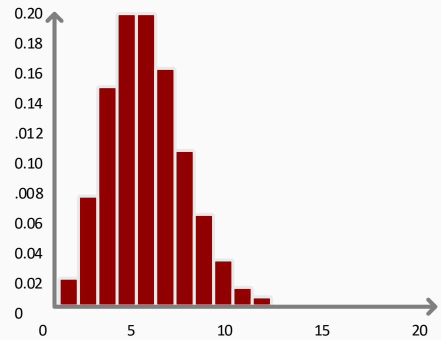
working with spatial data

Spatial Poisson Processes



The Poisson Distribution:

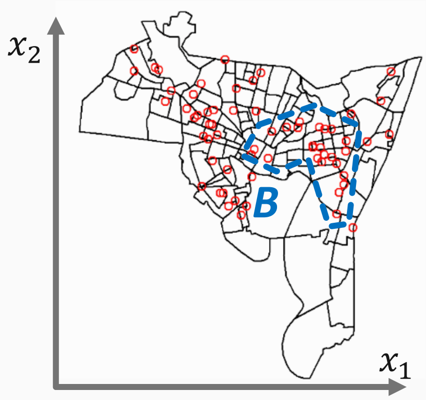
The **Probability Mass Function (PMF)** assigns a random variable ***X*** as **Poisson** with rate parameter lambda (***λ***)

$$X \sim \text{Pois}(\lambda)$$

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 1, 2, 3, \ldots$$

$$E(X) = \lambda, \quad \text{Var}(X) = \lambda$$

Assuming a model of crime rate occurring in a given month within area B designated spatially:



Choose any region B and any number of crimes n.

Determine how often the number of points in $B = n$.

$$P(\text{Number of points } in \ B = n) = \cdots$$

- Should depend on the size of area B
- Should depend on the value of n
- Should depend on the rate of events λ

$$P(\text{Number of points } in \ B = n) = \frac{(\lambda \,\text{size}(B))^{n}}{n!} e^{-\lambda \,\text{size}(B)}$$

The expression above is referred to as a **Homogeneous Point Process** with parameter λ

There is a flaw in the logic: The distribution assumes crimes can happen equally anywhere. If the crime rate should be represented as a non-constant, then lambda λ should be represented as non-constant; it should change depending on where the model is spatially. So when the latter is true, the expression above is becomes an **Inhomogeneous Point Process** with parameter $\lambda(x_1, x_2)$ the rate of events should λ depend on x_1, x_2. Considering the area B can have more concentrated rates of crime λ than others, the rate λ is integrated over the area B:

$$P(\text{Number of points } in \ B = n) = \cdots$$

$$\Lambda(B) = \int_{B} \lambda(x_1, x_2) dx_1 dx_2$$

$$P(\text{Number of points } in \ B = n) = \frac{(\Lambda(B))^{n}}{n!} e^{-\Lambda(B)}$$

possibly $\lambda(x_1, x_2) = ax_1 + bx_2 + c$

possibly $\lambda(x_1, x_2) = \text{polynomial}(x_1, x_2)$

possibly $\lambda(x_1, x_2) = \sum_i K_h([x_1, x_2], [x_1^i, x_2^i])$

possibly $\lambda(x_1, x_2) = f(\text{proximity to nearest drug house})$

The expression is similar to before with the rate integrated over B with various dependencies (above).

The scientist will do the modeling and the coding packages will fit the model accordingly. The code will determine the parameters using **Maximum Likelihood Estimate** and help determine the **bandwidth**. This process goes a step further than **Density Estimation**; a model is now created and parameters can be fitted accordingly.

Density of breaches in New Haven