# Probabilistic Species Co-occurrence

Veech, J. A. (2012). A probabilistic model for analysing species co-occurrence. Global Ecology and Biogeography, 22(2), 252–260. doi:10.1111/j.1466-8238.2012.00789.x

THAT darn DATA

# What's the probability of two species co-occurring?

| | |
|---|---|
| $N$ | number of sites |
| $N_1$ | number of sites occupied by species 1 |
| $N_2$ | number of sites occupied by species 2 |
| $p_j$ | probability that species 1 and 2 co-occur at exactly j sites, for $j = 0$ to $N$ |
| $Q_{obs}$ | observed number of sites with both species |

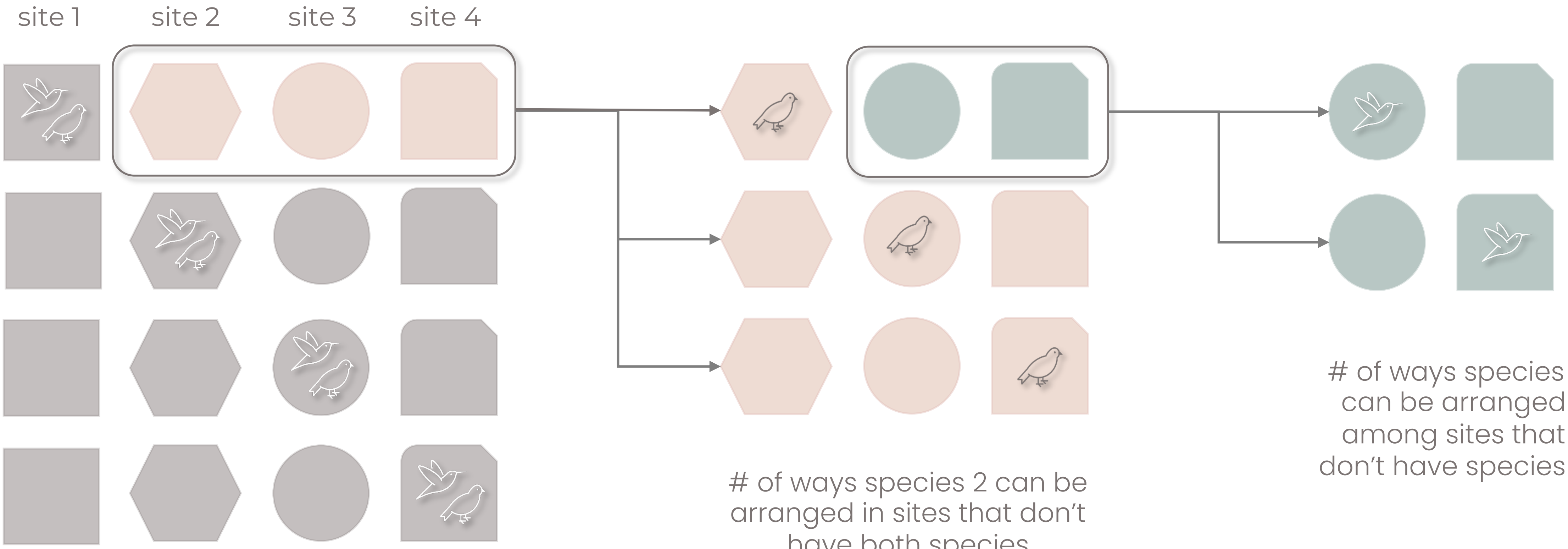
## NOTATION


## What's the probability that two species co-occur at **exactly** j sites?

$$p_j = \frac{\text{total \# of ways species 1 and 2 could be arranged among N sites for a given } N_1, N_2, \text{ and } j}{\text{total \# of ways species 1 and 2 could be arranged among N sites for a given } N_1 \text{ and } N_2, \text{ without considering } j}$$

