,

$$E(X_r^3) = \frac{1}{t^{3(r-1)}} \Bigg[n_1^{(3)} n_2^{(3)} \ldots n_r^{(3)} + \sum_{k=1}^{r-1} \sum^* n_{z_1} \ldots n_{z_k}^{(3)} n_{z_{k+1}}^{(3)} \ldots n_{z_r}$$

$$+ 3 \sum_{k=0}^{r-2} \sum_{m=1}^{r-k-1} \sum_{p=1}^{r-k-m} \sum^* n_{z_1} \ldots n_{z_k}^{(2)} n_{z_{k+1}}^{(2)} \ldots n_{z_{k+m+p}}^{(3)} n_{z_{k+m+p+1}}^{(3)} \ldots n_{z_r}$$

$$+ 3 \sum_{k=1}^{r-2} \sum_{m=1}^{r-k-1} \sum^* n_{z_1} \ldots n_{z_k}^{(2)} n_{z_{k+1}}^{(2)} \ldots n_{z_{k+m}}^{(3)} n_{z_{k+m+1}}^{(3)} \ldots n_{z_r}$$

$$+ \sum_{k=0}^{r-3} \sum_{m=1}^{r-k-2} \sum_{p=1}^{r-k-m-1} \sum_{q=1}^{r-k-m-p} \sum^* n_{z_1} \ldots n_{z_k}^{(2)} n_{z_{k+1}}^{(2)} \ldots n_{z_{k+m+p+q}}^{(3)} n_{z_{k+m+p+q+1}}^{(3)} \ldots n_{z_r} \Bigg]$$

$$+ \frac{1}{t^{2(r-1)}}\left[3n_1^{(2)}n_2^{(2)}\cdots n_r^{(2)} + \sum_{k=1}^{r-1}{}^{*} n_{z_1}\cdots n_{z_k}{}^{(2)} n_{z_{k+1}}\cdots n_{z_r}^{(2)}\right]$$

$$+ \frac{n_1 n_2 \cdots n_r}{t^{r-1}} \quad . \tag{5.6}$$

where $\sum^{*}$ denotes summation over all combinations of the r series.

hen r=3,

$$\begin{aligned}
E(X^3) = \frac{1}{t^6}[&n_1^{(3)}n_2^{(3)}n_3^{(3)} + 3n_1^{(2)}n_2^{(3)}n_3^{(3)} + 3n_1^{(3)}n_2^{(2)}n_3^{(3)} + \\
&3n_1^{(3)}n_2^{(3)}n_3^{(2)} + 9n_1^{(2)}n_2^{(2)}n_3^{(3)} + 9n_1^{(2)}n_2^{(3)}n_3^{(2)} + \\
&9n_1^{(3)}n_2^{(2)}n_3^{(2)} + n_1n_2^{(3)}n_3^{(3)} + n_1^{(3)}n_2n_3^{(3)} + n_1^{(3)}n_2^{(3)}n_3 + \\
&n_1n_2n_3^{(3)} + n_1n_2^{(3)}n_3 + n_1^{(3)}n_2n_3 + 24n_1^{(2)}n_2^{(2)}n_3^{(2)} + \\
&3n_1n_2^{(2)}n_3^{(3)} + 3n_1n_2^{(3)}n_3^{(2)} + 3n_1^{(2)}n_2n_3^{(3)} + 3n_1^{(3)}n_2n_3^{(2)} + \\
&3n_1^{(2)}n_2^{(3)}n_3 + 3n_1^{(3)}n_2^{(2)}n_3 + 6n_1n_2^{(2)}n_3^{(2)} + 6n_1^{(2)}n_2n_3^{(2)} + \\
&6n_1^{(2)}n_2^{(2)}n_3] \\
+ \frac{1}{t^4}[&3n_1^{(2)}n_2^{(2)}n_3^{(2)} + 3n_1n_2^{(2)}n_3^{(2)} + 3n_1^{(2)}n_2n_3^{(2)} + \\
&3n_1^{(2)}n_2^{(2)}n_3 + n_1n_2n_3^{(2)} + n_1n_2^{(2)}n_3 + n_1^{(2)}n_2n_3\ ] \\
+ \frac{1}{t^2}[&n_1n_2n_3] \quad .
\end{aligned}$$

The asymptotic distribution of X_r has not yet been determined and thus, the appropriate test of significance is not known. Monte Carlo simulations will need to be performed to determine the appropriate approximate significance test.

5.2. A second extension of the test to r series

The second r series statistic, X_p, is a pooled statistic. The statistic is formed by summing over all of the tests for concordant clustering between pairs of the r series.

clustering between pairs of the r series. This second extension may be less restrictive than the first in the sense that it may be possible to obtain a significant result when more than one pattern is present among the r series, provided the patterns are not too dissimilar and provided that most of the series share one common pattern. How dissimilar patterns may be and how many different patterns there may be and yet obtain a significant result has not been determined. The appropriate test of significance has not yet been determined.

5.2.1. Description of the test

Again, consider r series of incidence data occurring in r populations over t time cells, and containing n_1, n_2, ..., n_r cases. As for the two series case, under the null hypothesis of no clustering and no concordance, the cases are mutually independent and are distributed randomly among the t time cells with the probability of a case being in a given cell being 1/t. The pooled statistic is the sum of the pairwise statistics, that is,

$$X_p = \sum_{s=1}^{r} \sum_{u=s+1}^{r} X_{su} , \qquad (5.7)$$

where r is the total number of series, and X_{su} is the pairwise test statistic for series s and u that was defined in section 2.1. of Chapter 2.

5.2.2. The Moments of X_p

The first two moments of the statistic have been derived.

(i). The first moment of X_p

The expected value of X_p can be obtained easily.

$$E(X_p) = E\left(\sum_{s=1}^{r}\sum_{u=s+1}^{r} X_{su}\right) = \sum_{s=1}^{r}\sum_{u=s+1}^{r} E(X_{su})$$

$$= \sum_{s=1}^{r}\sum_{u=s+1}^{r} \frac{Ns_u}{t} \qquad (5.8)$$

where $N_{su} = n_s n_u$ and n_s and n_u are the population sizes of the sth and uth populations.

