Spatial Data Analysis in R

Preparation for Point Patterns and Analysis

Eco 697DR – University of Massachusetts, Amherst – Spring 2022
Michael France Nelson

Unit 3

Unit 3 Topics

- Functions and Statistics Review
- Point patterns and processes
- Descriptive spatial statistics

For today:

- Functions and stats refresher

- For next week:
  - Read F+F chapter 3 (up to the R examples)
  - Browse F+F chapter 4 (up to the R examples)

Probability theory and stats concepts

The book uses some probability theory oriented concepts, terminology, and notation.

I want to refresh our memory on key probability theory and statistics concepts so that we are all on the same page.

Theoretical functions and numerical estimates

Theoretical functions are pure objects that don't depend on our ability to write down an equation or estimate values from real data.

Often they have symbolic formulas: $\mathbf{y = mx + b}$

Other times descriptions may be presented verbally: "The number of points within a radius of r."

Symbolic function expressions can be very complicated, converting the components to words can help.

Population and sample

Populations are usually too large to sample fully. We often model them with theoretical distributions characterized by *parameters*.

Samples are subsets of populations that we use to estimate *statistics*.

Statistics are our best educated guesses for the true population parameter values.

Population and sample

Population parameters, i.e. quantities defined by theoretical functions, are represented by variables with no decoration.

Estimates of population parameters, i.e. quantities we estimated from real data, are usually notated with a ‘hat’: $\hat{y}$

Probability Density (or Mass) Functions (PDFs) are theoretical functions whose parameters we can estimate with statistics we calculate from data.

Center and spread

We are accustomed to using mean, median, or mode as a measure of the *center* of our data. There are formulas we can use to estimated these quantities from data.

Probability theory uses the concept *Expected Value*, which has a precise mathematical definition. In most cases it is equivalent to the *mean*.

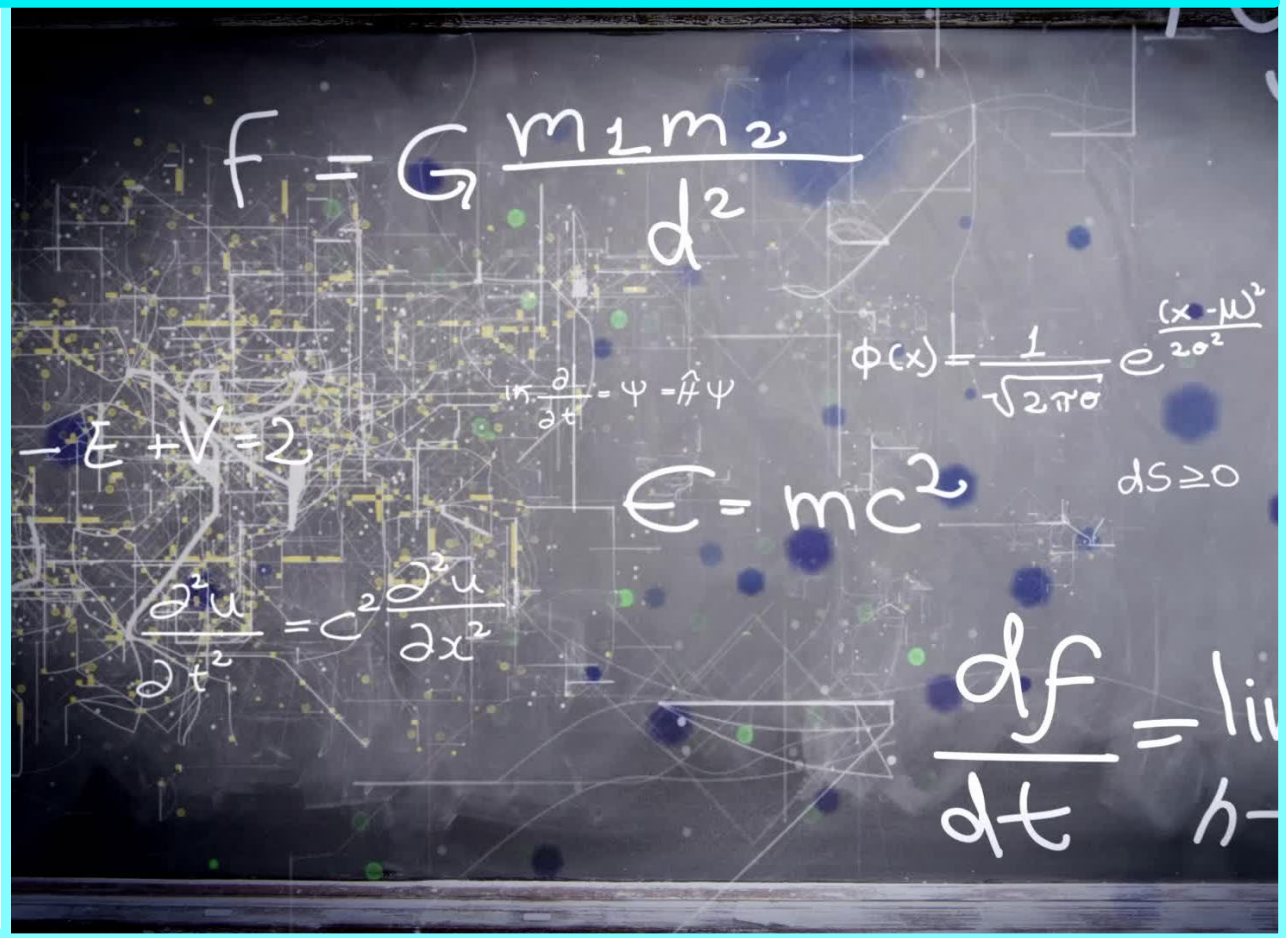
The Expected Value function is often notated with a capital E: $\mathbf{E}(\mathbf{x})$. The Fletcher and Fortin text uses this notation.

