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**

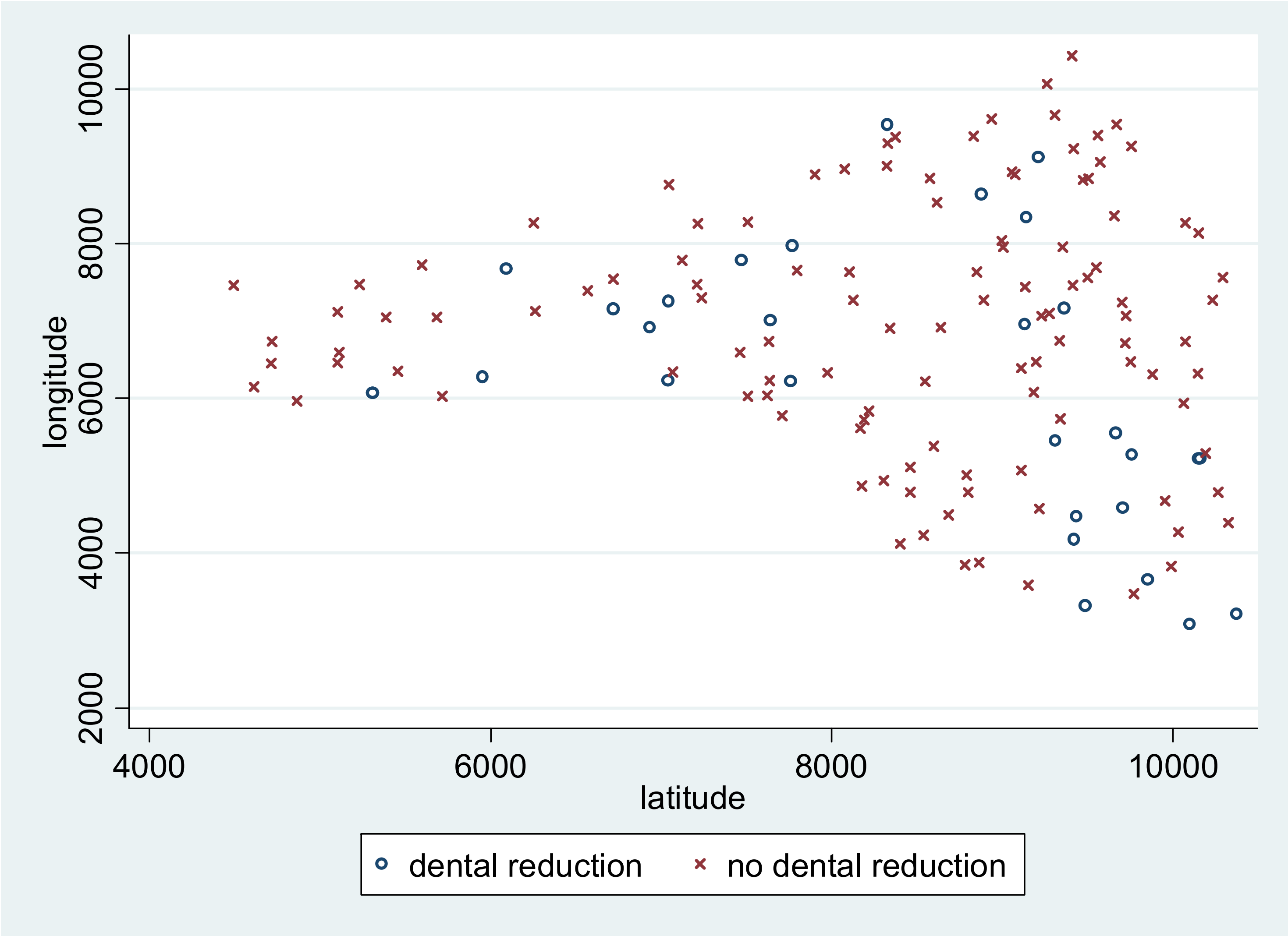
Motivation: spatial distribution (2)