(ii). The second moment of Xp

The second moment of X_p about zero,

$$E(X_p^2) = E\left(\sum_{s=1}^{r}\sum_{u=s+1}^{r}\sum_{v=1}^{r}\sum_{w=v+1}^{r} X_{su}X_{vw}\right),$$

can be partitioned into four special cases. The first case, in which $s \neq u \neq v \neq w$, can be evaluated easily since the series are mutually independent. Thus,

$$E(X_{su}X_{vw}) = E(X_{su})E(X_{vw}) = \frac{N_{su}}{t}\frac{N_{vw}}{t} . \qquad (5.9)$$

The second case, $s=v, s \neq u \neq w$, also can be evaluated easily because of the independence oi the series. Thus,

$$E(X_{su}X_{sw}) = \frac{N_{su}}{t}\frac{N_{sw}}{t} . \qquad (5.10)$$

The third case, $u=w$, $s \neq u \neq v$, can be solved like the second case. Thus,

$$E(X_{su}X_{vu}) = \frac{N_{su}}{t}\frac{N_{vu}}{t} \qquad (5.11)$$

To evaluate the fourth case, $s=v$, $u=w$, $s \neq u$, recall the identity,

$$Var(X) = E(X^2) - [E(X)]^2.$$

Then,

$$E(X_{su}^2) = \text{Var } X_{su} + [E(X_{su})]^2 = \frac{N_{su}}{t}\frac{t-1}{t} + \left(\frac{N_{su}}{t}\right)^2$$

$$= \frac{1}{t^2}\,[N_{su}(N_{su} + t-1)] \,. \qquad (5.12)$$

Thus,

$$E\Big(\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ s\neq u}}^{r}\sum_{v=1}^{r}\sum_{\substack{w=v+1\\ v\neq w}}^{r} X_{su}X_{vw}\Big) = 2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ u\neq v\neq w}}^{r}\sum_{v=s+1}^{r}\sum_{w=v+1}^{r} E(X_{su}X_{vu})$$

$$+ 2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ u\neq w}}^{r}\sum_{w=s+1}^{r} E(X_{su}X_{sw})$$

$$+ 2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ u\neq v}}^{r}\sum_{v=s+1}^{r} E(X_{su}X_{vu}) + \sum_{s=1}^{r}\sum_{u=s+1}^{r} E(X_{su}^2) \,.$$

$$= \frac{1}{t^2}\Big[2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ u\neq v\neq w}}^{r}\sum_{v=s+1}^{r}\sum_{w=v+1}^{r} N_{su}N_{vw} + 2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ w\neq w}}^{r}\sum_{w=s+1}^{r} N_{su}N_{sw}$$

$$+ 2\sum_{s=1}^{r}\sum_{\substack{u=s+1\\ u\neq v}}^{r}\sum_{v=s+1}^{r} N_{su}N_{vu} + \sum_{s=1}^{r}\sum_{u=s+1}^{r} N_{su}(N_{su} + t-1)\Big] \,. \qquad (5.13)$$

The variance of X_p is

$$\text{Var}(X_p) = \sum_{s=1}^{r}\sum_{u=s+1}^{r} N_{su}\,\frac{t-1}{t^2} \,. \qquad (5.14)$$

5.3 Effects of Permutations of the Time Order

Two alternative statistics to X, suggested by Ibrahim A. Salama in a private communication, which are not invariant to the time dimension,

were examined. Both of them are functions of cumulative sums of the data. As before, consider two series of incidence data, A and B, occurring over t time cells, containing n_1 and n_2 cases, respectively.

Let $w_h = \sum_{i=1}^{h} a_i$, $h=1,\ldots,t$

and

$z_h = \sum_{i=1}^{h} b_i$, $h=1,\ldots,t$

The two test statistics are based on $\underline{w}_h$ and $\underline{z}_h$ rather than on $\underline{a}_h$ and $\underline{b}_h$. The two statistics are

1) $X' = \sum_{h=1}^{t} |w_h - z_h|$ (6.76)

and

2) $X^* = \sum_{h=1}^{t} w_h z_h$. (6.87)

As examination of both of these statistics soon revealed, neither is satisfactory. While working with cumulative sums of the data had the desired effect of making the statistics sensitive to the time order, it also had some undesirable effects. The most important of these was that the cumulation resulted in unequal weighting of the time cells, the first cell being weighted most heavily and the last cell least heavily. This meant that the calculated value of the statistics for two series and for the reflection of the series were quite different, especially if $n_1 \neq n_2$. This effect is illustrated below:

t:	1	2	3	4
A:	10	5	5	5
B:	5	20	5	5

X' = 35
X* = 3075

t:	1	2	3	4
A':	5	5	5	10
B':	5	5	20	5

X' = 15
X* = 2250

The example above is a simple one and permutations of time cells 3 and 4 for A and B and time cells 1 and 2 for A' and B' will not affect X' or X*. However, in general, the counts in these cells would not be equal and the inconsistency demonstrated above will be compounded. Using $w_h'=w_h-n_1/2$ and $z_h'=z_h-n_2/2$, instead of w_h and z_h, did not make X' and X* invariant to time reversal.

Several other alternative statistics have been explored. None has been found that is sensitive to concordance and temporal clustering and also consistently sensitive to permutations of the time order. The possibility of developing a statistic that is sensitive to time order and combining it with X in some way is being explored. An alternative, perhaps simpler, that takes into account only the relative positions of the largest and next largest counts in each series also is being explored.

5.4 Summary

During the past twenty years, statisticians have developed a number of statistical tests to assist in the detection of clusters of disease, or other events, in space, in time, and in space and time. We have proposed and developed another disease clustering test. This new test is sensitive to concordant epidemic patterns in two series of incidence data from different locations, that is, it is significant when the morbidity or mortality experiences of two locations during a specified time period are parallel and show evidence of clustering in time. Thus, this test is sensitive to the same form of clustering as Goldstein and Cuzick's test (1978), but has the advantage of testing simultaneously for concordance and nonrandomness. The test also has the advantage of being intuitively appealing, and of being simple and easy to calculate. Like most disease

clustering tests, the proposed method requires no control or comparison geroup, and involves minimal assumptions. Unlike the two most commonly used temporal clustering tests, the method makes use of all of the data, not just the maximum counts. The test uses actual counts, and therefore, may be applied to existing morbidity and mortality data. The distributional properties of the test were developed and the appropriate approximate test of significance sought. In addition two possible extensions of the test to r series and some of their distributional properties were presented.

List of References

Abe G (1973): A note on the methodology of Knox's tests of "time and space interaction". Biometrics 29: 67-77

Babbott JG and Ingalls (1961) TH: Tracheoesophageal fistula occurring in Pennsylvania. Quarterly Review of Pediatrics 16: 86-92

Bailar III, JC, Eisenberg H, and Mantel N (19700): Time between pairs of leukemia cases. Cancer 25: 1301-1303

Barton DE and Dennis KE (1952): The conditions under which Gram-Charlier and Edgeworth curves are positive definite and unimodal. Biometrika 30: 28-30

Barton DE, and David FN (1962a): The analyses of chromosome patterns in the normal cell. Annals of Human Genetics 25: 323-329

Barton DE, and David FN (1962b): Randomization basis for multavariate tests. I. The bivariate case: randomness of n points in a plane. Bulletin of the International Statistics Institute i39 II: 455-467

Barton DE, and David FN (1966): The random intersection of two graphs. in: FN David, (ed.),Research Papers in Statistics, pp. 45-59, New York: John Wiley & Sons Inc., 1966

Barton DE, David FN, and Merrington M (1965): A criterion for testing contagion in time and space. Annals Human Genetics Lund 29: 97-103

Bhapkar VP (1961): A nonparametric test for the problem of several samples. Annals of Statistics 32: 1106-1117

Cramer, H(1946): Mathematical Methods of Statistics. Princeton University Press, Princeton