We usually think of the *spread* of our data in terms of variance or *standard deviation*.

Indicator Functions:

Logical functions: usually return values of 0 or 1

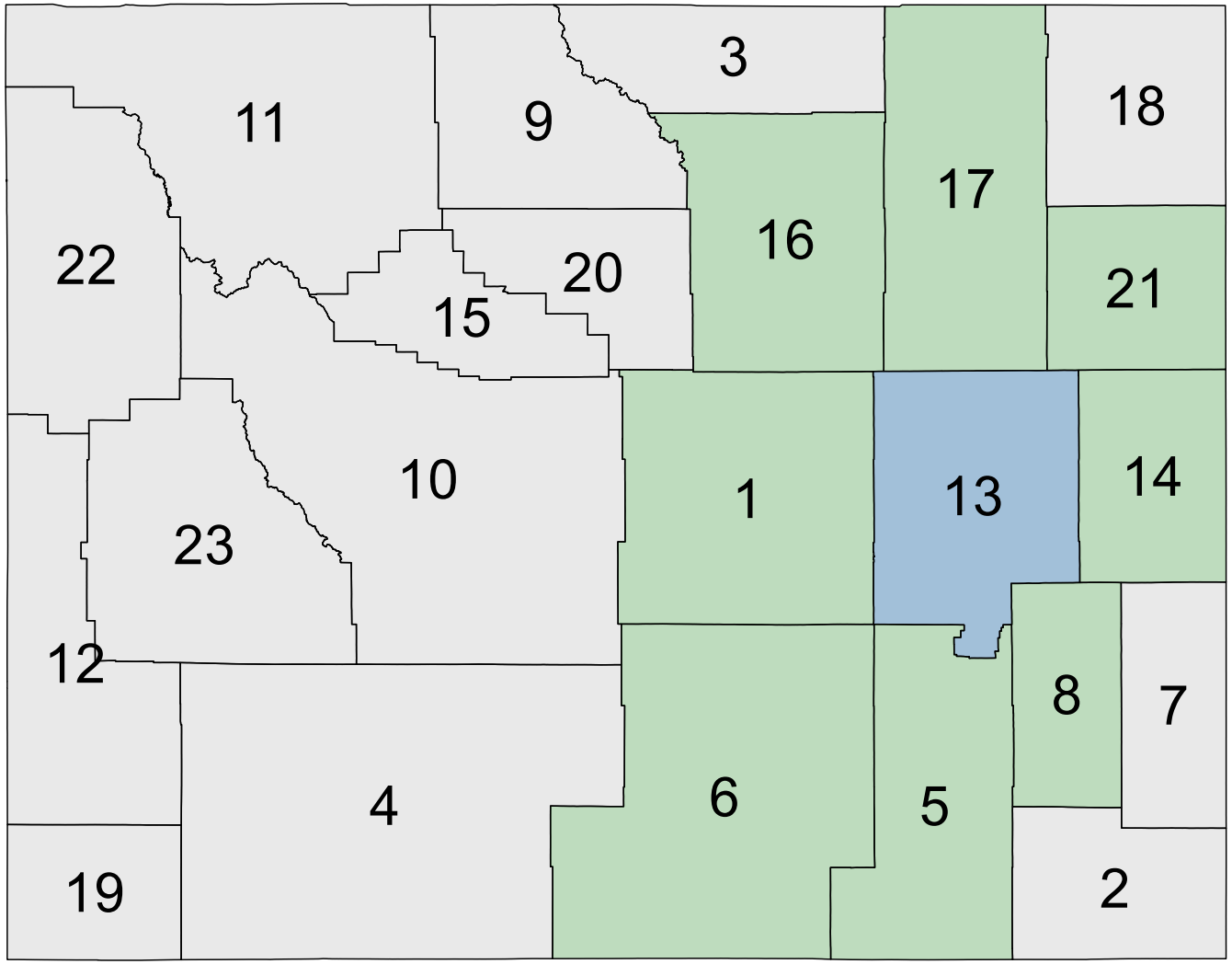
- Indicator functions are tests for whether a desired condition is true.
  - Prime number indicator function:
    - F(2) = 1
    - F(3) = 1
    - F(4) = 0
    - F(5) = 1
    - F(42) = 0



