M statistic commands: interpoint distance distribution analysis with Stata

Pietro Tebaldi  
Universita’ Bocconi  
pietro.tebaldi@studbocconi.it

Marco Bonetti  
Universita’ Bocconi  
marco.bonetti@unibocconi.it

Marcello Pagano  
Harvard School of Public Health  
pagano@hsph.harvard.edu
Motivation: spatial distribution (1)

In many situations we are interested in answering the following

QUESTION: Is a certain phenomenon more (less) concentrated in a certain area?

Typical question of cluster analysis
Motivation: spatial distribution (2)

More general problem: **spatial distribution analysis**

**QUESTION:** Does the group of interest have a different **spatial distribution** than the population (null distribution)?

**QUESTION:** Does group 1 have a different **spatial distribution** than group 2?
Motivation: spatial distribution (3)


Tebaldi, Bonetti, Pagano. M statistic with Stata
Motivation: spatial distribution (4)

Leukemia Data
In Massachusetts

H0: do locations w/ OBS>EXP and OBS<EXP have the same SD?

Breast Cancer Data
In Massachusetts


Tebaldi, Bonetti, Pagano. M statistic with Stata
New commands

*Mstat* and *Mtest* are Stata routines that can be used to test $H_0$. Based on the Euclidean distance in bi-dimensional spaces.

Applications

Epidemiology, Sociology, Economics, Demography, etc... whenever the fact that two phenomena are equally distributed over a given area of interest is not obvious and relevant at the same time.

Extensions

• K-dimensional spaces
• Non-Euclidean metrics or general dissimilarity measures
• $H_0$: is group j distributed as the underlying null (population) distribution (1-sample M statistic)
Theory: Interpoint Distance Distribution (IDD)

The main statistic on which M is based is the Empirical (Cumulative) Density Function (ECDF) of the Interpoint Distance Distribution.

\[
D = \begin{bmatrix}
    0 & d_{1,2} & \ldots & d_{1,n} \\
    d_{2,1} & 0 & \ddots & \\
    \vdots & \ddots & \ddots & d_{n-1,n} \\
    d_{n,1} & \ldots & d_{n,n-1} & 0
\end{bmatrix}
\]
Interpoint Distance Distribution (2)

From \( n \) observations \( \rightarrow \) \( D \) is an \( nxn \) symmetric matrix w/ zero main diagonal. We calculate

\[
d_{i,j}, \ i \neq j
\]

\[
D = \begin{bmatrix}
0 & d_{1,2} & \cdots & d_{1,n} \\
d_{2,1} & 0 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \ddots \\
d_{n,1} & \cdots & d_{n,n-1} & 0
\end{bmatrix}
\]

\[d\] is a sample of

\[
\binom{n}{2} = \frac{n(n-1)}{2}
\]

DEPENDENT distances

Tebaldi, Bonetti, Pagano. M statistic with Stata
We use all the \( \binom{n}{2} = \frac{n(n-1)}{2} \) distances.

From a sample of \( n \) observations we get roughly \( n^2/2 \) distances: COMPLEX RELATION.

Others use different, less informative statistics: distance to the nearest neighbor (or \( k<n-1 \) neighbors), average distance, etc...

Forsberg et al. [4] show that using all distances is more powerful.
Interpoint Distance Distribution (4)

We build the **Empirical Density Function** $f(d)$ for the IDD: a statistic we use to collect information on the (unknown) spatial distribution.

Tebaldi, Bonetti, Pagano. M statistic with Stata
The M statistic

Based on a discretized version of the ECDF:

- vector of cutoffs values of the BINS, whose number affects the power of our test [3].

With k bins:

$$\hat{F}(d) = \left[ \hat{F}(d_1), ..., \hat{F}(d_k) \right]$$

where

$$\hat{F}(d_\ell) = \binom{n}{2}^{-1} \sum_{i \neq j} 1\{d_{i,j} \leq d_\ell\}$$
The M statistic (2)

1-sample M statistic: one group vs population

\[ H_0: \ F(d) = F^0(d) \]

\( \hat{F}(d) \) being the observed ECDF:

\[ M = \left( \hat{F}(d) - F^0(d) \right)^T \Sigma^{-} \left( \hat{F}(d) - F^0(d) \right) \]

\( \Sigma^{-} \) is the Moore-Penrose (Mata \( \text{pinv}(() \)\)) generalized inverse of the variance covariance matrix of \( \hat{F}(d), \Sigma \).
The M statistic (3)

$$\Sigma = \begin{bmatrix} \sigma_{\ell,m} \end{bmatrix}, \text{ with}$$

$$\sigma_{\ell,m} = 4\left(\frac{n}{3}\right)^{-1} \sum_{i<j,k} 1\{d_{i,j} \leq d_\ell\} 1\{d_{i,k} \leq d_m\} - \hat{F}(d_\ell)\hat{F}(d_m)$$

Bonetti and Pagano (2005) show that

$$M \xrightarrow{d} \chi_k^2$$

Slow convergence $\iff$ Monte Carlo test.
The M statistic (4)

