5-1-2006

JMASM23: Cluster Analysis In Epidemiological Data (Matlab)

Andrés M. Alonso
Universidad Carlos III de Madrid, andres.alonso@uc3m.es

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
DOI: 10.22237/jmasm/1146457320
Available at: http://digitalcommons.wayne.edu/jmasm/vol5/iss1/23

This Algorithms and Code is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
Matlab functions for testing the existence of time, space and time-space clusters of disease occurrences are presented. The classical scan test, the Ederer, Myers and Mantel’s test, the Ohno, Aoki and Aoki’s test, and the Knox’s test are considered.

Key words: Time cluster, space cluster, time-space cluster, epidemiology, Monte Carlo.
or 6 months, many quotients \( r = \frac{t}{T} \) can be reduced to the fraction \( 1/L \) with \( L = 4, 6, 8, 12, 15 \) or the 24. If \( N \) is greater than 10 and smaller than 100, then tables in Wallenstein (1980) give the critical values of the distribution of \( n \).

Example 1: The following table shows the number of cases of trisomia and spontaneous abortion in the city of New York between July/1975 and June/1977 (see Bailar et al., 1970).

```
function p = ProbabilityOfScanTest(n, N, t, T, B)

% Inputs:
% -------
% n : Maximum number of cases observed in t periods.
% N : Number of cases observed in T periods.
% t : Epidemic duration time.
% T : Total observation time.
% B : Number of replications.
% Output:
% -------
% p : Probability of having a value bigger or equal to n.

% Cases are B independents replicas of a uniform distribution
% of N cases in T periods.
Cases = zeros(T, B);
for b = 1:B
    X = rand(N, 1);
    for ii = 1:N
        for tt = 1:T
            if ((tt-1)/T < X(ii, 1) & X(ii, 1) < tt/T)
                Cases(tt, b) = Cases(tt, b) + 1;
            end
        end
    end
end

% Calculating the scan statistics using the B generates replicas
% stored in variable Cases.
ScanStatistics = zeros(B, 1);
for b = 1:B
    for tt = 1:T-t+1
        if (ScanStatistics(b, 1) < sum(Cases(tt:tt+t-1, b)))
            ScanStatistics(b, 1) = sum(Cases(tt:tt+t-1, b));
        end
    end
end

% Estimating the probability of having a scan statistics bigger
% or equal to the observed value, n.
p = sum(ScanStatistics >= n)/B;
```

Figure 1. Matlab Function  \( p \)
Therefore, N = 62, T = 24 months and the epidemic duration is fixed to t=2 months. Then n=14 and Pr(14,62,1/12) can be calculated. The Matlab function in Figure 1 obtain the probability p = Pr(n, N, r) by a Monte Carlo simulation procedure. The results of the above function for the data in Example 1 is Pr(14,62,2,24)= 0.0113. It supports the conclusion of a time cluster.

Test of Ederer, Myers and Mantel

The period under study is divided in k disjoints intervals. Under the null hypothesis of no grouping, the n cases will have to be distributed uniformly in the k intervals. The test statistics, m, is the maximum number of cases in an interval. Mantel et al. (1976) calculated tables for the expectation and variance of m under the null hypothesis of no group and for selected values of k and n. In the following table, the approximated estimators of E(m) and Var(m) are shown when the number of cases is greater than 100 (see Mantel et al., 1976).

<table>
<thead>
<tr>
<th>Number of intervals, k</th>
<th>E(m)</th>
<th>Var(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>n/2 + 0.3989*n^{1/2}</td>
<td>0.09084*n</td>
</tr>
<tr>
<td>3</td>
<td>n/3 + 0.4886*n^{1/2}</td>
<td>0.07538*n</td>
</tr>
<tr>
<td>4</td>
<td>n/4 + 0.5147*n^{1/2}</td>
<td>0.06043*n</td>
</tr>
<tr>
<td>5</td>
<td>n/5 + 0.5201*n^{1/2}</td>
<td>0.04951*n</td>
</tr>
</tbody>
</table>