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?*

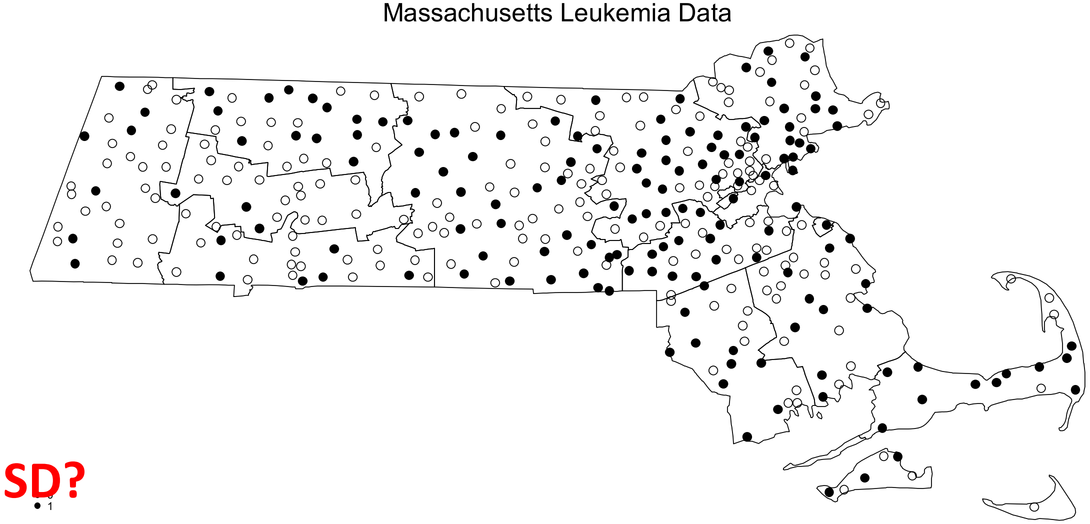
Motivation: spatial distribution (3)



Source: Alt and Vach Data, Waller, L. and Gotway, C., Applied Spatial Statistics for Public Health Data. Wiley-IEEE, 2004.

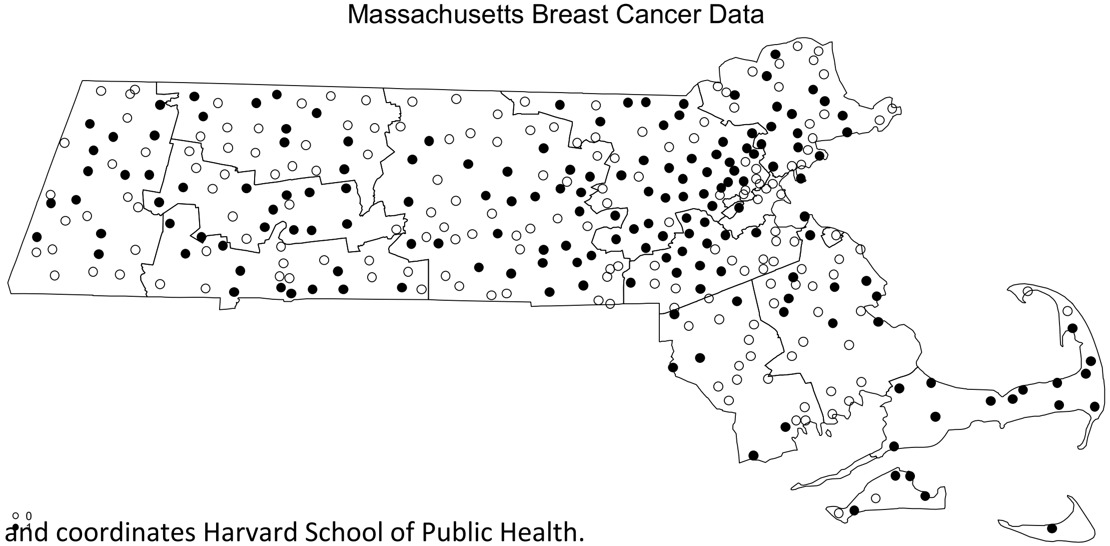
Motivation: spatial distribution (4)

Leukemia Data
In Massachusetts



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

Breast Cancer Data
In Massachusetts



Source: Massachusetts Cancer Report 2006, elaboration and coordinates Harvard School of Public Health.