Neighbor Indicator Functions: Edge or Vertex

Which counties share an edge or vertex with lucky county # 13?

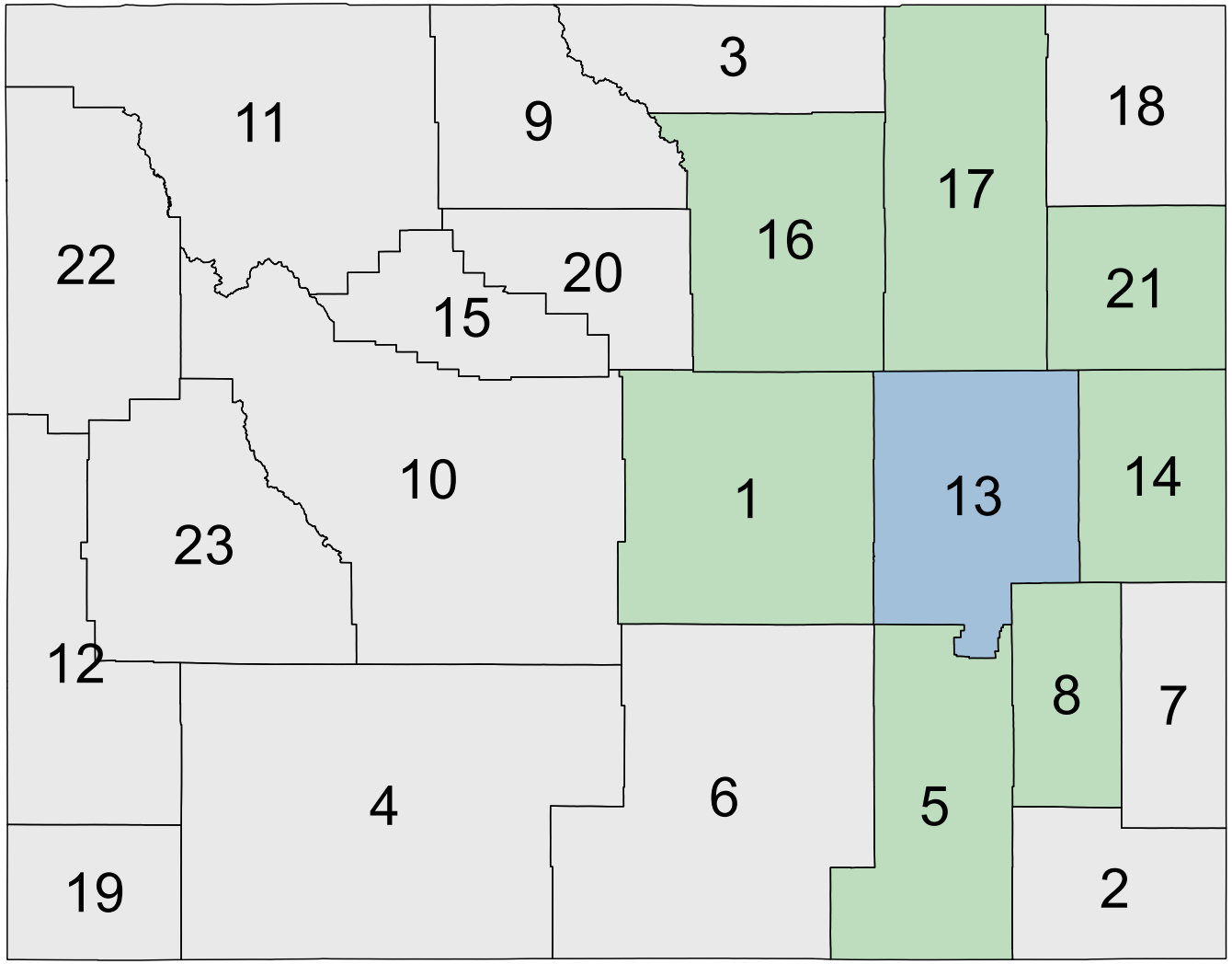
- F(13, 1) = 1
- F(13, 6) = 1
- F(13, 11) = 0