# THE NUMERATOR

= species 1 = species 2 $N = 4$ $N_1 = 2$ $N_2 = 2$ $j = 1$




# of ways j sites can be arranged among N sites

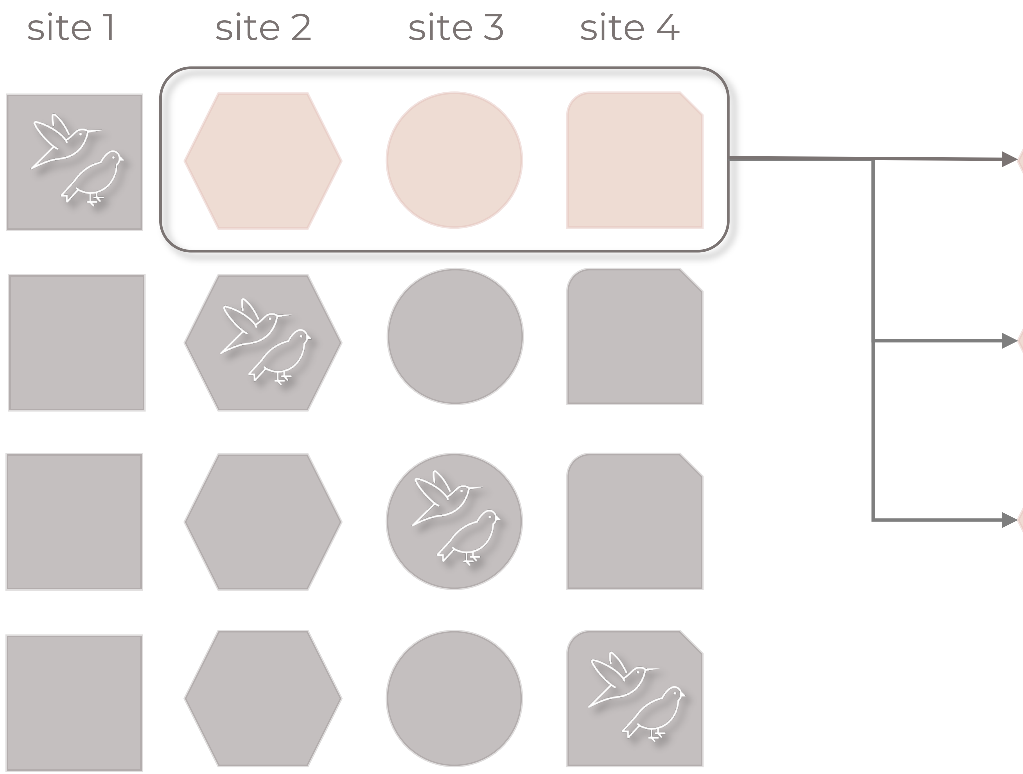
# of ways species 2 can be arranged in sites that don't have both species

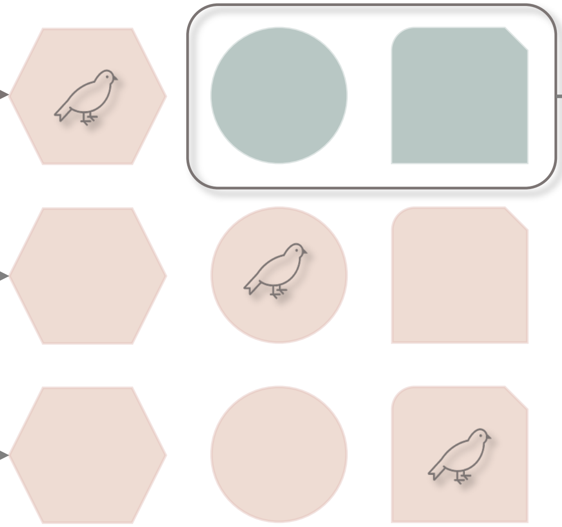
# of ways species 1 can be arranged among sites that don't have species 2