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

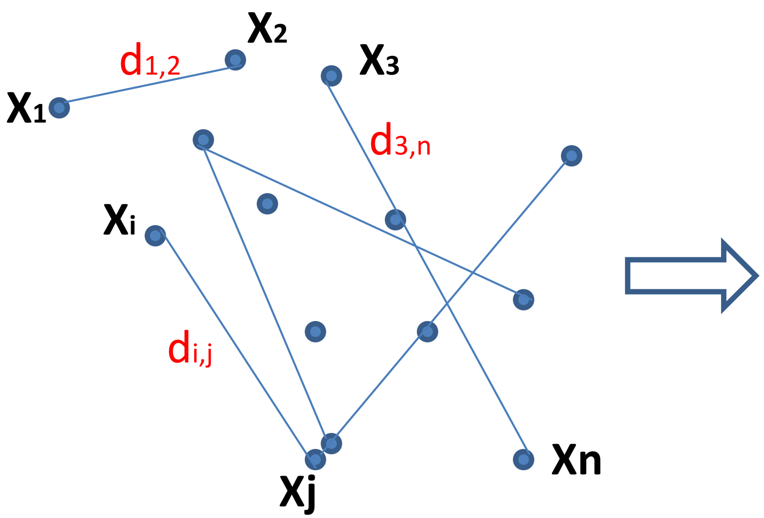
Epidemiology, Sociology, Economics, Demography, etc… whenever the fact that two phenomena are equally distributed over a given area of interest is not obvious and relevant at the same time.

Extensions

- K-dimensional spaces
- Non-Euclidean metrics or general dissimilarity measures
- H_0 : is group j distributed as the underlying null (population) distibution (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.



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

Interpoint Distance Distribution (2)

From *n* observations $\Longrightarrow$ D is an *nxn* symmetric matrix w/ zero main diagonal. We calculate

$d_{i,j}$, $i \neq j$

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

d is a sample of

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

DEPENDENT distances

Interpoint Distance Distribution (3)

We use **all** the $\binom{n}{2} = \frac{n(n-1)}{2}$ distances.

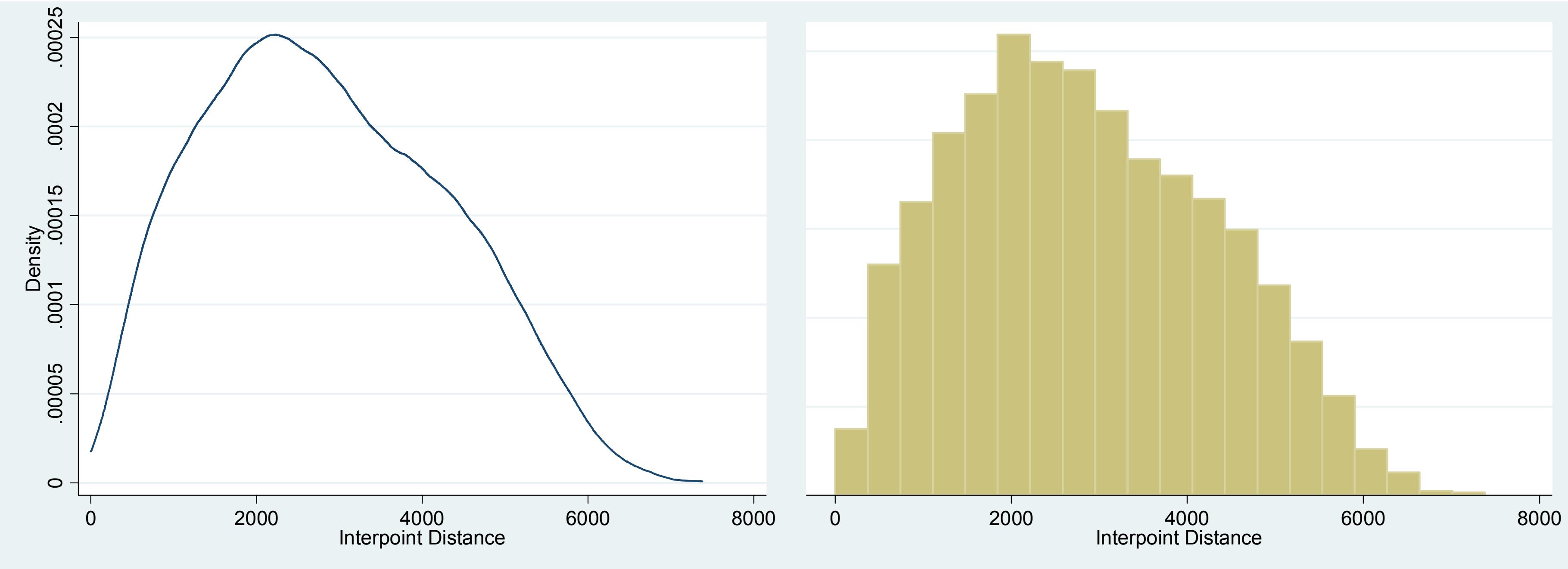
From a sample of ***n*** **observations** we get roughly ***n²/2*** **distances**: COMPLEX RELATION.