Neighbor Indicator Functions: Edge Only

Which counties share an edge or vertex with lucky county # 13?

- F(13, 1) = 1
- F(13, 6) = 0
- F(13, 11) = 0



Quantifying Variation

Dispersion

How might we measure variability in a collection of numbers?

Dispersion

Difference between min and max values?

- 0^{th} and 100^{th} percentiles

Difference between 1^{st} and 3^{rd} quartile?

- Interquartile range

Differences between mean and observations?

- Absolute values, sum of absolute values
- Sum of squared values (SS)?
- Average of the (SS)?

Variance and Covariance: Key Concepts

- Variance and covariance are measures of *spread* or *dispersion.*
  - What are some other measures of dispersion that we know about?
- Variance and covariance are sample statistics - we can use them to estimate population parameters.
- Variance is a univariate statistic.
- Covariance is a bivariate statistic.
- Covariance measures an association between two variables: this is directly analogous to correlation!
- Pearson correlation is built up from components of variance and covariance. It is like a normalized version of covariance.

Variance: Notation

First some notational conventions:

Our values in set notation:

$X = \{x_1, x_2, x_3, \ldots, x_n\}$ - note the capital X for the set, and the lowercase x for the elements

The sum of values in sigma notation:

$$\sum_{i=1}^{n} x_i$$

The sample size: n

Starting Simple: Formula Chunks

The mean formula chunk

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The SS chunk:

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The sigma decorations are often dropped:

$$SS = \sum (x_i - \bar{x})^2$$

Variance

Variance is a measure of *dispersion* or *spread*

- In words: the average of the squared differences
- It's just the SSE normalized by the [adjusted] sample or population size.
  - For a population we use N, a sample uses n - 1

Formulae: Populations

- for populations

$$Var(x) = \frac{1}{N}\sum(x_i - \bar{x})^2 = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N}$$

- for samples:

$$Var(x) = \frac{1}{N-1}\sum(x_i - \bar{x})(x_i - \bar{x}) = \frac{\sum(x_i - \bar{x})^2}{N-1}$$

Variance

Variance is a measure of *dispersion* or *spread*

- In words: “The variance is the average of the squared differences from the mean.”
- It’s just the SSE normalized by the [adjusted] sample or population size.
  - For a population we use N, a sample uses n - 1

Why do we want to use squared differences?

$$(x_i - \bar{x})^2$$

- What is the sign of this term?
- What would happen to the sum if we used the unsquared differences?
- Why not just use the absolute value?

Variance

Variance is a measure of *dispersion* or *spread*

- In words: “The variance is the average of the squared differences from the mean.”
- It’s just the SSE normalized by the [adjusted] sample or population size.
  - For a population we use N, a sample uses n - 1

Why do we want to use squared differences?

$$(x_i - \bar{x})^2$$

- What is the sign of this term?
  - It is always positive!
- What would happen to the sum if we used the unsquared differences?
  - They would sum to zero for both dispersed and clustered data.
- Why not just use the absolute value?
  - The squaring penalizes large deviations, this has desirable theoretical and practical consequences.

Covariance

Covariance measures the *dispersion* of one variable, x, in the context of the *dispersion* of a second variable, y.

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

- Covariance tells us the amount by which the changes in one variable are *coordinated* with changes in another.
- $(x_i - \bar{x})(y_i - \bar{y})$ is like a *crossed* version of the squared errors…
  - But the term *cross product* is already taken.