David FN and Barton DE (1966) Two space-time interaction tests for epidemicity. British Journal of Social Medicine 20: 44-48

Ederer F, Myers MH, and Mantel N(1964) A statistical problem in space and time. Do leukemia cases come in clusters? Biometrika 20: 626-638

Elderton WP and Johnson NL (1969): Systems of Frequency Curves. Cambridge University Press, Aberdeen

Fraser DAS (1957): Nonparametric Methods in Statistics. John Wiley & Sons, Inc., New York

Goldstein IF and Cuzick J (1978): Application of a time-space clustering methodology to the assessment of acute enviromental effects on respiratory illnesses. Unpublished manuscript

Grimson R. and Ingram DD (1982): A test for concordant epidemic patterns. (Submitted for publication)

Glass AG, Mantel N, Gunz FW, and Spears GFS (1971): Time-space clustering of childhood leukemia in New Zealand. Journal National Cancer Institute 47: 329-336

Gregory GG (1977): Large sample theory for U-statistics and tests of fit. Annals of Statistics 5: 110-123The

Hall p (1979): On the invariance principle for U-statistics. Stochastic Processesses and their Applications 9: 163-174

Hoeffding W (1948): A class of statistics with asymptotically normal distributions. Annals of Mathematical Statistics 19: 293-325

Johnson NL and Kotz, S (1969): Discrete Distributions. Houghton Miffin Company, Boston

Kendall M.G and Stuart A (1964): The advanced theory of statistics Vol I: distribution theory. Hafner Publishing Co. 3rd. ed. New Yor

Klauber MR (1971): Two sample randomization test for space-time clustering. Biometrics 27: 129-142

Klauber MR (1975) 380: Space-time clustering tests for more than two samples. Biometrics 31: 719-726

Klauber IR and Mustachi P (1970): Space-time clustering of childhood leukemia in San Francisco. Cancer Research 30: 1969-1973

Knox G (1959): Secular patterns in congenital oesophaghal atresia. British Journal of Preventive and Social Medicine 13: 222-226

Knox G (1964a): Detection of space-time interactions. Applied Statistics 13: 25-30

Knox, G (1971): Ep]demics of rare diseases. British Medical Bulletin 27: 43-47

Lehmann ee (1951): Consistency and unbiasedness of certain nonparametric tests. Annals of Mathematical Statistics 22: 165-179

Lloyd S and Roberts CJ (1973): A test for space clustering and its applications to congenital limb defects in Cardiff. British Journal of Preventive and Social Medicine 27: 186-191.

Lloyd S and Roberts CJ (1973): Space-time clustering of limb defects in Cardiff. British Journal of Social Medicine 77: 67

Mantel N (1967): The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209-220

Mantel N, Krysoto RJ, and Myers MH (1976): Tables and formulas for extended use of the Ederer-Myers-Mantel disease clustering procedure. American Journal of Epidemiology 104: 576-583

Naus JI (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 JI (1966a): Some probabilities, expectations, and variance for the size of the smallest intervals and largest clusters. Journal of the American Statistical Association 61:1191-1199

Naus JI(1966b): A power comparison of two tests of non-random clustering. Technometrics 8: 493-517

Neuhaus G (1977): Functional limit theorems for U-statistics in the degenerate case. Journal of Multivariate Analysis 7:424-439

Pike MC and Smith PG (1968): Disease clustering: a generalization of Knox's approach to the detection of space-time interactions. Biometrics 24: 541-546

Pike MG and Smith PG (1974): A case-control approach to examine diseases for evidence of contagion including diseases with long latent periods . Biometrics 30: 263-279

Pinkel D and Nefzger D (1959): Some epidemiological features of childhood leukemia in the Buffalo, NY area. Cancer 12: 351-357

Pinkel D, Dowd JE, and Bross IDJ (1963): Some epidemiological features of malignant solid tumors of children in the Buffalo, NY area. Cancer 16: 28-33

Puri ML and Sen PK (1971): Nonparametric methods in multivariate analysis John Wiley & Sons, Inc., New York

Sen PK (1960): On some convergence properties of U-statistics. Calcutta Statistical Association Bulletin 10: 1-18

Siemiatycki, J and McDonald AD (1972): Neural tube defects in Quebec: A search for evidence of clustering in time and space. British Journal of Preventive and Social Medicine 26: 40-44

Siemiatycki J (1978): Mantel's space-time clustering statistic: computing higher moments and a comparison of various data transforms. Journal of Statistical Computation and Simulation 7: 13-31

Smith PG and Pike MC (1974): A note on a "close pairs" test for space clustering. British Journal of Preventive and Social Medicine 28: 63-64

Stark CR and Mantel N (1967a): Temporal-spatial distribution of birth defects for Michigan children with leukemia. Cancer Research 27: 1744-1755

Stark CE and Mantel N (1967b): Lack of seasonal or temporal spatial clustering of Down's syndrome births in Michigan. American Journal of Epidemiology 86: 199-213

Wallenstein S (1980): A test for detection of clustering over time. American Journal of Epidemiology 111: 367-372

APPENDIX 1

Examples of the FORTRAN computer program used to simulate the distribution of X for t=5, 7, and 9

t=5

```
ISN 0002              INTEGER IR(200),K,NR,X,T1,T2,T3,T4,T5,
                      S1,S2,S3,S4,S5,XX(200)
ISN 0003              INTEGER T6,T7,S6,S7
ISN 0004              DOUBLE PRECISION DSEED
ISN 0005              N=0
ISN 0006              K=5
ISN 0007              DSEED=12347.0D0
ISN 0008              DO 9 J=1,5000,1
ISN 0009              T1=0
ISN 0010              T2=0
ISN 0011              T3=0
ISN 0012              T4=0
ISN 0013              T5=0
ISN 0014              NR=100
ISN 0015              CALL GGUD(DSEED,K,NR,IR)
ISN 0016              DO 1 I=1,100,1
ISN 0017              IF (IR(I)-2) 2,3,4
ISN 0018        2     T1=T1+1
ISN 0019              GO TO 1
ISN 0020        3     T2=T2+1
ISN 0021              GO TO 1
ISN 0022        4     IF (IR(I)-4) 5,6,7
ISN 0023        5     T3=T3+1
ISN 0024              GO TO 1
ISN 0025        6     T4=T4+1
ISN 0026              GO TO 1
ISN 0027        7     T5=T5+1
ISN 0028        1     CONTINUE
ISN 0029              S1=0
ISN 0030              S2=0
ISN 0031              S3=0
ISN 0032              S4=0
ISN 0033              S5=0
ISN 0034              NR=100
ISN 0035              CALL GGUD(DSEED,K,NR,IR)
ISN 0036              DO 11 I=1,100,1
ISN 0037              IF (IR(I)-2) 12,13,14
ISN 0038        12    S1=S1+1
ISN 0039              GO TO 11
ISN 0040        13    S2=S2+1
ISN 0041              GO TO 11
ISN 0042        14    IF (IR(I)-4) 15,16,17
ISN 0043        15    S3=S3+1
ISN 0044              GO TO 11
ISN 0045        16    S4=S4+1
ISN 0046              GO TO 11
```