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.



The M statistic

Based on a discretized version of the ECDF:

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

With k bins:

$$\hat{F}(\mathrm{d}) = \left[\hat{F}(\mathrm{d}_1), \ldots, \hat{F}(\mathrm{d}_k)\right]$$

where $\hat{F}(\mathrm{d}_\ell) = \binom{n}{2}^{-1} \sum_{i \neq j} 1\{d_{i,j} \leq \mathrm{d}_\ell\}$

The M statistic (2)

1-sample M statistic: one group vs population

H0: $F(\mathrm{d}) = F^0(\mathrm{d})$

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

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

Σ^- is the Moore-Penrose (Mata *pinv()*) generalized inverse of the variance covariance matrix of $\hat{F}(\mathrm{d}), \Sigma$.

The M statistic (3)

$$\Sigma = \left[\sigma_{\ell,m} \right], \text{ with}$$

$$\sigma_{\ell,m} = 4\binom{n}{3}^{-1} \sum_{i<j,k} 1\{d_{i,j} \le \mathrm{d}_\ell\} 1\{d_{i,k} \le \mathrm{d}_m\} - \hat{F}(\mathrm{d}_\ell)\hat{F}(\mathrm{d}_m)$$

Bonetti and Pagano (2005) show that

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

Slow convergence $\Rightarrow$ Monte Carlo test.

The M statistic (4)

2-samples M statistic: Group 1 vs Group 2

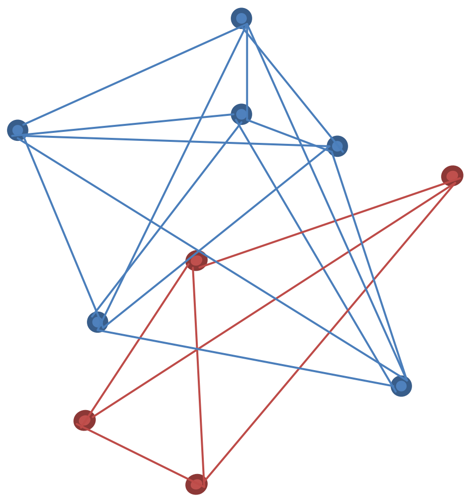
H0: $F_1(\mathrm{d}) = F_2(\mathrm{d})$

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

Equal variance assumption:

$$\sigma_{\ell,m} = \left(\frac{n_1 + n_2}{n_1 n_2}\right) 4 \binom{n}{3}^{-1} \sum_{i<j,k} 1\{d_{i,j} \le \mathrm{d}_\ell\} 1\{d_{i,k} \le \mathrm{d}_m\}$$