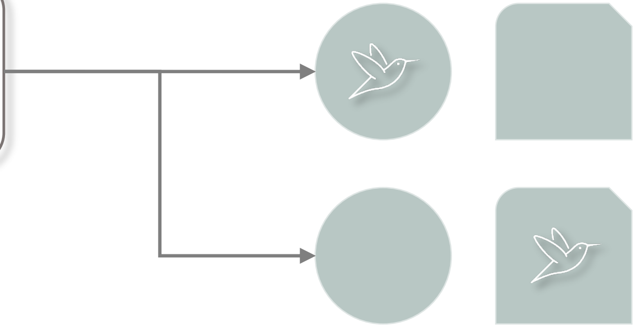
$$4 \times 3 \times 2 = 24$$

# THE NUMERATOR

= species 1 = species 2 $N = 4$ $N_1 = 2$ $N_2 = 2$ $j = 1$





# of ways j sites can be arranged among N sites

$$\binom{N}{j} = \binom{4}{1} = 4$$

# of ways species 2 can be arranged in sites that don't have both species

$$\binom{N-j}{N_2-j} = \binom{4-1}{2-1} = \binom{3}{1} = 3$$

# of ways species 1 can be arranged among sites that don't have species 2

$$\binom{N-j-(N_2-j)}{N_1-j} =$$

$$\binom{N-N_2}{N_1-j} = \binom{4-2}{2-1} = \binom{2}{1} = 2$$

$\binom{y}{x} = \frac{y!}{x!(y-x)!}$ = # of ways to select x items from y items

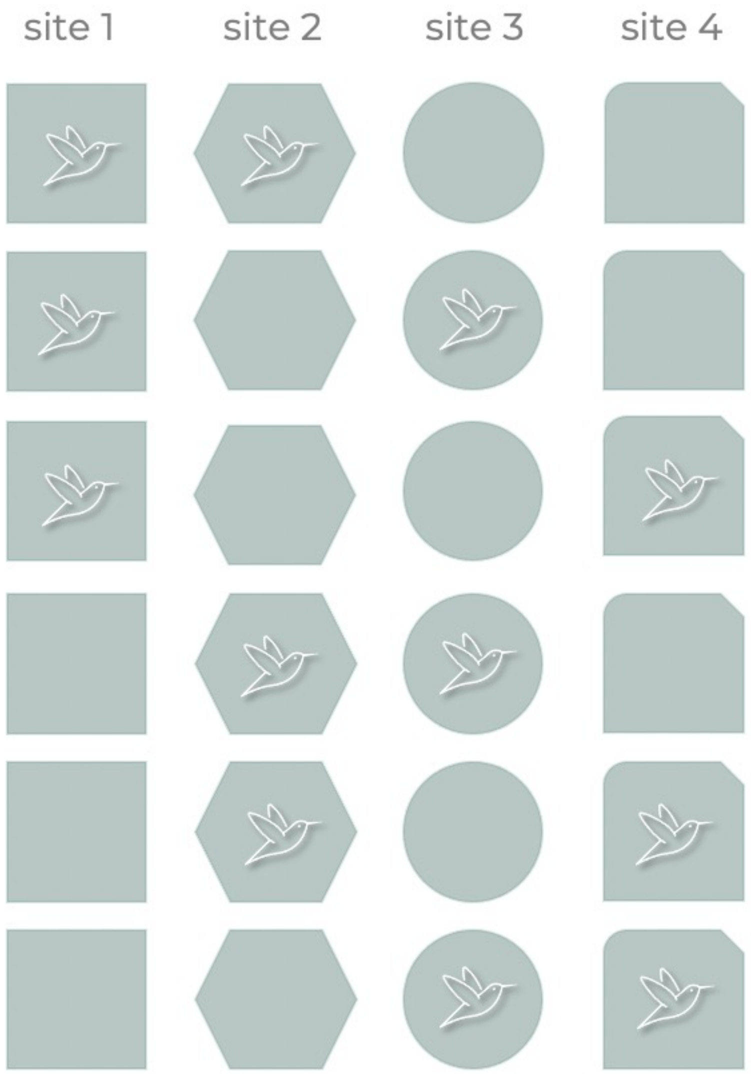
# Rewriting the numerator

total # of ways species 1 and 2 could be arranged among $N$ sites for a given $N_1$, $N_2$, and $j$ =

$$\binom{N}{j} \times \binom{N-j}{N_2-j} \times \binom{N-N_2}{N_1-j}$$

# THE DENOMINATOR

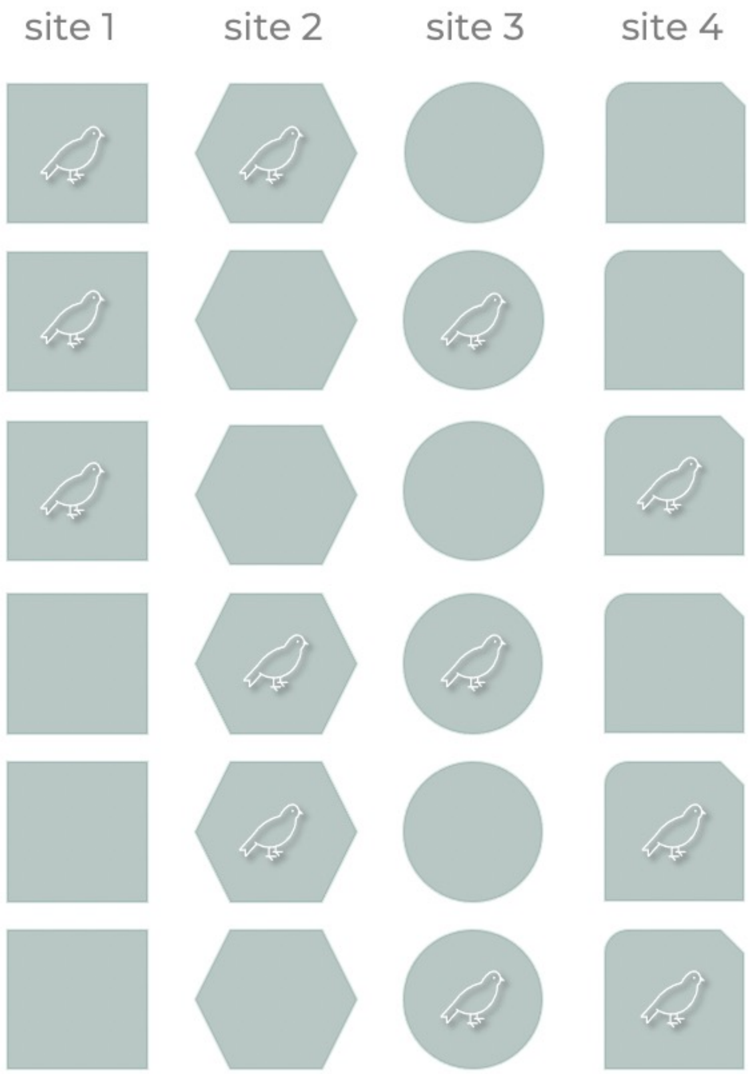
= species 1 = species 2 $N = 4$ $N_1 = 2$ $N_2 = 2$




