Point patterns

- A spatial point pattern is a set of locations generated by some random process. They are distributed within a selected region. The region is usually two-dimensional (but it can be one- or three-dimensional). Examples: lightning strikes, earthquake epicenters, locations of pine trees, etc. We refer to the locations as **events**.

- Consider the point process $\{Z(s) : \mathbf{s} \in D \subset R^2\}$. A realization of this process consists of a pattern (arrangement) of point in D. (D is a random set.). These points are called the events of the point process.

- A point pattern is called completely random pattern (hypothesis of **complete spatial randomness** - CSR) if the following criteria hold:
  - The average number of events (the intensity, $\lambda(s)$) is homogeneous throughout D.
  - The number of events in two non-overlapping subregions A_1 and A_2 are independent.
  - The number of events in any subregion follows the Poisson distribution.

- Analysis of point pattern data begin with a test of CSR hypothesis. If a particular pattern does not reject CSR then usually no further statistical analysis is needed. If CSR is rejected then further investigation is needed to explain the nature of the spatial point pattern.

- Most processes don't follow a complete spatial random pattern. Events may be independent in non-overlapping subregions, but the intensity $\lambda(s)$ is not homogeneous throughout D. For example, more events will be present in regions where the intensity is high and less will be present where the intensity is low. The intensity may be constant, but the present of an event can attract or drive away other events nearby.

- `R` packages for the analysis of spatial point patterns:

  `spatial`
  `splancs`
  `spatstat`
  `maptools`

  See Chapter 7 of Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R, Use R!, Springer.

- Preliminary analysis of a point pattern:
  It is focused on the spatial distribution of the observed events to make inference on the process that generated them. We are interested in (a) the distribution of the events in space and (b) existence of possible interaction between them.

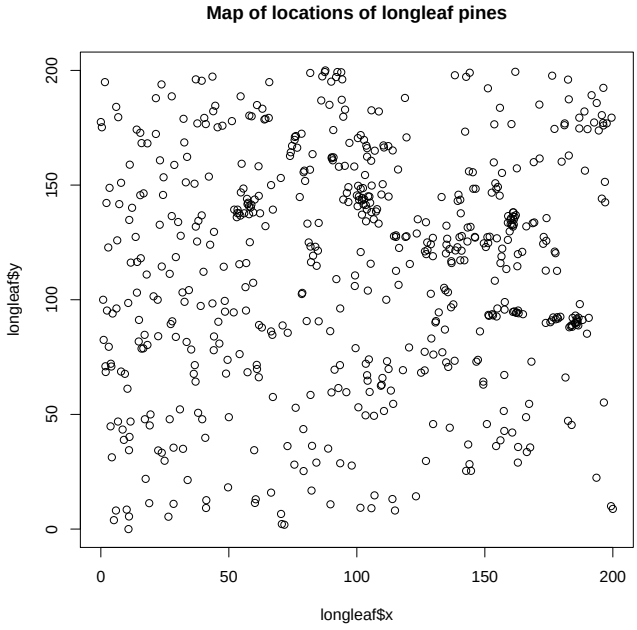
- Poisson process
  There are many types of Poisson processes: Homogeneous Poisson process (HPP), inhomogeneous Poisson process (IPP), Poisson cluster process, and the compound Poisson process. A process is called homogeneous Poisson process if the two following properties hold:
  - If $N(A)$ denotes the number of events in subregion $A \subset D$, then $N(A) \sim \text{Poisson}(\lambda v(A))$, where $0 < \lambda < \infty$ is the constant intensity of the process.
  - If A_1 and A_2 are two disjoint subregions of D, then $N(A_1)$ and $N(A_2)$ are independent.