t=5 (Contd.)

```
ISN 0047      17    S5=S5+1
ISN 0048      11    CONTINUE
ISN 0049            X=(T1*S1) + (T2*S2) + (T3*S3) + (T4*S4)
                    + (T5*S5)
ISN 0050            N=N+1
ISN 0051            XX(N)=X
ISN 0052            IF (N.LT.200) GO TO 9
ISN 0054            WRITE(11,100) XX
ISN 0055      100   FORMAT(20I6)
ISN 0056            N=0
ISN 0057      9     CONTINUE
ISN 0058            WRITE(3,101) DSEED
ISN 0059      101   FORMAT(3X,D15.6)
ISN 0060            STOP
ISN 0061            END
```

t=7

```
ISN 0002            INTEGER IR(100),K,NR,X,T1,T2,T3,T4,T5,
                   1S1,S2,S3,S4,S5,XX(200)
ISN 0003            INTEGER T6,T7,S6,S7
ISN 0004            DOUBLE PRECISION DSEED
ISN 0005            N=0
ISN 0006            K=7
ISN 0007            DSEED=12347.000
ISN 0008            DO 29 J=1,5000,1
ISN 0009            T1=0
ISN 0010            T2=0
ISN 0011            T3=0
ISN 0012            T4=0
ISN 0013            T5=0
ISN 0014            T6=0
ISN 0015            T7=0
ISN 0016            NR=100
ISN 0017            CALL GGUD(DSEED,K,NR,IR)
ISN 0018            DO 1 I=1,100,1
ISN 0019            IF (IR(I)-2) 2,3,4
ISN 0020       2    T1=T1+1
ISN 0021            GO TO 1
ISN 0022       3    T2=T2+1
ISN 0023            GO TO 1
ISN 0024       4    IF (IR(I)-4) 5,6,7
ISN 0025       5    T3=T3+1
ISN 0026            GO TO 1
ISN 0027       6    T4=T4+1
ISN 0028            GO TO 1
ISN 0029       7    IF (IR(I)-6) 8,9,10
ISN 0030       8    T5=T5+1
ISN 0031            GO TO 1
ISN 0032       9    T6=T6+1
ISN 0033            GO TO 1
ISN 0034       10   T7=T7+1
ISN 0035       1    CONTINUE
ISN 0036            S1=0
ISN 0037            S2=0
ISN 0038            S3=0
ISN 0039            S4=0
ISN 0040            S5=0
ISN 0041            S6=0
ISN 0042            S7=0
ISN 0043            NR=100
ISN 0044            CALL GGUD(DSEED,K,NR,IR)
ISN 0045            DO 11 I=1,100,1
ISN 0046            IF (IR(I)-2) 12,13,14
ISN 0047       12   S1=S1+1
ISN 0048            GO TO 11
ISN 0049       13   S2=S2+1
ISN 0050            GO TO 11
```

t=7 (Contd.)

```
ISN 0051      14    IF (IR(I)-4) 15,16,17
ISN 0052      15    S3=S3+1
ISN 0053            GO TO 11
ISN 0054      16    S4=S4+1
ISN 0055            GO TO 11
ISN 0056      17    IF (IR(I)-6) 18,19,20
ISN 0057      18    S5=S5+1
ISN 0058            GO TO 11
ISN 0059      19    S6=S6+1
ISN 0060            GO TO 11
ISN 0061      20    S7=S7+1
ISN 0062      11    CONTINUE
ISN 0063            X=(S1*T1) + (S2*T2) + (S3*T3) + (S4*T4)
                   1+ (S5*T5) + (S6*T6) + (S7*T7)
ISN 0064            N=N+1
ISN 0065            XX(N)=X
ISN 0066            IF (N.LT.200) GO TO 29
ISN 0068            WRITE(11,100) XX
ISN 0069      100   FORMAT(200I6)
ISN 0070            N=0
ISN 0071      29    CONTINUE
ISN 0072            WRITE(3,101) DSEED
ISN 0073      101   FORMAT(3X,D15.6)
ISN 0074            STOP
ISN 0075            END
```

t=9

```
ISN 0002            INTEGER IR(100),K,NR,X,T1,T2,T3,T4,T5,
                   1S1,S2,S3,S4,S5,XX(200)
ISN 0003            INTEGER S6,S7,S8,S9,T6,T7,T8,T9
ISN 0004            DOUBLE PRECISION DSEED
ISN 0005            N=0
ISN 0006            K=9
ISN 0007            DSEED=12347.0D0
ISN 0008            DO 29 J=1,5000,1
ISN 0009            T1=0
ISN 0010            T2=0
ISN 0011            T3=0
ISN 0012            T4=0
ISN 0013            T5=0
ISN 0014            T6=0
ISN 0015            T7=0
ISN 0016            T8=0
ISN 0017             T9=0
ISN 0018            NR=100
ISN 0019            CALL GGUD(DSEED,K,NR,IR)
ISN 0020            DO 1 I=1,100,1
ISN 0021            IF (IR(I)-2) 2,3,4
ISN 0022       2    T1=T1+1
ISN 0023            GO TO 1
ISN 0024       3    T2=T2+1
ISN 0025            GO TO 1
ISN 0026       4    IF (IR(I)-4) 5,6,7
ISN 0027       5    T3=T3+1
ISN 0028            GO TO 1
ISN 0029       6    T4=T4+1
ISN 0030            GO TO 1
ISN 0031       7    IF (IR(I)-6) 8,9,10
ISN 0032       8    T5=T5+1
ISN 0033            GO TO 1
ISN 0034       9    T6=T6+1
ISN 0035            GO TO 1
ISN 0036       10   IF (IR(I)-8) 30,31,32
ISN 0037       30   T7=T7+1
ISN 0038            GO TO 1
ISN 0039       31   T8=T8+1
ISN 0040            GO TO 1
ISN 0041       32   T9=T9+1
ISN 0042       1    CONTINUE
ISN 0043            S1=0
ISN 0044            S2=0
ISN 0045            S3=0
ISN 0046            S4=0
ISN 0047            S5=0
ISN 0048            S6=0
ISN 0049            S7=0
ISN 0050            S8=0
```