$$\binom{N}{N_1} = \binom{4}{2} = 6$$

$$\binom{N}{N_2} = \binom{4}{2} = 6$$

$$6 \times 6 = 36$$



# Rewriting the denominator

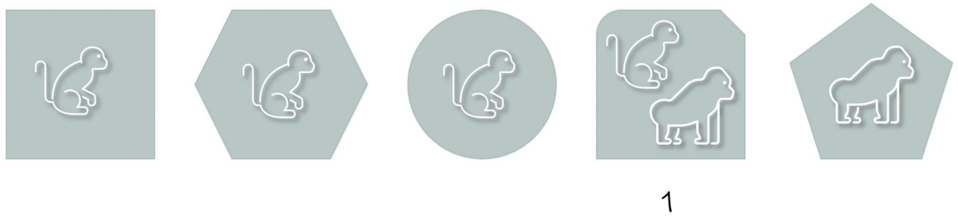
total # of ways species 1 and 2 could be arranged among N sites for a given N1 and N2, without considering j

$$= \binom{N}{N_2} \times \binom{N}{N_1}$$

# Putting it all together

$$\underbrace{p_j}_{\substack{\text{probability that species 1} \\ \text{and 2 co-occur at exactly} \\ j \text{ sites}}} = \frac{\overbrace{\binom{N}{j}}^{\substack{\text{\# of ways } j \text{ sites can be} \\ \text{arranged among } N \text{ sites}}} \times \overbrace{\binom{N-j}{N_2-j}}^{\substack{\text{\# of ways species 2 can} \\ \text{be arranged in sites that} \\ \text{don't have both species}}} \times \overbrace{\binom{N-N_2}{N_1-j}}^{\substack{\text{\# of ways species 1 can be} \\ \text{arranged among sites that} \\ \text{don't have species 2}}}}{\underbrace{\binom{N}{N_2}}_{\substack{\text{\# of ways species 2} \\ \text{can be arranged}}} \times \underbrace{\binom{N}{N_1}}_{\substack{\text{\# of ways species 1} \\ \text{can be arranged}}}}$$

# Not all values of j make sense…




$$\max\{0, N_1 + N_2 - N\} \le j$$


$$j \le \min\{N_1, N_2\}$$

# AN EXAMPLE

Imagine we've sampled 40 sites and found two llama species co-occur at 4 sites. Species 1 is present at 10 sites and species 2 at 25 sites. Do these species occur more, or less, frequently than expected by chance?

$N$ — **40** sites

$N_1$ — **10** sites occupied by species 1

$N_2$ — **25** sites occupied by species 2

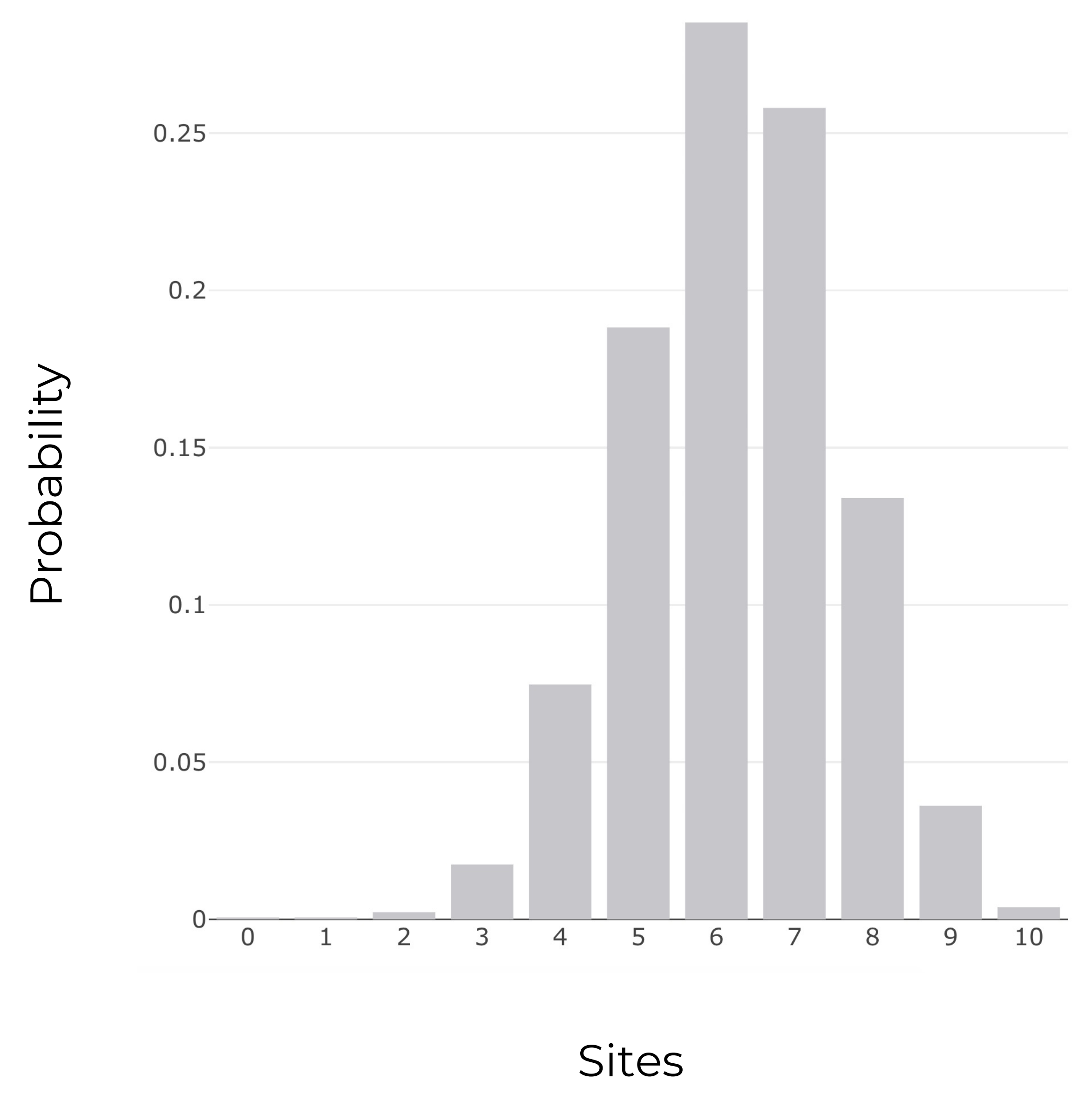
$Q_{obs}$ — **4** sites with both species