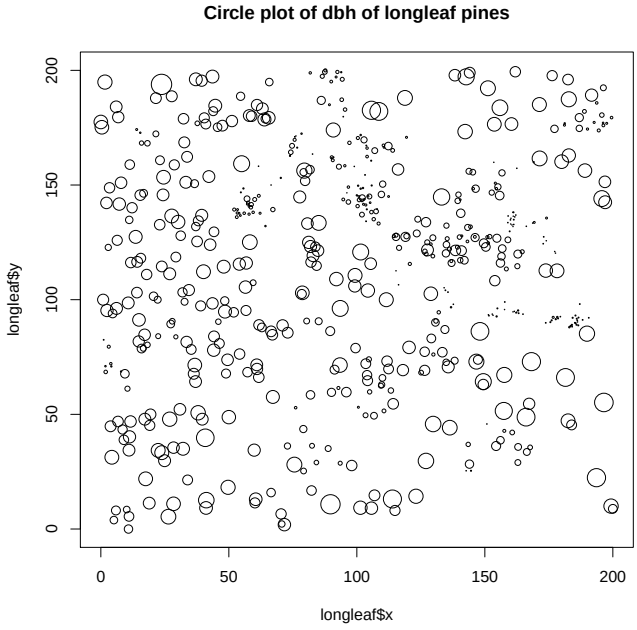
  If the intensity function $\lambda(s)$ varies spatially then the first condition does not hold, but the second condition may still hold. In this case the process is called inhomogeneous Poisson process. (The homogeneous Poisson process is a special case of the inhomogeneous Poisson process.). The homogeneous Poisson process is also called the stationary Poisson process, while the inhomogeneous Poisson process is also called the non-stationary Poisson process.

- Testing for complete spatial randomness:
  We want to test if the observed point pattern is a realization of a homogeneous Poisson process. The statistical tests are based on counts of events in regions (quadrats) or based on distances. When the sampling distributions are difficult one can rely on simulations methods. There are two methods of simulations: The Monte Carlo test and simulation envelopes.

- Test based on quadrats

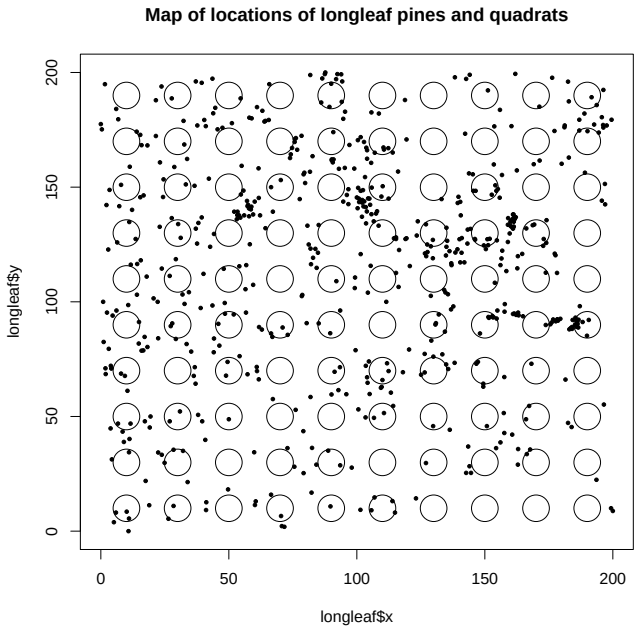
  Data set `longleaf`: 584 long-leaf pine trees from the Wade Tract, a forest in Thomas County, Georgia. The data consists of the location of each tree (x, y) coordinates and its diameter at breast height (`dbh`) in centimeters. Area covered 200m $\times$ 200m.





  Do the spatial locations in these plots appear completely random or are they clustered?

One hundred non-overlapping quadrats (each one with radius 6 meters) were randomly chosen in the area of study and the number of trees were counted in each quadrat. The next plot is only an example! (The quadrats are not randomly chosen.)