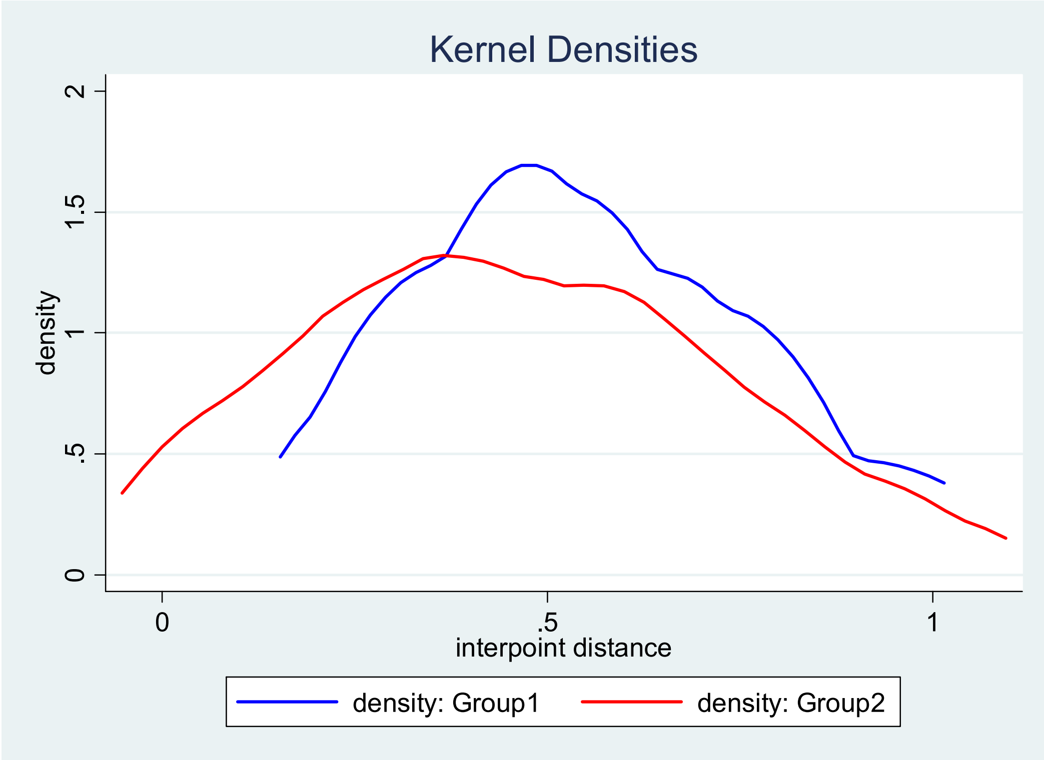
2-samples M Test



10 observations

⇩

45 distances



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 Σ , 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 Σ^{-} from step (2) and (4) of Mstat algorithm compute M;

(E) Iterate (A)-(D) **P** times, generating a vector $[M, M_1, M_2, \ldots, M_P]$;

(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

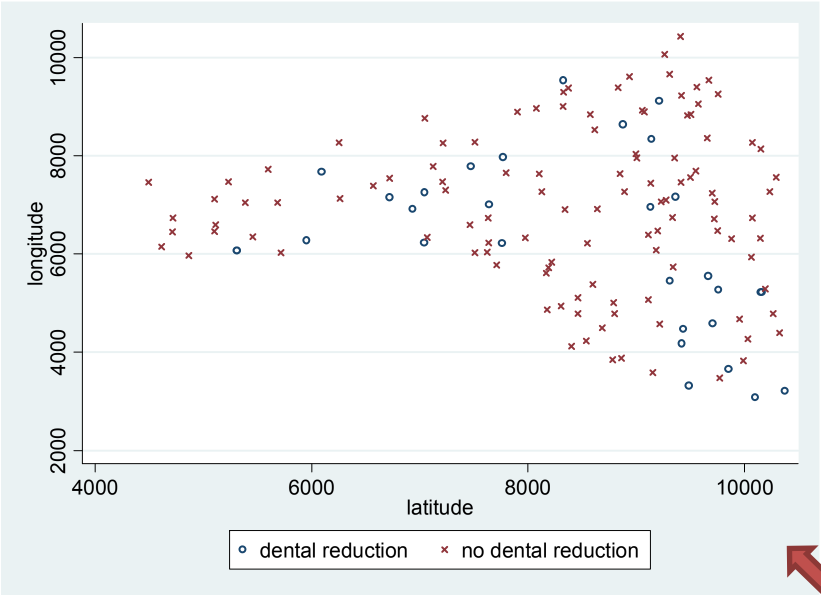
- x-coordinates
- y-coordinates
- Group dummy variable

