Spatial point pattern data - Part II

Applied Spatial Statistics

Spatial point pattern data

- In the previous lecture we introduced spatial point pattern data
- Point patterns have two types of dependece: clustering and repulsion
- The K function is an exploratory tool to detect these types of dependence
- In this lecture we will formally test for dependence

CRS test

- A formal test for a completely random sample (CRS) is often a goal of a study
- Example: Do mountain lions interact with each other?
- Example: Do tumors cluster in the brain?
- This test is also useful for model building
- The hypotheses are

$$\begin{aligned} \mathcal{H}_0 &: \text{completely random sample} \\ \mathcal{H}_1 &: \text{not a completely random sample} \end{aligned}$$

CRS test – Ripley's K

Ripley's K can be used to test for a CRS

1. Generate N data sets of size n over $\mathcal{D}$ using CRS
2. For each dataset, compute the K function, $K_1(t), ..., K_N(t)$
3. For each t, compute the 95% interval of the $K_1(t), ..., K_N(t)$
4. Reject $\mathcal{H}_0$ if the observed K function is outside the interval

CRS test – quadrat

- Split the sampling window into m equally-sized subregions
- Let O_j be the number of observations in region j
- The expected count under $\mathcal{H}_0$ is $E_j = n/m$
- Pick m so $E_j > 5$
- The chi-squared test statistic ($m - 1$ dof) is

$$\chi = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}$$

CRS test – quadrat

- Clustering gives large χ
- Inhibition gives small χ
- Reject $\mathcal{H}_0$ if $\chi < \chi_{m-1,0.025}$ or $\chi < \chi_{m-1,0.975}$
- Usually try the test for a few values of m

CRS test – Clark/Evans

- Let Y_i be the distance between $\mathbf{s}_i$ and its nearest neighbor
- The test statistic is $\bar{Y} = \sum_{i=1}^{n} Y_i/n$
- $\bar{Y}$'s mean and variance under $\mathcal{H}_0$ are

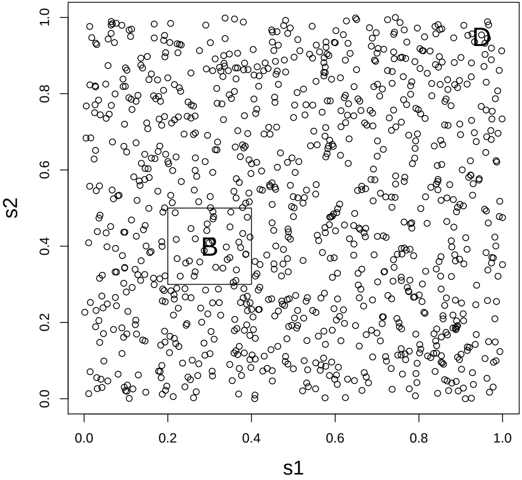
$$\mu = \frac{n}{|\mathcal{D}|} \quad \text{and} \quad \sigma^2 = \frac{4-\pi}{4\pi n^2}|\mathcal{D}|$$

- Clustering gives $\bar{Y} < \mu$ and inhibition gives $\bar{Y} > \mu$
- Reject $\mathcal{H}_0$ if $|\bar{Y} - \mu|/\sigma > 1.96$

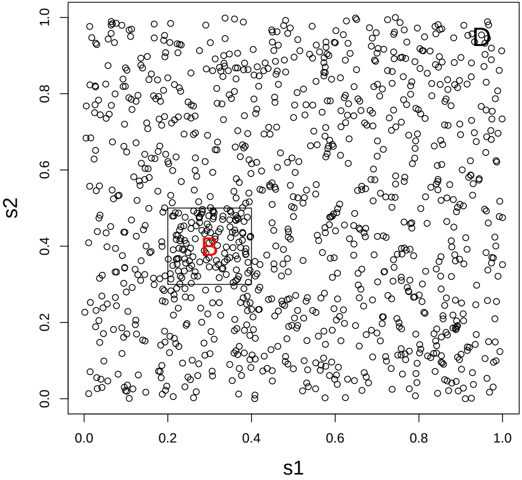
CRS test – scan statistics

- The scan statistic tests for a "hot spot"
- Let $\mathcal{B} \subset \mathcal{D}$ be a subregion
- Example, $\mathbf{s}_i$ is the location of a brain cancer case, $\mathcal{B}$ is Wake Co and $\mathcal{D}$ is North Carolina
- The sampling rate of $\mathcal{B}$, $r(\mathcal{B})$, is the expected proportion of the samples that fall in $\mathcal{B}$
- Under CSR, rate is proportional to area $r(\mathcal{B}) = |\mathcal{B}|/|\mathcal{D}|$
- If $r(\mathcal{B}) > |\mathcal{B}|/|\mathcal{D}|$ then $\mathcal{B}$ is a hot spot

$\mathcal{B}$ is not a hot spot



$\mathcal{B}$ is a hot spot



CRS test – scan statistics

- A scan statistic tests a slightly different set of hypotheses than other CRS tests
- It tests for whether than is a hot spot somewhere in $\mathcal{D}$
- That is, the location of the hotspot is not known
- In words:

$$\begin{aligned} \mathcal{H}_0 &: \text{completely random sample} \\ \mathcal{H}_1 &: \text{there is a hot spot} \end{aligned}$$

- In math:

$$\begin{aligned} \mathcal{H}_0 &: r(\mathcal{B}) = |\mathcal{B}|/|\mathcal{D}| \text{ for all } \mathcal{B} \subset \mathcal{D} \\ \mathcal{H}_1 &: \text{there exists some } \mathcal{B} \subset \mathcal{D} \text{ so that } r(\mathcal{B}) > |\mathcal{B}|/|\mathcal{D}| \end{aligned}$$

CRS test – scan statistics

- For a given $\mathcal{B}$, let $t(\mathcal{B})$ be the test statistic that it is a hotspot
- Typically the scan statistic uses the likelihood ratio statistic
- The test statistic "scans" over all possible hotspots
- The final test statistic is

$$T = \max_{\mathcal{B}} t(\mathcal{B})$$

- Typically it only scans over $\mathcal{B}$ that are circles with radius r

CRS test – scan statistics

The p-value for the test is approximated as

- Generate N datasets of size n over $\mathcal{D}$ under CRS
- For each dataset, compute the scan stat $t_1, ..., t_N$
- The p-value is the proportion of the N scan stats that are larger than the observed scan stat
- If the p-value is less then 0.05 we reject $\mathcal{H}_0$ and conclude there is a hotspot somewhere in $\mathcal{D}$

Advantages of CRS tests

- K-function: good exploratory tool, but must do a test for each distance
- Quadrat: Good text for global homogeneity, but does not pick up local features
- Clark/Evans: great local test, but will miss global features
- Spat stat: good at finding clusters, but narrow alternative hypothesis