The following frequencies are obtained:

Trees per quadrat	Observed frequency	Estimated frequency
0	34	23.93
1	33	34.22
2	17	24.47
3	7	11.66
4	3	4.17
5	1	1.19
6	1	0.28
7	2	0.06
8	1	0.01
9	0	0.00
10	1	0.00

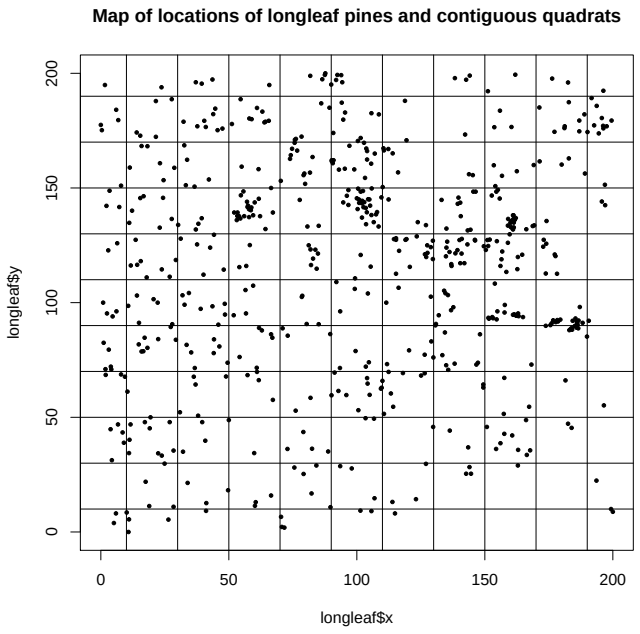
Estimate Poisson parameter $\hat{\lambda} = \frac{34\times0+33\times1+...+1\times10}{100} = 1.43$. The expected frequencies are computed using the Poisson probability mass function, e.g., the expected number of quadrats with zero trees will be $100 \times P(Y = 0) = 100\frac{1.43^0 exp(-1.43)}{0!} = 23.93$.

To test the hypothesis of a complete spatial randomness (which is synonymous with homogeneous Poisson process) one can use the χ^2 goodnes-of-fit test.

$$X^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(34 - 23.93)^2}{23.93} + \ldots + \frac{(6 - 1.54)^2}{1.54} = 21.67.$$

Since $21.67 > \chi^2_{0.95;4} = 9.49$ the null hypothesis of homogeneous Poisson process is rejected.

- If quadrats are contiguous then we have lattice data.



Using Moran's I and Geary's c statistics we can test for clustering. Here we identify regions with high values in short distances which suggests clustering.

- Test based on distances
  Distance methods use the exact location of the events. They do not dependent on the arbitrary choice of quadrat size or shape. See Cressie (1993), p. 604 for various test statistics based on distances along with their asymptotic distributions.

- Test based on simulations

  - Monte Carlo tests
    These can be used for many statistical analysis spatial or non-spatial. The general idea: Compute a test statistic from the observed data, call it q_0. Then simulate the random process say, g times. For each realization compute the test statistics $q_1, \ldots, q_g$. Then rank the simulated test statistics and place the observed test statistic q_0 in the ordered array and compute the p-value. For example the average nearest neighbor distance can be used. It can be computed using simulations by generating points independently and uniformly in the area of interest.

  - Simulation envelopes
    Find the nearest neighbor distance for event $i = 1, \ldots, n$. Let $d_1, d_2, \ldots, d_n$ be the nearest neighbor distances. Compute the estimate of the distribution function $\hat{G}(d)$ of nearest-neighbor event distances as follows: Let $I(di \leq r)$ be the indicator function that takes the value 1 if $di \leq r$ or 0 if $di > r$. Compute $\hat{G}(r) = \frac{1}{n} \Sigma_{i=1}^{n} I(di \leq r)$ for various n distance r.

    For the same distances r compute the theoretical G function. Under csr it is equal to $G(r) = 1 - e^{-\lambda \pi r^2}$.

Finally, plot $\hat{G}(r)$ against $G(r)$. Under complete spatial randomness the plot should be roughly linear.

To measure the departure from linearity we should find the sampling distribution of $\hat{G}(r)$ under complete spatial randomness, which not easy, because of the dependence between the distances (for example, if the nearest neighbor for point 1 is point 2 then the nearest neighbor for point 2 will be point 1, and so on). We therefore assess linearity using simulations. We compute $\hat{G}(r)$ for many simulations. Each simulation consists of n independent uniformly distributed points in the area of interest. For each simulation we compute the minimum and maximum value of $\hat{G}(r)$ to construct the simulation envelope as shown below.