2-samples M statistic: Group 1 vs Group 2

\[ H_0: \quad F_1(d) = F_2(d) \]

\[
M = \left( \hat{F}_1(d) - \hat{F}_2(d) \right)^T \Sigma^- \left( \hat{F}_1(d) - \hat{F}_2(d) \right)
\]

Equal variance assumption:

\[
\sigma_{\ell,m} = \left( \frac{n_1 + n_2}{n_1 n_2} \right) \frac{n}{3} \sum_{i<j,k} 1\{d_{i,j} \leq d_{\ell}\} 1\{d_{i,k} \leq d_{m}\}
\]

Tebaldi, Bonetti, Pagano. M statistic with Stata
2-samples M Test

Kernel Densities

10 observations

45 distances

Tebaldi, Bonetti, Pagano. M statistic with Stata
Mstat and Mtest algorithm

**Mstat:**
(1) Generate the distance matrix $D$;
(2) Generate the cutoff vector $d$ so to have EQUIPROBABLE BINS (wrt the population);
(3) Compute the ECDF in Group 1 and 2 at $d$;
(4) Compute the matrix $\sum$, take its (generalized) inverse and compute $M$.

**Mtest:**
(A) Execute steps (1)-(4) of Mstat algorithm;
(B) Permute the Group indicator variable (dummy 0-1);
(C) Execute step (3) of Mstat algorithm;
(D) Using $d$ and $\sum$ from step (2) and (4) of Mstat algorithm compute $M$;
(E) Iterate (A)-(D) $P$ times, generating a vector $[M,M_1,M_2,...,M_P]$;
(F) Compute the Monte Carlo p-value= $(#M_i\geq M)/P$, and its exact binomial confidence interval.
Mstat and Mtest commands

DATASET: must contain three variables
- x-coordinates
- y-coordinates
- Group dummy variable

Syntax:

Mstat, x(varname) y(varname) g(varname) bins(#) scatter density chi2

Mtest, x(varname) y(varname) g(varname) bins(#) sc den iter(#) level(#)

Tebaldi, Bonetti, Pagano. M statistic with Stata
Mstat and Mtest commands (2)

```
.Mtest, x(X) y(Y) g(CASE) iter(1000) scatter density
```

M statistic
Monte Carlo permutation results
H0: The two groups have the same spatial distribution
Number of bins = 20
Number of permutations = 1000

<table>
<thead>
<tr>
<th></th>
<th>T(obs)</th>
<th>c</th>
<th>n</th>
<th>p=c/n</th>
<th>SE(p) [95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>79.63794</td>
<td>23</td>
<td>1000</td>
<td>0.0230</td>
<td>0.0047 .0146346 .0343123</td>
</tr>
</tbody>
</table>

Note: confidence interval is with respect to p=c/n.
Note: c = #{T >= T(obs)}
Mstat options

\[ x*(\text{varname}) \quad \text{x-coordinates} \]
\[ y*(\text{varname}) \quad \text{y-coordinates} \]
\[ g*(\text{varname}) \quad 0-1 \text{ dummy} \]

\[ \text{bins(#)} \quad \text{number of bins} \]
\[ \text{scatter} \quad \text{scatter plot} \]
\[ \text{density} \quad \text{Kernel density} \]
\[ \text{chi2} \quad \text{asymptotic Chi2 pvalue} \]
## Mtest options

- **x*(varname)**: x-coordinates
- **y*(varname)**: y-coordinates
- **g*(varname)**: 0-1 dummy
- **bins(#)**: number of bins
- **scatter**: scatter plot
- **density**: Kernel density
- **iter(#)**: # of permutations
- **level(#)**: conf level for pvalue C.I.
Cancer Data in Massachusetts

Datasets are fully compatible with Pisati’s *spmap*

**Leukemia Data**

From MA cancer report, we have 348 locations (census tracts) with

**EXPECTED EVENTS**

**OBSERVED EVENTS**

We build the dummy

Group1: EXP<OBS

Group2: OBS≤EXP

Plot with *spmap*

Run *Mtest*

---

Tebaldi, Bonetti, Pagano. M statistic with Stata
Cancer Data in Massachusetts (2)

Breast Cancer Data

Massachusetts Breast Cancer Data

<table>
<thead>
<tr>
<th>T</th>
<th>T(obs)</th>
<th>c</th>
<th>n</th>
<th>p=c/n</th>
<th>SE(p)</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>38.56281</td>
<td>4</td>
<td>1000</td>
<td>0.0040</td>
<td>0.0020</td>
<td>.0010909</td>
</tr>
</tbody>
</table>

Note: confidence interval is with respect to p=c/n.
Note: c = #(T >= T(obs))

Tebaldi, Bonetti, Pagano. M statistic with Stata
Future Developments

• Two-samples M with general, non-Euclidean dissimilarity. Possible for the user to input the matrix D.

• One-sample M. The user need to be familiar with the underlying theory: specification of the null distribution critical.
Acknowledgments

• Harvard University, School of Public Health, Biostatistics Department.
• Universita’ Commerciale Luigi Bocconi, Dipartimento di Scienze delle Decisioni.
• Research supported in part by a grant from the NIH, P01 CA 134294.
• A special thanks to Justin Manjourides, PhD and Al Ozonoff, PhD for the helpful insights and suggestions during my Summer 2010 visit at HSPH.
Selected References