It turns out that *variance* is a special case of *covariance*.

$$Cov(x,x) = Var(x)$$

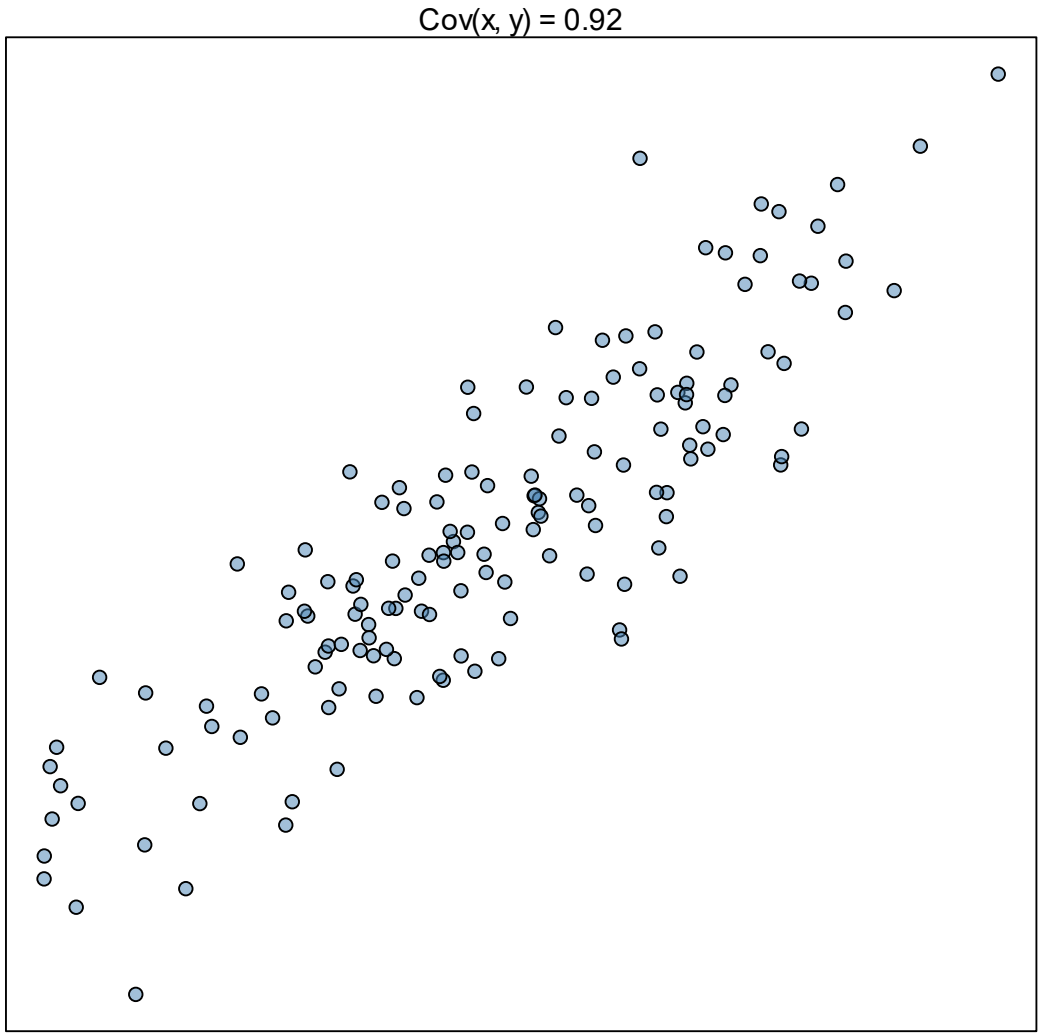
Covariance

Case 1: Positive covariance: High x-values tend to co-occur with high y-values.

Most terms will be positive	Few terms will be negative
$(x_i > \bar{x})$ AND $(y_i > \bar{y}) =$ positive	$(x_i > \bar{x})$ AND $(y_i < \bar{y}) =$ negative
$(x_i < \bar{x})$ AND $(y_i < \bar{y}) =$ positive	$(x_i < \bar{x})$ AND $(y_i > \bar{y}) =$ negative