Example 2: Assume that the number of children with congenital malformations born in the same year is as follows: 1st trimester: 100 cases, 2nd trimester: 50 cases, 3rd trimester: 50 cases and 4th trimester: 70 cases. If k=4 and n=270, then one can use the estimators of the previous table: E(m)= 270/4+0.5147*√270 ≈ 75.95 and Var(m) =0.06043*270 ≈ 16.32. The following statistic is calculated,

\[
\chi^2 = \frac{(m - E(m))^2}{Var(m)} = \frac{(100 - 75.95)^2}{16.32} = 35.44,
\]

and it may be concluded that it exists a time cluster.
The Matlab function in Figure 2 obtains the estimators of $E(m)$ and $Var(m)$ by a Monte Carlo simulation procedure. The results of this function for the data in Example 2 is $E(m) = 76.07$ and $Var(m) = 17.52$.

Detection of Space Clusters

A space cluster is defined as a non-uniform distribution of the cases in the area under study relative to the distribution of the population under study. The presence of clusters suggests a possible environmental etiology. The simplest analysis of space cluster is the comparison of the incidence or the prevalence of a particular disease in different geopolitical areas.

Test of Ohno, Aoki and Aoki

The test proposed by Ohno et al. (1979) determines if the obtained geographic pattern is different from the expected geographic pattern under the assumption of a uniform random distribution of the cases in the area under study. The procedure is as follows:

1. Define $k > 2$ disjoint categories of the incidence rates.
2. Identify the adjacent geographic areas in a map of the area under study.

3. Count the number of concordant area pairs.

4. Calculate the expected number of concordant adjacent pairs for each category: Let be $N$ the number of areas and $N_i$ the number of areas in the $i$-th category, then the number of concordant pairs in category $i$ is $N_i(N_i-1)/2$. Let $A$ be the number of adjacent pairs of regions, then the expected number of adjacent pairs with the $i$-th category is

$$E(C_i) = \frac{A}{N(N-1)}N_i(N_i-1).$$

5. Calculate the expected number of concordant adjacent pairs:

$$E(C) = \sum_{i=1}^{k} E(C_i).$$

Finally a $\chi^2$ test statistics, $\chi^2 = \frac{(C - E(C))^2}{E(C)}$, is calculated.

**Example 3:** The mortality rates of vesicle and esophagus cancer in Japan (1967-71) is categorized according to the following criterion:

Category 1. Rate $\geq 140$ by 10000 inhabitants.
Category 2. $120 \leq$ Rate $\leq 139.9$ by 10000 inhabitants.
Category 3. $80 \leq$ Rate $\leq 119.9$ by 10000 inhabitants.
Category 4. $60 \leq$ Rate $\leq 79.9$ by 10000 inhabitants.
Category 5. Rate $\leq 60$ by inhabitants.

In 1970, Japan had $N = 1,123$ cities and towns, without counting the prefecture of Okinawa, with $A=2840$ adjacent pairs of regions. The number of regions by category was: $N_1 = 293$, $N_2 = 78$, $N_3 = 256$, $N_4 = 116$ and $N_5 = 380$. In the following table, the calculation required for Ohno, Aoki and Aoki’s test is presented.

<table>
<thead>
<tr>
<th>Concordant pairs</th>
<th>Observed, $C_i$</th>
<th>Expected, $E(C_i)$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td>201</td>
<td>192.84</td>
<td>0.35</td>
</tr>
<tr>
<td>(2,2)</td>
<td>17</td>
<td>13.54</td>
<td>0.89</td>
</tr>
<tr>
<td>(3,3)</td>
<td>170</td>
<td>147.14</td>
<td>3.55</td>
</tr>
<tr>
<td>(4,4)</td>
<td>25</td>
<td>30.07</td>
<td>0.85</td>
</tr>
<tr>
<td>(5,5)</td>
<td>315</td>
<td>324.61</td>
<td>0.28</td>
</tr>
<tr>
<td>Total</td>
<td>728</td>
<td>708.20</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Finally, $\chi^2=0.55$ and it is concluded that evidence does not exist for the geographic association of the vesicle and esophagus cancer in men for these years in Japan. The following Matlab function obtain the value of Ohno, Aoki and Aoki’s test statistics given $N$, $A$, $C$ and the $N_i$.
Detection of Space-Time Clusters

A space-time cluster is defined as a non-uniform distribution of the cases in space and time, simultaneously. In general, the test of space-time cluster of health events needs a more sophisticated elaboration because one needs to prove that if the cases are associated in space they are also significantly near in the time, and vice versa (see, e.g., Kleinbaum et al., 1982).