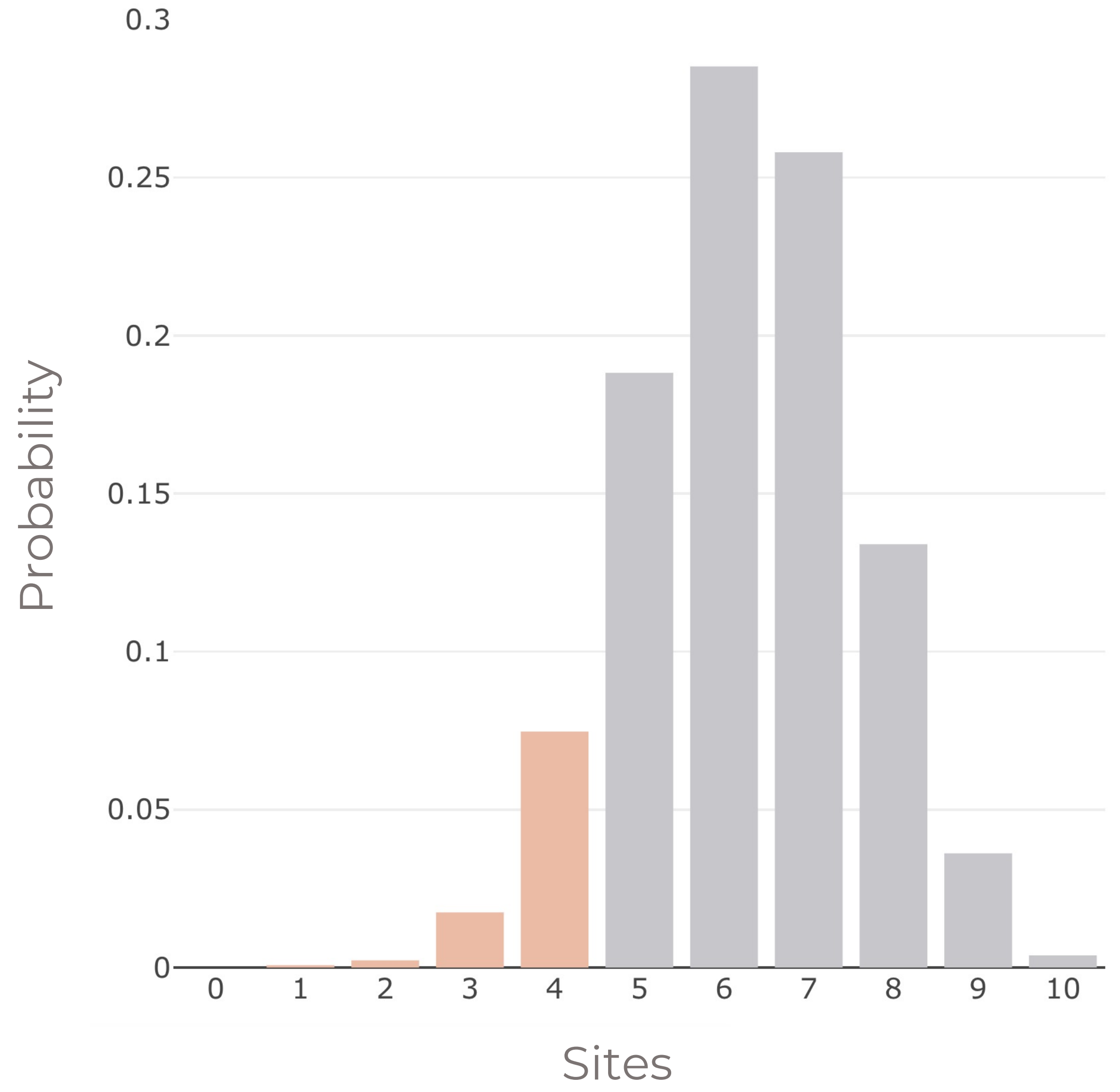
# Code along in RStudio

```
# Define the number of sites.
N = 40
# Define the number of sites occupied by species 1.
N1 = 10
# Define the number of sites occupied by species 2.
N2 = 25
# Number of sites species 1 and 2 co-occur at.
j = 0:10

# Calculate probability that species 1 and 2 occur at exactly j sites.
pj <- choose(N, j) * choose(N - j, N2 - j) * choose(N - N2, N1 - j)/
  (choose(N, N2) * choose(N, N1))

# Construct data frame of j, pj, and cumulative sum of pj.
co <- data.frame(j, Pj = pj, sumPj = cumsum(pj))
```