Covariance

Positive Covariance



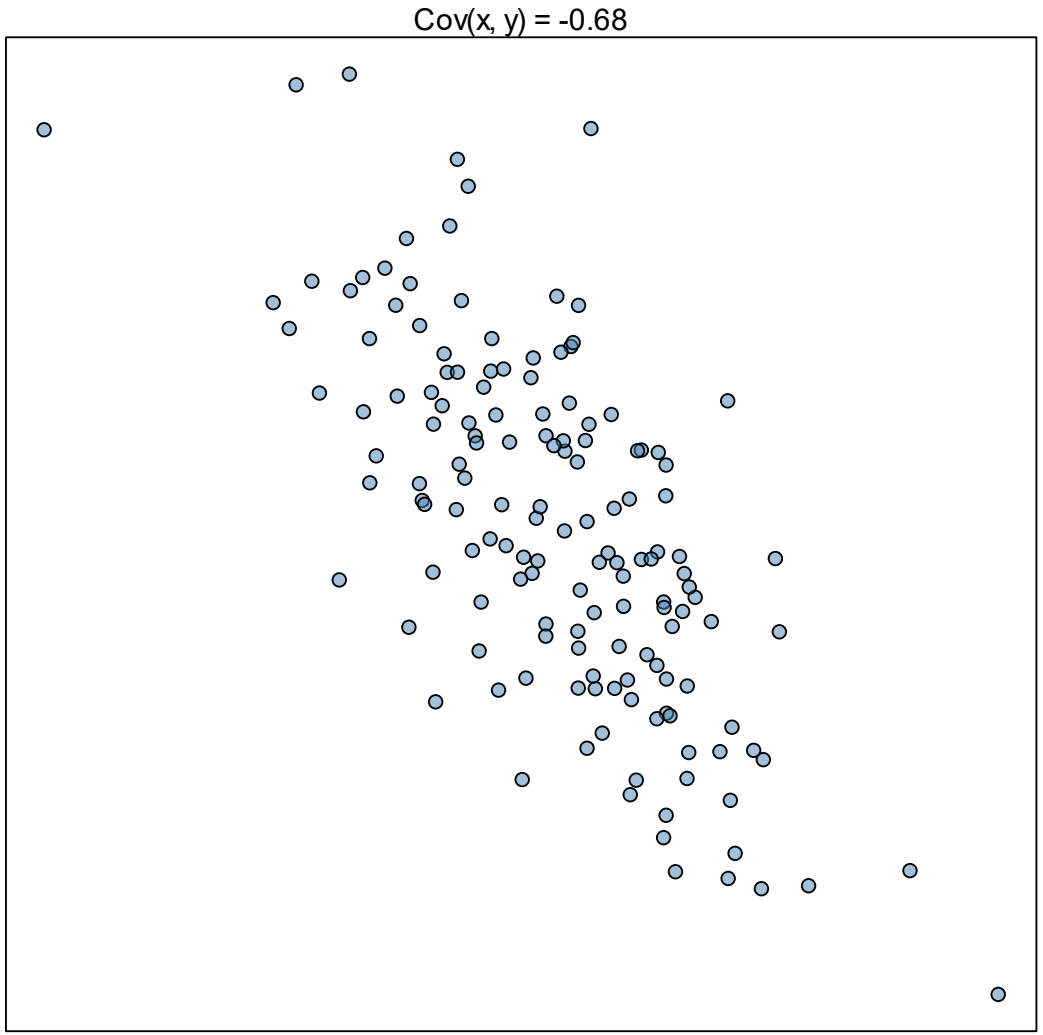
Covariance

Case 2: Negative covariance: High x-values tend to co-occur with low y-values.

Few terms will be positive	Most terms will be negative
$(x_i > \bar{x})$ AND $(y_i > \bar{y}) =$ positive $(x_i < \bar{x})$ AND $(y_i < \bar{y}) =$ positive	$(x_i > \bar{x})$ AND $(y_i < \bar{y}) =$ negative $(x_i < \bar{x})$ AND $(y_i > \bar{y}) =$ negative