t=9 (Contd.)

```
ISN 0051              S9=0
ISN 0052              NR=100
ISN 0053              CALL GGUD(DSEED,K,NR,IR)
ISN 0054              DO 11 I=1,100,1
ISN 0055              IF (IR(I)-2) 12,13,14
ISN 0056        12    S1=S1+1
ISN 0057              GO TO 11
ISN 0058        13    S2=S2+1
ISN 0059              GO TO 11
ISN 0060        14    IF (IR(I)-4) 15,16,17
ISN 0061        15    S3=S3+1
ISN 0062              GO TO 11
ISN 0063        16    S4=S4+1
ISN 0064              GO TO 11
ISN 0065        17    IF (IR(I)-6) 18,19,20
ISN 0066        18    S5=S5+1
ISN 0067              GO TO 11
ISN 0068        19    S6=S6+1
ISN 0069              GO TO 11
ISN 0070        20    IF (IR(I)-8) 21,22,23
ISN 0071        21    S7=S7+1
ISN 0072              GO TO 11
ISN 0073        22    S8=S8+1
ISN 0074              GO TO 11
ISN 0075        23    S9=S9+1
ISN 0076        11    CONTINUE
ISN 0077              X=(S1*T1) + (S2*T2) + (S3*T3) + (S4*T4)
                     1+ (S5*T5) + (S6*T6) + (S7*T7) + (S8*T8)
                      + (S9*T9)
ISN 0078              N=N+1
ISN 0079              XX(N)=X
ISN 0080              IF (N.LT.200) GO TO 29
ISN 0082              WRITE(11,100) XX
ISN 0083        100   FORMAT(200I6)
ISN 0084              N=0
ISN 0085        29    CONTINUE
ISN 0086              WRITE(3,101) DSEED
ISN 0087        101   FORMAT(3X,D15.6)
ISN 0088              STOP
ISN 0089              END
```

APPENDIX 2

Tables of the mean, mode, median, and variance of the simulated distributions of X and of the exact mean and variance of the distribution of X

t=5

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
10	20	40	40	40.0	31.4	40	32
10	30	60	60	60.0	46.6	60	48
10	40	80	80	79.9	64.1	80	64
10	50	100	100	99.8	78.4	100	80
10	60	120	120	119.8	92.8	120	96
10	70	140	140	139.9	109.4	140	112
10	80	160	160	159.7	126.7	160	128
10	90	180	180	179.9	147.0	180	144
10	100	200	200	199.9	160.9	200	160
20	20	80	80	80.2	62.5	80	64
20	30	120	120	120.1	95.7	120	96
20	40	160	160	159.9	124.7	160	128
20	50	200	200	200.0	159.3	200	160
20	60	240	240	240.2	194.2	240	292
20	70	280	280	279.9	229.5	280	224
20	80	320	320	319.9	250.2	320	256
20	90	360	360	359.8	289.4	360	288
20	100	399	400	400.1	308.6	400	320
30	30	179	180	180.0	139.7	180	144
30	40	238	240	240.0	193.0	240	192
20	50	300	300	299.8	239.4	300	240
20	60	360	360	359.8	289.6	360	288
30	70	418	420	419.9	329.3	420	336
30	80	480	480	479.6	393.3	480	384
30	90	540	540	539.7	437.3	540	432
30	100	600	600	600.3	468.5	600	480
40	40	322	320	320.0	251.5	320	256
40	50	404	400	399.9	311.2	400	320
40	60	480	479	479.7	382.2	480	384
40	70	560	559	559.4	461.4	560	448
40	80	642	640	639.7	527.6	640	512
40	90	722	720	720.1	579.8	720	576
40	100	804/800	800	799.7	639.7	800	640
50	50	500	499	499.3	402.7	500	400
50	60	602	599	599.5	481.1	600	480
50	70	700	700	699.8	584.0	700	560
50	80	800	800	800.3	621.6	800	640
50	90	900	900	900.2	721.2	900	720
50	100	1000	1000	999.6	806.0	1000	800
60	60	720	720	720.1	579.8	720	576
60	70	836	840	840.0	667.4	840	672

t=5 (Condt.)

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
60	80	956	960	960.1	763.5	960	768
60	90	1080	1080	1079.8	849.7	1080	864
60	100	1200	1200	1200.1	968.2	1200	960
70	70	979	980	979.9	754.9	980	784
70	80	1118	1120	1119.9	886.5	1120	896
70	90	1258	1260	1260.0	1003.8	1260	1008
70	100	1400	1401	1400.4	1098.2	1400	1120
80	80	1280	1280	1280.0	1024.0	1280	1024
80	90	1444	1440	1439.2	1124.5	1440	1157
80	100	1600	1600	1600.3	1287.1	1600	1280
90	90	1618	1619	1620.3	1311.2	1620	1296
90	100	1800	1799	1798.6	1424.1	1800	1440
100	100	2000	1999	1999.3	1588.9	2000	1600
100	200	4020	3999	3999,5	3231.4	4000	3200
200	200	8027	8001	8000.2	6597.3	8000	6400
200	500	19988	19999	19996.9	16739.0	20000	16000
500	500	49985	49998	49996.7	39627.7	50000	40000

t=7

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
20	20	57	57	57.2	48.1	57.1	49.0
20	30	87	86	85.7	73.8	85.7	73.4
20	40	114	114	114.1	96.0	114.3	98.0
20	50	142	143	142.8	122.5	142.9	122.5
20	60	170	171	171.3	149.0	171.4	148.9
20	70	200	200	199.8	170.4	200.0	171.4
20	80	226	228	228.5	200.9	228.6	195.9
20	90	259	257	256.8	215.4	257.1	220.4
20	100	280	286	285.7	237.9	285.7	246.9
30	30	128	128	128.5	108.9	128.6	110.2
30	40	174	171	171.3	147.7	171.4	148.9
30	50	214	214	214.1	178.7	214.3	183.7
30	60	260	257	256.9	219.2	257.1	220.4
30	70	301	300	300.0	260.4	300.0	257.1
30	80	340/343	342	342.5	287.1	342.0	293.9
30	90	392	385	385.3	331.1	385.7	330.6
30	100	426	428	428.7	356.0	428.6	367.3
40	40	227	229	228.5	189.3	228.6	195.9
40	50	284	285	285.6	244.8	285.7	246.9
40	60	339	343	342.7	296.2	342.9	293.9
40	70	399	399	399.3	346.8	400.0	342.9

t=7 (Contd.)

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
40	80	460	457	456.6	397.7	457.1	391.8
40	90	511	514	514.1	444.0	514.3	440.8
40	100	573	571	571.1	498.6	571.4	489.8
50	50	355	357	356.8	302.2	357.1	306.1
50	60	430	428	428.0	363.7	428.6	367.3
50	70	497	499	499.5	433.8	500.0	428.6
50	80	571	571	571.4	487.4	571.4	489.8
50	90	642	643	642.9	541.9	642.9	551.0
50	100	709	713	713.9	592.2	714.3	612.2
60	60	514	514	514.1	440.6	514.3	440.8
60	70	606	600	599.8	521.6	600.0	514.3
60	80	686	686	685.5	586.4	685.7	587.8
60	90	772	771	771.3	661.9	771.4	661.2
60	100	858/861	857	857.4	716.7	857.1	734.7
70	70	700	700	699.7	585.7	700.0	600.0
70	80	799	799	799.8	680.9	800.0	685.7
70	90	902	900	900.2	771.4	900.0	771.4
70	100	1000	1000	1000.3	833.4	1000.0	857.1
80	80	918	915	914.6	791.1	914.3	783.7
80	90	1020	1028	1028.3	863.0	1028.6	881.6
80	100	1145	1143	1142.8	943.5	1142.9	979.6
90	90	1158/1164	1158	1157.2	985.0	1157.1	971.9
90	100	1286	1285	1285.1	1068.5	1285.7	1102.0
100	100	1426	1428	1428.0	1224.5	1428.6	1224.5
100	200	2851	2856	2856.1	2484.6	2857.1	2449.0
200	200	5693/5704	5713	5715.0	4867.7	5714.3	4898.0
200	500	14289	14284	14283.6	12082.3	14285.7	12244.9
500	500	35690/35725	35709	35709.8	30701.0	35714.3	30612.2

t=9

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
20	20	45	44	44.4	39.1	49.4	39.5
20	30	66	67	66.6	59.1	66.7	59.3
20	40	88	89	88.6	78.0	88.9	79.0
20	50	108	111	111.0	97.7	111.1	9878
20	60	137	133	133.3	118.8	133.3	118.5
20	70	143	155	155.2	137.5	155.6	138.3
20	80	177	177	177.7	158.9	177.8	158.0
20	90	199	200	199.8	174.9	200.0	176.8
20	100	220	222	221.7	198.7	222.2	197.5

t=9 (Contd.)