| | j | Pj | sumPj |
|---|---|---|---|
| 1 | 0 | 3.542692e-06 | 3.542692e-06 |
| 2 | 1 | 1.476122e-04 | 1.511549e-04 |
| 3 | 2 | 2.277445e-03 | 2.428600e-03 |
| 4 | 3 | 1.746041e-02 | 1.988901e-02 |
| 5 | 4 | 7.469175e-02 | 9.458076e-02 |
| 6 | 5 | 1.882232e-01 | 2.828040e-01 |
| 7 | 6 | 2.851867e-01 | 5.679907e-01 |
| 8 | 7 | 2.580261e-01 | 8.260167e-01 |
| 9 | 8 | 1.339751e-01 | 9.599918e-01 |
| 10 | 9 | 3.615200e-02 | 9.961438e-01 |
| 11 | 10 | 3.856214e-03 | 1.000000e+00 |

```r
library(plotly)
# Plot pj distribution.
plot_ly(x = as.factor(co$j), y = co$Pj,
        type = "bar", color = I("#C7C6CA"))
```



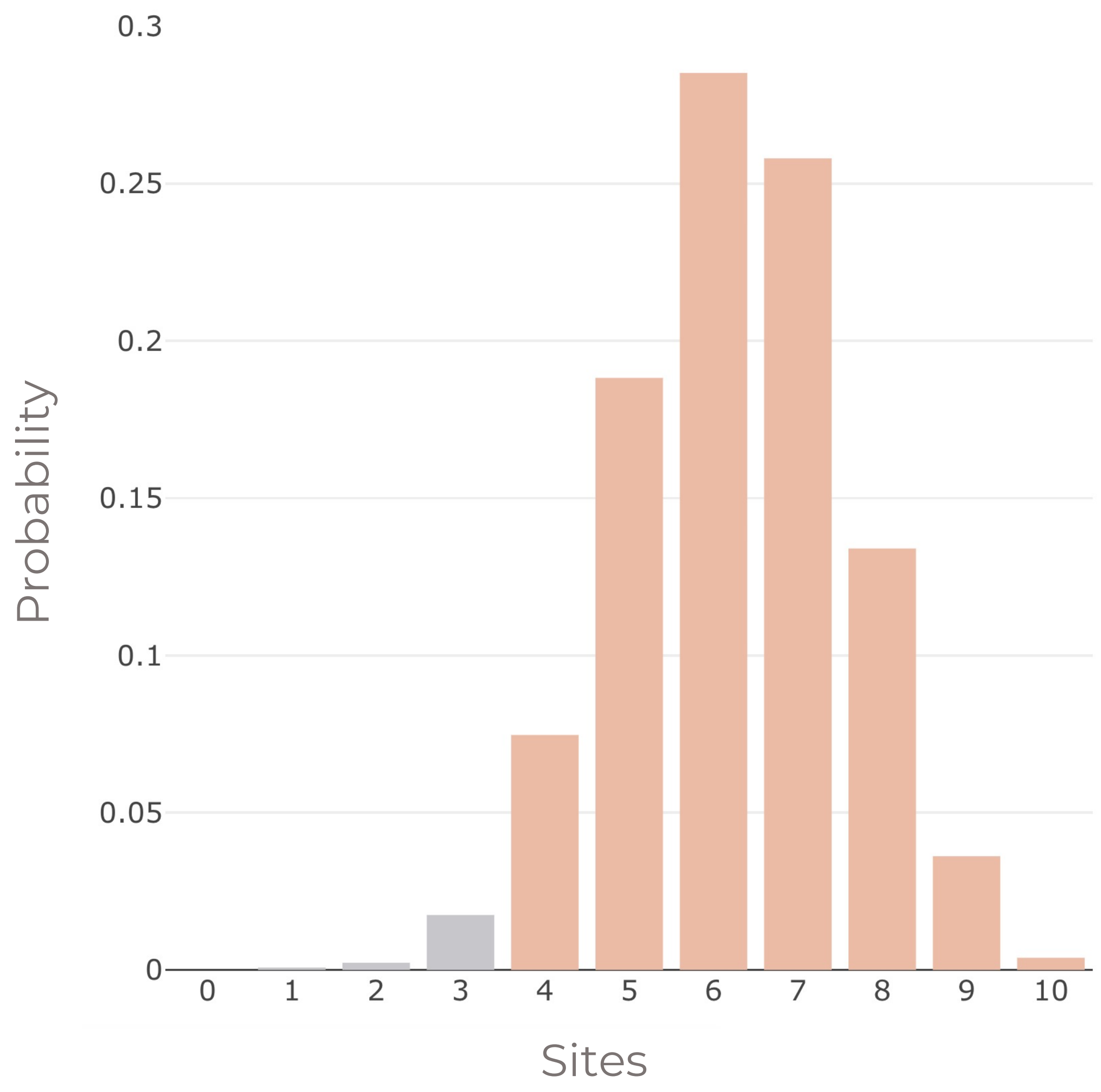


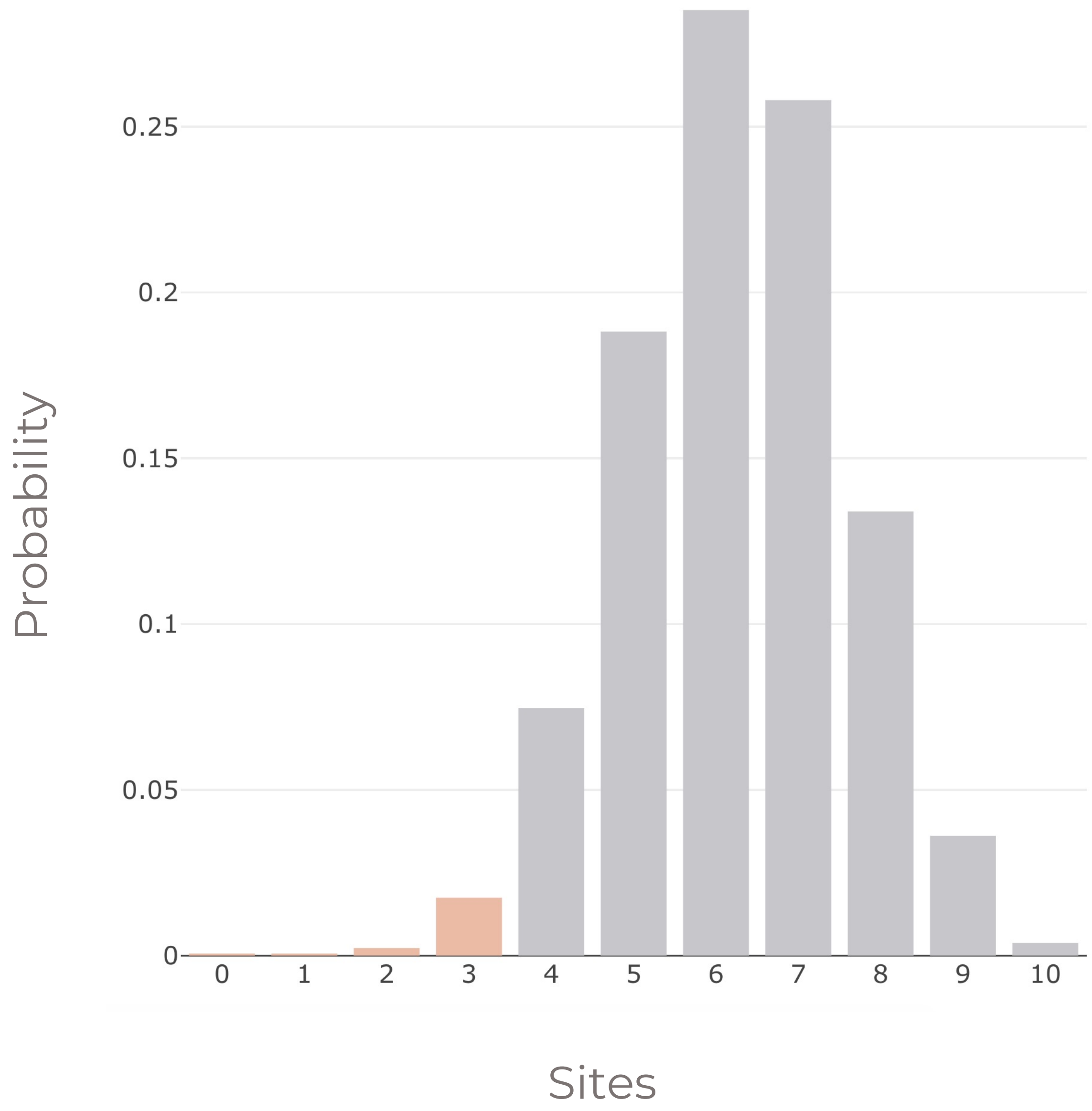

# Less than expected

$$\sum_{0}^{4} p_j \nless 0.05$$

# More than expected

$$\sum_{4}^{10} p_j \not< 0.05$$



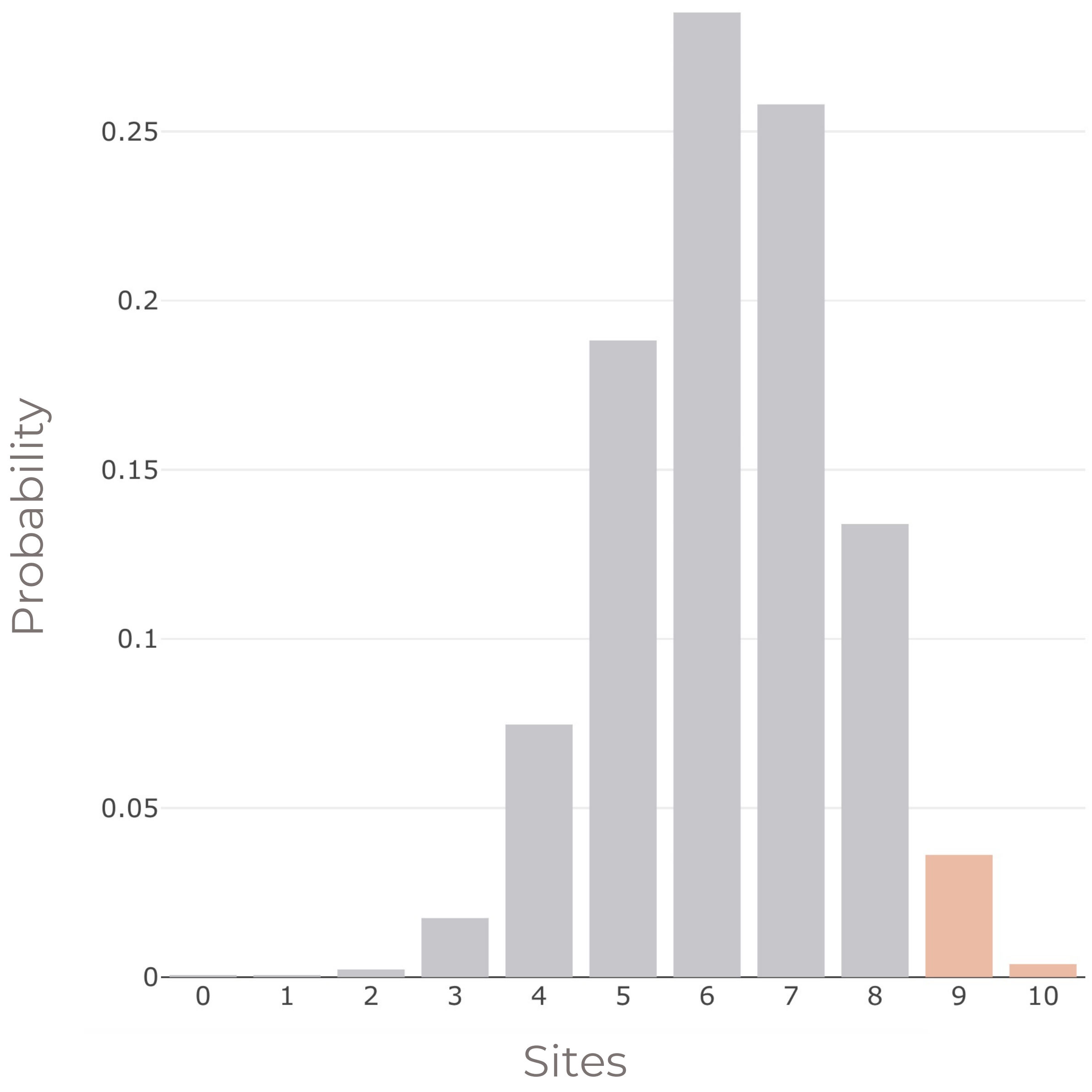



# Less than expected

$$\sum_{j=\max\{0,\, N_1+N_2-N\}}^{Q_{obs}} p_j < 0.05$$

# More than expected

$$\sum_{j = Q_{obs}}^{\min(N_1, N_2)} p_j < 0.05$$



Journal of Statistical Software
February 2016, Volume 69, Code Snippet 2. doi: 10.18637/jss.v069.c02