Covariance

Negative Covariance



Covariance

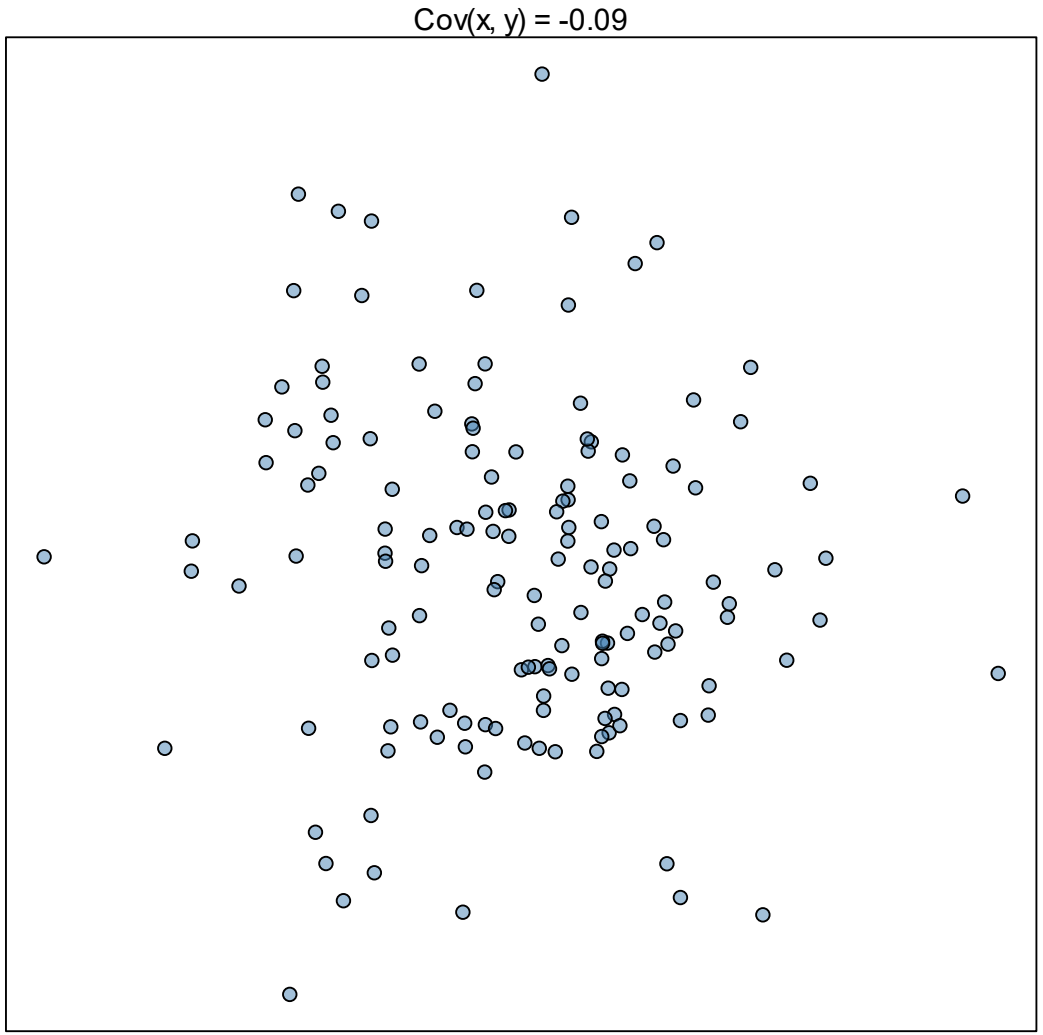
Case 3: no covariance: No association between above average x and above average y

Negative and positive values cancel – sum is near zero

About half the terms will be positive	About half the terms will be negative
$(x_i > \bar{x})$ AND $(y_i > \bar{y}) =$ positive	$(x_i > \bar{x})$ AND $(y_i < \bar{y}) =$ negative
$(x_i < \bar{x})$ AND $(y_i < \bar{y}) =$ positive	$(x_i < \bar{x})$ AND $(y_i > \bar{y}) =$ negative

Covariance

Zero Covariance



Recap

Common chunks

- Means: $\bar{x} = \frac{\sum x}{n}$
- Sums of errors: $\sum(x_i - \bar{x})$
  - Remember this sum is zero!
- Sums of squared errors: $\sum(x_i - \bar{x})^2$
- Sums of squared errors: $\sum(x_i - \bar{x})(x_i - \bar{x})$
- Sums of crossed errors: $\sum(x_i - \bar{x})(y_i - \bar{y})$
- Normalizing by population size:
  - $\frac{1}{N}$ and $\frac{1}{N-1}$