n_1	n_2	Simulation distribution				Exact	
		Mode	Median	Mean	Variance	Mean	Variance
30	30	103	100	99.9	86.3	100.0	88.9
30	40	132	133	133.0	142.0	133.3	118.5
30	50	164	166	166.4	175.6	166.7	148.1
30	60	197	200	199.7	204.5	200.0	177.8
30	70	234	233	233.3	235.0	233.3	207.4
30	80	266	266	266.4	266.1	266.7	237.0
30	90	300	299	299.5	266.1	300.0	266.7
30	100	340	232	333.6	294.6	333.3	296.3
40	40	178	177	177.6	159.8	177.8	158.0
40	50	225	222	221.9	195.2	222.2	197.5
40	60	267	267	266.5	235.1	266.7	237.0
40	70	310	311	310.7	279.0	311.1	276.5
40	80	354	355	355.2	321.5	355.6	316.0
40	90	445	444	444.1	355.7	444.4	395.1
40	100	398	400	400.3	374.1	400.0	355.6
50	50	281	278	277.6	241.3	277.8	246.9
50	60	335	333	333.0	307.9	333.3	296.3
50	70	389	288	388.5	350.2	388.9	345.7
50	80	448/452	444	444.6	394.7	444.4	395.1
50	90	503	500	500.9	434.3	500.0	444.4
50	100	555	555	555.3	485.7	555.6	493.8
60	60	398	400	399.7	359.1	400.0	355.6
60	70	463	467	466.7	413.6	466.7	414.8
60	80	535	533	533.1	462.1	533.3	474.1
60	90	603	600	599.8	526.8	600.0	533.3
60	100	662	667	667.0	571.3	666.7	592.6
70	70	545	544	544.0	461.4	544.4	484.0
70	80	623	622	622.0	545.3	622.2	553.1
70	90	700	700	700.2	603.4	700.0	622.2
70	100	782	778	787.5	684.5	777.8	691.4
80	80	711	711	711.3	617.2	711.1	632.1
80	90	796	800	799.9	684.5	800.0	711.1
80	100	890	889	888.1	767.2	888.9	790.1
90	90	897	900	900.0	780.1	900.0	800.0
90	100	997	1000	999.8	871.1	1000.0	888.9
100	100	1109	1110	1110.6	998.5	1111.1	987.7
100	200	2217	2221	2221.9	2029.0	2222.2	1975.3
200	200	4447	4445	4444.4	4108.1	4444.4	3950.6
200	500	11104	11110	11109.6	9538.5	11111.1	9876.5
500	500	27785	27778	27776.5	24910.4	27777.8	24691.4

APPENDIX 3

Cirtical values of X for given sample sizes and values of t.
t-5

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
10	10	27	27	28	31	33	34	---*	---	---	---	---	---	25	27	28	29	30	31
10	20	47	49	52	54	57	60	---	---	---	---	---	---	47	49	51	53	55	56
10	30	68	71	74	78	80	84	---	---	---	---	---	---	69	71	74	76	78	80
10	40	90	93	96	100	103	107	---	---	---	---	---	---	90	93	96	99	101	103
10	50	111	114	118	122	125	131	---	---	---	---	---	---	111	115	118	121	123	125
10	60	131	136	140	145	149	152	---	---	---	---	---	---	133	136	139	143	145	148
10	70	153	157	161	167	172	175	---	---	---	---	---	---	154	157	161	165	167	170
10	80	173	178	183	190	194	196	---	---	---	---	---	---	175	179	182	186	189	192
10	90	194	200	205	212	217	221	---	---	---	---	---	---	195	200	204	208	211	214
10	100	215	221	225	232	239	245	---	---	---	---	---	---	216	221	225	229	233	236
20	20	90	93	96	101	105	109	90	93	96	101	104	108	90	93	96	99	101	103
20	30	132	137	141	146	151	156	132	136	140	145	149	153	133	136	139	143	145	148
20	40	173	178	183	189	193	200	174	178	183	189	193	198	175	179	182	186	189	192
20	50	215	221	226	232	238	243	215	221	225	232	237	242	216	221	225	229	233	236
20	60	257	264	269	276	281	287	257	262	268	275	281	286	258	263	267	272	276	279
20	70	298	305	311	319	325	332	298	304	310	318	324	330	299	305	309	315	319	322
20	80	339	346	352	362	366	372	339	346	352	360	367	373	341	346	351	357	361	365
20	90	380	388	393	404	412	421	381	387	394	403	410	417	382	388	393	399	404	408
20	100	421	430	436	445	452	462	422	436	426	445	452	460	423	429	435	442	446	451
30	30	194	200	204	211	216	222	195	199	204	210	215	220	195	200	204	208	211	214
30	40	257	263	269	276	283	286	257	262	268	275	281	286	258	263	267	272	276	279
20	50	318	326	332	340	347	351	319	325	331	339	345	352	320	325	330	336	340	349
20	60	380	388	395	404	411	417	381	387	394	403	410	417	382	388	394	399	404	408
30	70	441	450	458	468	474	483	442	450	457	466	474	481	444	450	456	463	467	472
30	80	503	512	520	531	540	547	504	512	519	529	537	545	505	512	518	526	530	535
30	90	565	574	582	592	604	614	565	574	582	592	601	609	567	574	581	588	594	599
30	100	627	637	647	659	667	674	627	635	644	655	664	673	628	636	643	651	656	662
40	40	339	346	354	362	367	375	339	346	352	360	367	373	341	346	351	357	361	365
40	50	421	430	437	446	454	462	422	429	436	445	452	460	423	429	435	442	446	451
40	60	503	512	521	532	539	549	504	512	519	529	537	545	505	512	518	526	530	535
40	70	585	594	604	615	621	637	586	594	602	613	622	631	587	595	601	609	615	620
40	80	667	677	687	702	713	724	667	677	685	697	706	715	669	677	684	693	698	704
40	90	750	759	771	786	794	805	749	759	768	780	790	800	751	759	767	776	782	788
40	100	831	842	851	865	875	884	831	841	851	864	874	884	832	842	850	859	865	871

t=5 (Contd.)