Test of Knox

The test proposed by Knox (1964) is used to determine if there exists a significant interaction between the sites and the moments of appearance of the disease. It divides the dimensions in space-time into two parts, for which the critical distance in space, $E$, and the critical distance in time, $T$, must be defined. In a contingency table, each pair of cases is classified in one of the following categories: (i) near only in space, (ii) near only in time, (iii) near in space-time, and (iv) distant both in space and in time. The procedure is as follows:

1. Let be $n$ the number of cases. For each case, one knows its position in the space and in the time, then there are $N = n(n-1)/2$ possible pairs of cases.
2. Determine the distances in space, $e$, and in time, $t$, for each pair of cases.
3. Classify the $N$ pairs according to the following criterion:

   (a) A pair is near in space if $e < E$.

   (b) A pair is near in time if $t < T$.

   (c) A pair is near in space-time if it fulfills (a) and (b), simultaneously.

   (d) When a pair satisfies neither (a) nor (b), then we say that it is not near nor in space nor in time.

function OAAtest = OhnoAokiAokiTest(N, A, Ni, C)
% Inputs:
% -------
% N : Total number of regions.
% A : Number of adjacent regions.
% Ni : Number of regions in the ith category (k x 1 vector).
% C : Observed number of concordant adjacent regions.
% Output:
% -------
% OOAAtest : Ohno, Aoki and Aoki test statistics.
% Numbers of categories.
k = length(Ni);

% Expected number of adjacent regions in the ith category.
ECi = A*Ni.*(Ni-1)/(N*(N-1));

% Expected number of concordant adjacent regions.
EC = sum(ECi);

% Ohno, Aoki and Aoki test statistics.
OAAtest = (C-EC)^2/EC;

Figure 3. Matlab Function OAAtest
4. Construct the following table:

<table>
<thead>
<tr>
<th></th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Near</td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>Near</td>
<td>$X$</td>
</tr>
<tr>
<td>Non-Near</td>
<td>$N_e - X$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_e$</td>
</tr>
</tbody>
</table>

where $N_e$ is the number of pairs near in the space, $N_t$ the near ones in the time, and $X$ the near pairs in space-time.

5. The test statistic is the observed number of pairs near in space-time, $X$. In Knox (1964) it is assumed that $X$ distributes as a Poisson, therefore,

$$ p = \Pr(X \geq x) = \sum_{i=x}^{N} \frac{e^{-\lambda} \lambda^i}{i!}, $$

where $\lambda = \frac{N_e N_t}{N}$.

Example 4: The following table shows the results of the method of Knox for 5 cases of meningococcal disease in a territory given in a period of one year, it takes like critical distance in space 500 meters and in time 5 days.

<table>
<thead>
<tr>
<th></th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Near</td>
</tr>
<tr>
<td>Time</td>
<td></td>
</tr>
<tr>
<td>Near</td>
<td>$X=4$</td>
</tr>
<tr>
<td>Non-Near</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>$N_e=5$</td>
</tr>
</tbody>
</table>

Therefore, $\lambda = \frac{5 \times 4}{10} = 2$ and $\Pr(X \geq 4) = 0.142$. The Matlab function in Figure 4 obtains the value of above $p$-value given $X, N_e, N_t$, and $N$.

function pktest = KnoxTest(X, Ne, Nt, N)
    lambda = Ne*Nt/N;
    pktest = 0;
    for i = X:N
        pktest = pktest + exp(-lambda)*lambda^i/factorial(i);
    end
end

Figure 4. Matlab Function pktest
References