Distributions: The Poisson

The Poisson Distribution

A key distribution in spatial analysis is the Poisson

It has a single parameter: λ

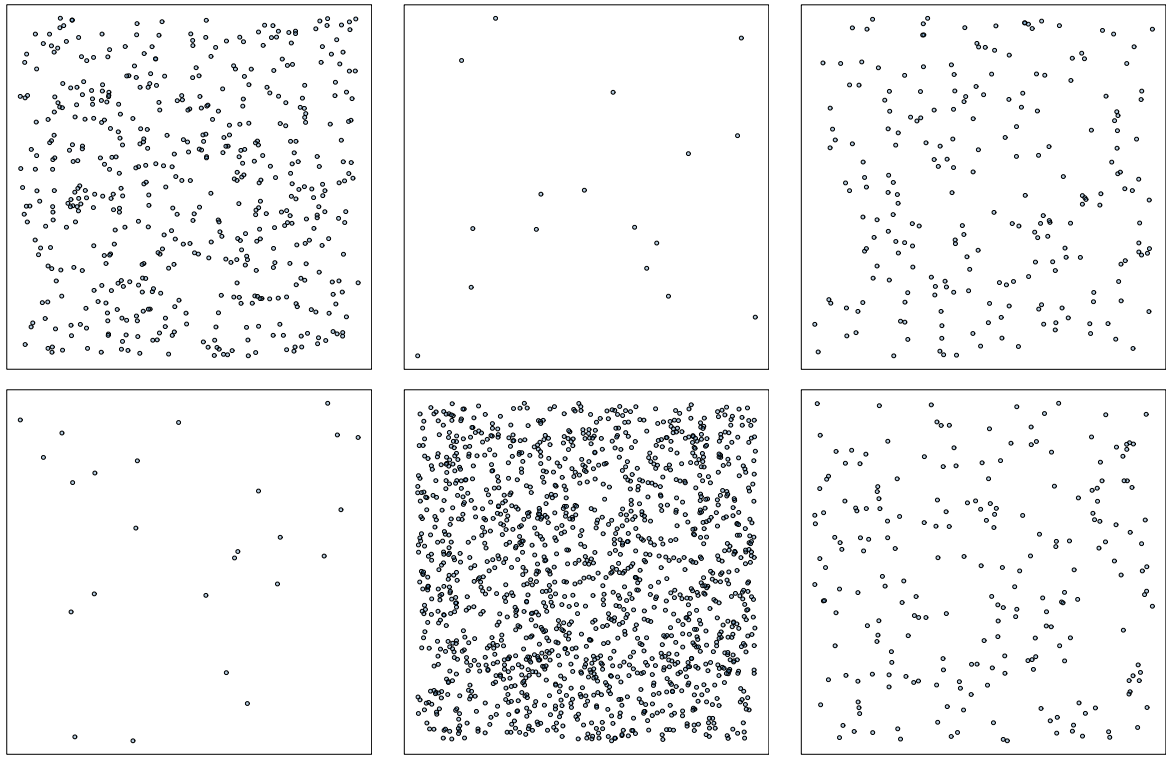
- The Poisson distribution describes counts:
  - A Poisson event is a count, or census.
  - The sample space is $\{0,1,2,\ldots,\infty\}$
  - It has an *infinite sample space*!

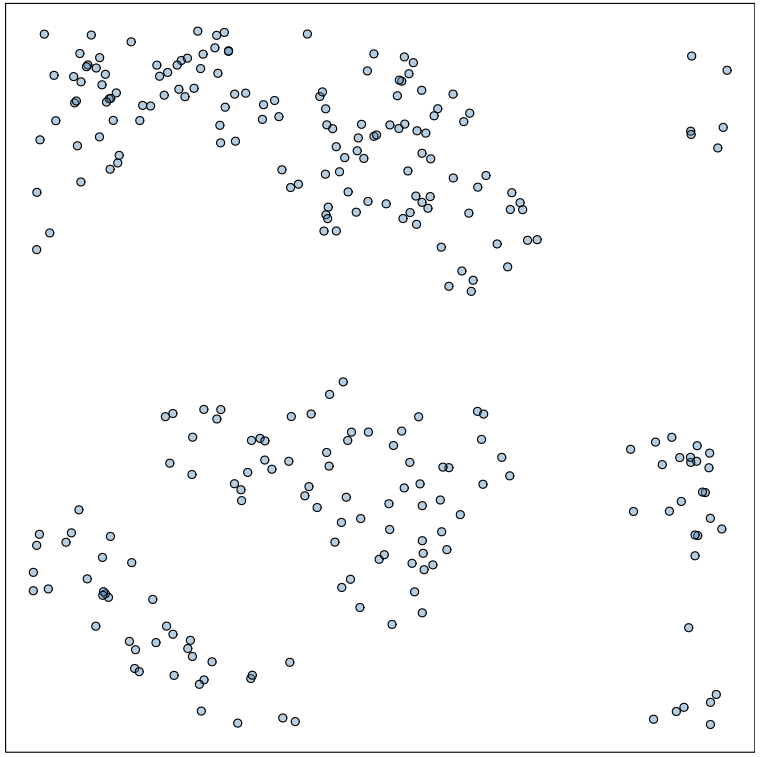
Key property: the standard deviation is equal to the mean!

Poisson distribution is often appropriate for things that occur *randomly* but at a certain *constant rate*.

The Poisson Distribution

The Poisson distribution is very important *null model* in spatial statistics





The Poisson Distribution

Completely Spatially Random (CSR) point patterns follow a Poisson Distribution – It is a great model for **point processes.**