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
50	50	524	532	540	552	557	587	524	534	540	550	558	567	526	533	539	547	552	557
50	60	626	635	645	657	667	675	627	635	644	655	664	673	628	636	643	651	656	662
50	70	728	739	751	764	774	790	729	738	747	760	769	779	730	739	746	755	761	767
50	80	830	841	852	869	878	889	831	841	851	864	874	884	832	842	850	859	965	871
50	90	932	943	954	971	983	997	933	943	954	968	978	989	934	944	953	962	969	976
50	100	1033	1046	1058	1074	1088	1102	1034	1046	1057	1071	1082	1094	1036	1047	1055	1066	1073	1080
60	60	749	760	770	783	795	806	749	759	768	780	790	800	751	759	767	776	782	787
60	70	871	882	892	909	923	934	871	882	892	905	916	926	873	883	891	900	907	913
60	80	993	1004	1016	1033	1043	1056	994	1005	1016	1030	1041	1052	996	1006	1014	1024	1031	1038
60	90	1115	1126	1140	1155	1168	1181	1116	1127	1139	1154	1166	1178	1118	1128	1138	1148	1156	1163
60	100	1237	1250	1264	1278	1291	1306	1238	1250	1262	1278	1290	1303	1240	1251	1261	1272	1280	1288
70	70	1013	1024	1035	1051	1064	1073	1014	1025	1036	1050	1062	1073	1016	1026	1035	1045	1052	1059
70	80	1155	1168	1181	1197	1208	1218	1156	1168	1180	1195	1207	1220	1158	1169	1179	1190	1197	1205
70	90	1298	1311	1324	1339	1356	1373	1298	1311	1324	1340	1352	1366	1301	1312	1322	1334	1342	1350
70	100	1440	1453	1467	1484	1497	1510	1441	1454	1467	1484	1498	1511	1443	1455	1466	1478	1486	1495
80	80	1319	1333	1344	1363	1377	1394	1319	1332	1344	1361	1373	1386	1321	1333	1343	1354	1362	1370
80	90	1479	1493	1506	1525	1537	1554	1481	1495	1508	1525	1539	1553	1484	1496	1507	1519	1527	1536
80	100	1642	1659	1675	1694	1708	1744	1643	1658	1672	1690	1704	1719	1646	1659	1670	1683	1692	1701
90	90	1663	1679	1693	1717	1734	1767	1664	1678	1692	1711	1725	1740	1666	1679	1691	1704	1713	1722
90	100	1844	1859	1876	1894	1905	1928	1846	1861	1876	1895	1911	1926	1849	1862	1875	1888	1897	1907
100	100	2048	2065	2080	2100	2117	2141	2048	2065	2080	2101	2116	2133	2051	2066	2078	2093	2103	2112
100	200	4069	4094	4118	4150	4168	4193	4069	4091	4113	4142	4164	4187	4073	4093	4111	4132	4146	4159
200	200	8097	8132	8168	8215	8244	8265	8097	8129	8160	8201	8233	8266	8103	8132	8157	8186	8206	8225
200	500	20154	20206	20255	20312	20371	20411	20153	20204	20253	20317	20368	20420	20162	20208	20248	20294	20326	20355
500	500	50233	50322	50404	50502	50567	50602	50242	50322	50399	50502	50581	50664	50256	50329	50392	50465	50515	50562

*Not calculated

t=7

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
20	20	66	69	71	75	78	81	66	69	71	75	77	80	66	69	71	73	75	77
20	30	96	100	102	106	110	114	96	100	103	107	110	114	97	100	103	106	108	110
20	40	126	130	133	138	142	146	127	131	134	139	143	146	127	131	134	137	140	142
20	50	156	161	165	171	174	178	157	161	165	171	175	179	157	161	165	169	171	174
20	60	186	191	195	200	205	209	186	191	196	202	206	211	187	191	195	200	203	206
20	70	216	221	226	232	238	240	216	222	226	233	238	242	217	222	226	230	234	237
20	80	246	251	257	263	268	276	246	252	257	264	269	274	247	252	256	261	265	268
20	90	275	281	286	293	301	306	275	282	287	294	300	305	276	282	286	292	296	295
20	100	305	311	317	324	330	336	305	312	317	325	330	336	306	311	316	322	326	330
30	30	141	145	149	154	158	161	---*	---	---	---	---	---	142	146	149	153	156	158
30	40	186	191	195	201	206	209	---	---	---	---	---	---	187	191	195	200	203	206
30	50	231	236	241	247	252	255	---	---	---	---	---	---	232	237	241	246	249	253
30	60	275	281	288	295	299	305	---	---	---	---	---	---	276	282	286	292	295	299
30	70	320	326	332	339	347	357	---	---	---	---	---	---	321	326	331	337	341	345
30	80	363	370	377	384	390	396	---	---	---	---	---	---	365	371	376	383	387	391
30	90	408	415	421	430	438	448	---	---	---	---	---	---	409	416	421	428	433	437
30	100	452	459	467	476	483	494	---	---	---	---	---	---	453	460	466	473	478	483
40	40	246	251	256	262	266	272	---	---	---	---	---	---	247	252	226	261	265	268
40	50	305	311	318	326	331	338	---	---	---	---	---	---	306	311	316	322	326	330
40	60	364	370	376	385	395	399	---	---	---	---	---	---	365	371	376	383	387	391
40	70	422	430	436	444	451	458	---	---	---	---	---	---	424	430	436	443	448	452
40	80	481	489	497	507	515	523	---	---	---	---	---	---	483	490	496	503	508	513
40	90	540	548	557	567	577	585	---	---	---	---	---	---	541	549	555	563	568	574
40	100	598	608	617	628	637	644	---	---	---	---	---	---	600	608	615	623	628	634
50	50	378	385	391	399	406	411	---	---	---	---	---	---	380	386	391	398	402	407
50	60	451	460	466	476	483	489	---	---	---	---	---	---	453	460	466	473	478	483
50	70	525	534	541	552	561	572	---	---	---	---	---	---	527	534	541	548	553	558
50	80	599	608	616	628	638	645	599	607	616	626	635	643	600	608	615	623	628	634
50	90	671	681	690	702	709	717	672	681	690	701	710	718	673	681	689	697	703	709
50	100	744	754	763	776	787	797	745	755	764	776	785	794	746	755	763	772	778	784
60	60	540	548	557	567	576	586	540	548	556	567	574	582	541	549	555	563	568	574
60	70	628	637	646	758	668	677	628	637	645	656	665	776	629	637	644	653	658	664
60	80	715	724	733	747	756	765	716	725	734	746	755	764	717	726	733	742	748	754
60	90	802	813	823	837	847	855	803	813	823	835	845	854	805	814	822	831	838	844
60	100	890	902	912	927	932	947	890	901	911	924	934	944	892	902	910	920	927	934