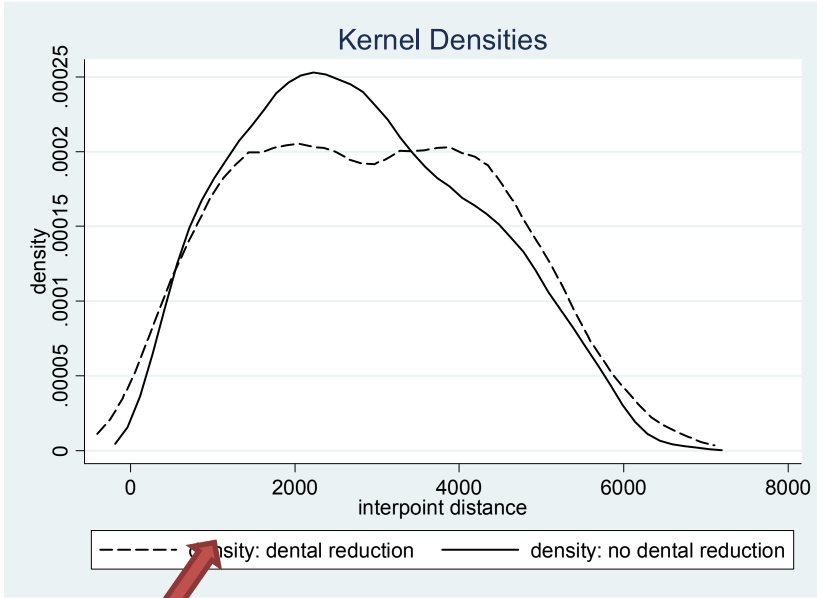
Syntax:

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

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

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
------------------------------------------------------------------------------
T              T(obs)       c       n    p=c/n    SE(p)  [95% Conf. Interval]
-------------+----------------------------------------------------------------
           M |   79.63794      23    1000  0.0230   0.0047   .0146346   .0343123
------------------------------------------------------------------------------
Note:  confidence interval is with respect to p=c/n.
Note:  c = #{T >= T(obs)}
```

Mstat options

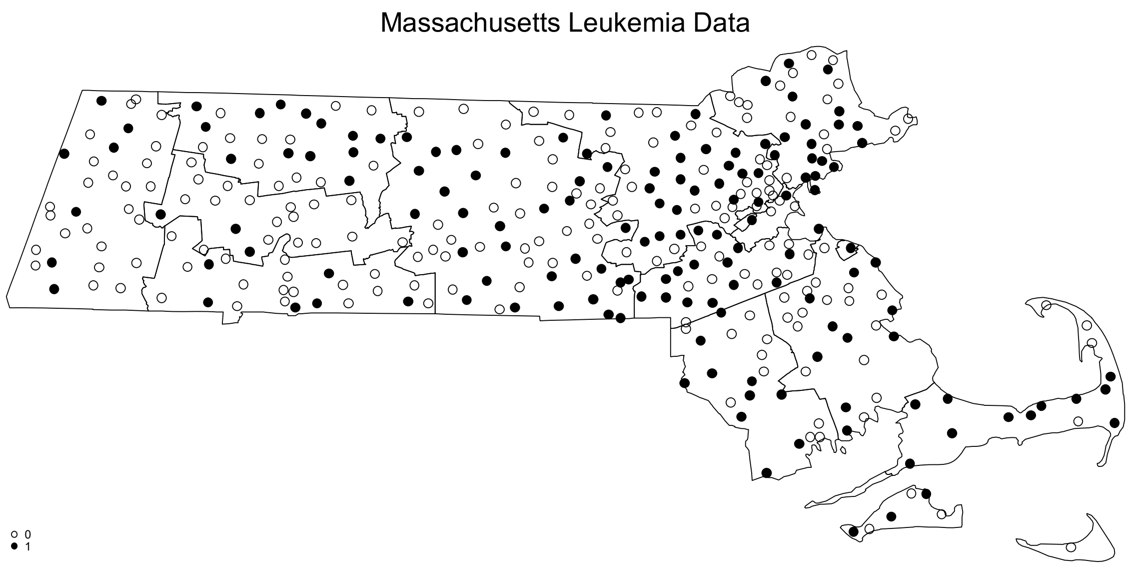
x*(*varname*)	*x-coordinates*
y(varname)*	*y-coordinates*
g(varname)*	*0-1 dummy*
bins(#)	*number of bins*
scatter	*scatter plot*
density	*Kernel density*
chi2	*asymptotic Chi2 pvalue*

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*

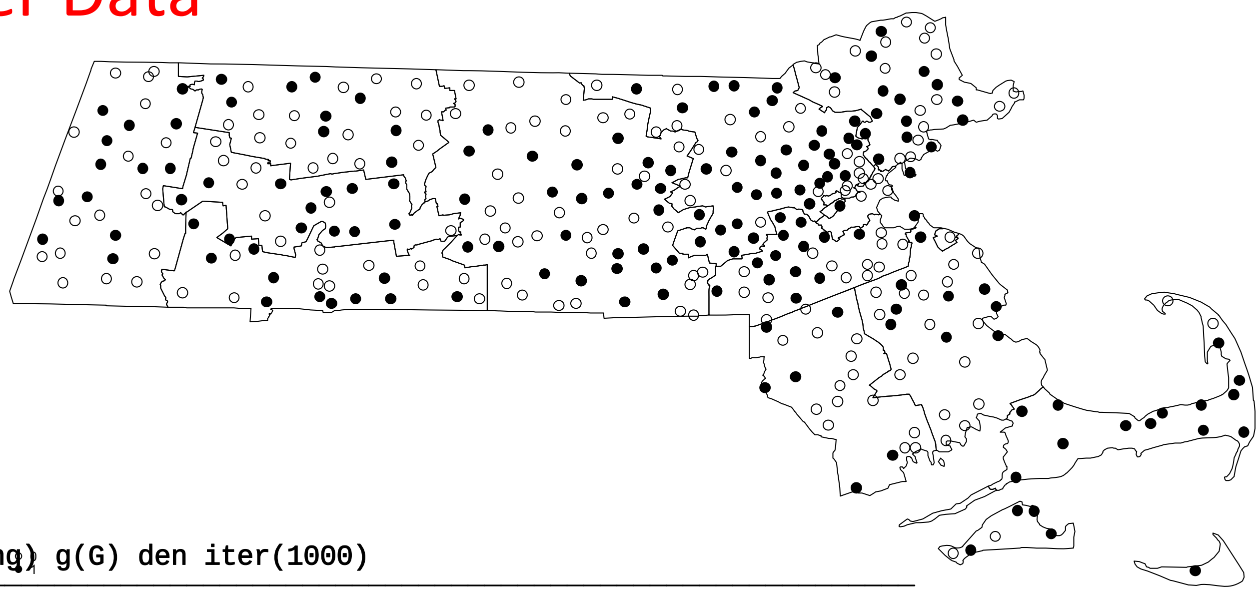
```
. Mtest , x(Lat) y(Long) g(G) den iter(1000)

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

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf.	Interval]
M	26.28204	159	1000	0.1590	0.0116	.1368657	.1831623

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

Cancer Data in Massachusetts (2)

Breast Cancer Data




```
. Mtest , x(Lat) y(Long) g(G) den iter(1000)

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

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf.	Interval]
M	38.56281	4	1000	0.0040	0.0020	.0010909	.0102097

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

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

[1] Bonetti, M. and Pagano, M., *The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering.* Stat Med 2005; 24(5):753-773.

[2] Manjourides, J. and Pagano, M., *A test of the difference between two interpoint distance distributions.* Submitted

[3] Forsberg, L., Bonetti, M. and Pagano, M., *The choice of the number of bins for the M statistic.* Computational Statistics & Data Analysis 2009; 53(10): 3640-3649.

[4] Bonetti, M., Forsberg, L., Ozonoff, A. and Pagano, M., *The distribution of interpoint distances. Mathematical Modeling Applications in Homeland Security.* HT Banks and C Castillo-Chavez, Eds. ; 2003:87-106.

[5] Forsberg, L., Bonetti, M., Jeffery, C., Ozonoff, A. and Pagano, M., *Distance-Based Methods for Spatial and Spatio-Temporal Surveillance. Spatial and Syndromic Surveillance for Public Health (ch.8).* John Wiley & Sons, Ltd; 2005.