t=7

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
70	70	729	737	748	762	770	779	---	---	---	---	---	---	731	740	748	757	763	769
70	80	832	843	852	869	877	885	---	---	---	---	---	---	834	843	851	861	867	874
70	90	934	946	957	972	981	992	---	---	---	---	---	---	936	946	954	965	972	978
70	100	1035	1048	1058	1075	1084	1104	---	---	---	---	---	---	1038	1048	1057	1068	1075	1083
80	80	948	961	973	985	993	1002	---	---	---	---	---	---	950	960	969	979	986	993
80	90	1064	1075	1086	1104	1115	1128	---	---	---	---	---	---	1067	1077	1087	1098	1105	1112
80	100	1181	1193	1204	1222	1233	1246	---	---	---	---	---	---	1183	1194	1204	1216	1223	1231
90	90	1196	1208	1221	1236	1249	1260	---	---	---	---	---	---	1198	1209	1219	1230	1238	1246
90	100	1324	1338	1351	1366	1376	1392	---	---	---	---	---	---	1328	1340	1351	1363	1371	1379
100	100	1471	1484	1499	1515	1527	1539	1472	1486	1499	1515	1528	1541	1473	1486	1497	1510	1519	1527
100	200	2916	2936	2955	2982	2998	3014	2918	2938	2956	2980	2998	3016	2921	2939	2954	2972	2985	2996
200	200	5800	5831	5858	5902	5924	5947	5800	5828	5854	5888	5913	5939	5804	5829	5851	5877	5895	5911
200	500	14419	14468	14510	14560	14590	14631	14422	14465	14506	14560	14600	14641	14428	14468	14503	14543	14571	14597
200	500	35920	35993	36069	36144	36194	36240	35929	35998	36063	36147	36211	36275	35939	36002	36057	36121	36165	36206

*Not calculated

t=9

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
20	20	52	54	57	61	63	65	52	55	57	60	63	65	53	55	57	59	61	62
20	30	76	79	82	86	89	92	76	79	82	86	89	91	77	79	80	85	87	88
20	40	99	103	106	111	115	119	100	104	109	111	114	117	100	104	106	110	112	114
20	50	123	127	131	136	139	141	124	127	131	136	139	143	124	128	131	134	137	139
20	60	147	151	155	159	163	166	147	151	155	160	164	168	147	151	155	159	161	164
20	70	170	174	179	183	188	192	170	175	179	185	189	193	171	175	179	183	186	189
20	80	193	198	203	209	213	218	194	198	203	209	213	218	194	199	202	207	210	213
20	90	216	221	226	232	237	242	217	222	227	233	238	242	217	222	226	237	234	238
20	100	239	244	249	257	262	267	240	245	250	257	262	267	240	245	250	255	258	262
30	30	111	115	119	123	127	133	---*	---	---	---	---	---	112	116	119	122	124	127
30	40	147	151	155	160	164	168	---	---	---	---	---	---	147	151	155	159	161	164
30	50	181	186	190	196	199	202	---	---	---	---	---	---	182	187	191	195	198	201
30	60	216	222	226	232	237	242	---	---	---	---	---	---	217	222	226	232	234	238
30	70	251	256	262	269	274	279	---	---	---	---	---	---	252	257	262	267	270	274
30	80	285	292	297	306	311	315	---	---	---	---	---	---	286	292	297	303	306	310
30	90	319	326	332	341	348	354	---	---	---	---	---	---	321	327	332	338	342	346
30	100	355	363	369	377	385	389	---	---	---	---	---	---	355	362	367	373	378	382
40	40	193	198	203	208	212	216	---	---	---	---	---	---	194	199	202	207	210	213
40	50	239	245	250	257	263	267	---	---	---	---	---	---	240	245	250	255	258	262
40	60	285	291	297	305	311	316	---	---	---	---	---	---	286	292	297	303	306	310
40	70	332	338	345	352	357	365	---	---	---	---	---	---	332	339	344	350	354	358
40	80	377	385	391	400	407	416	---	---	---	---	---	---	378	385	390	397	401	406
40	90	423	432	440	447	454	466	---	---	---	---	---	---	424	431	437	444	449	453
40	100	469	477	485	492	498	508	---	---	---	---	---	--	470	477	483	491	496	501
50	50	297	304	308	315	320	327	---	---	---	---	---	---	298	304	309	314	318	322
50	60	354	363	369	377	381	388	---	---	---	---	---	---	355	362	367	373	378	382
50	70	412	419	426	435	439	446	---	---	---	---	---	---	413	419	425	432	437	441
50	80	470	478	485	494	502	512	469	477	484	494	500	507	470	477	483	491	496	501
50	90	526	534	542	551	559	568	526	534	541	551	558	567	527	535	541	549	554	560
50	100	583	591	600	612	618	625	583	592	600	610	618	626	584	592	599	607	613	618
60	60	424	431	438	447	454	464	422	431	438	446	453	460	424	431	437	444	449	453
60	70	491	501	509	519	526	530	492	500	507	517	524	531	493	500	506	514	519	524
60	80	560	568	576	587	596	605	560	569	577	587	595	602	561	569	576	584	589	595
60	90	628	638	647	659	668	675	629	638	646	657	665	673	630	638	645	654	659	665
60	100	697	706	715	726	732	740	697	706	715	727	735	744	698	707	714	723	729	735

t=9(Contd.)

n_1	n_2	Critical values from simulation						Critical values from Pearson curves						Critical values from normal approx					
		90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75	90	95	97.5	99	99.5	99.75
70	70	571	580	587	600	607	613	---	---	---	---	---	---	573	581	588	596	601	607
70	80	651	660	669	680	688	701	---	---	---	---	---	---	652	661	668	677	683	689
70	90	730	741	750	762	769	776	---	---	---	---	---	---	732	741	749	750	764	770
70	100	809	821	832	844	853	860	---	---	--	---	---	---	811	821	829	839	846	852
80	80	742	752	762	772	780	788	---	---	---	---	---	---	743	752	760	770	776	782
80	90	833	843	854	867	879	883	---	---	---	---	---	---	834	844	852	862	869	875
80	100	923	933	945	957	966	978	---	---	---	---	---	---	925	935	944	954	961	968
90	90	935	946	956	969	979	991	---	---	---	---	---	---	936	947	955	966	973	980
90	100	1036	1048	1058	1073	1082	1091	---	---	---	---	---	---	1030	1049	1058	1069	1077	1084
100	100	1150	1162	1174	1187	1196	1206	1150	1162	1174	1188	1199	1211	1151	1163	1173	1184	1192	1200
100	200	2278	2296	2312	2331	2346	2358	2279	2295	2311	2331	2347	2363	2299	2295	2309	2326	2337	2348
200	200	4522	4548	4571	4604	4630	4651	4522	4547	4570	4599	4619	4643	4525	4548	4568	4591	4606	4622
200	200	11231	11270	11308	11342	11371	11397	11234	11273	11309	11355	11389	11425	11239	11275	11306	11342	11367	11392
500	500	27970	28033	28086	28156	28217	28271	27972	28034	28091	28163	28216	28274	27979	28036	28086	28143	28183	28222

*Not calculated