# cooccur: Probabilistic Species Co-Occurrence Analysis in R

**Daniel M. Griffith**
Wake Forest University

**Joseph A. Veech**
Texas State University

**Charles J. Marsh**
University of Leeds

**Abstract**

The observation that species may be positively or negatively associated with each other is at least as old as the debate surrounding the nature of community structure which began in the early 1900's with Gleason and Clements. Since then investigating species co-occurrence patterns has taken a central role in understanding the causes and consequences of evolution, history, coexistence mechanisms, competition, and environment for community structure and assembly. This is because co-occurrence among species is a measurable metric in community datasets that, in the context of phylogeny, geography, traits, and environment, can sometimes indicate the degree of competition, displacement, and phylogenetic repulsion as weighed against biotic and environmental effects promoting correlated species distributions. Historically, a multitude of different co-occurrence metrics have been developed and most have depended on data randomization procedures to produce null distributions for significance testing. Here we improve upon and present an R implementation of a recently published model that is metric-free, distribution-free, and randomization-free. The R package, **cooccur**, is highly accessible, easily integrates into common analyses, and handles large datasets with high performance. In the article we develop the package's functionality and demonstrate aspects of co-occurrence analysis using three sample datasets.

*Keywords*: co-occurrence, species niche, community ecology.

## 1. Introduction

The analysis of species co-occurrence patterns is a fundamental task for many ecological investigations. Species coexistence, community structure and assembly, and the maintenance of biodiversity are all essentially founded on the ways in which species co-occur with one another. Even the very early Clementsian and Gleasonian perspectives on the organization of plant communities can be put in the context of species co-occurrence (Hoagland and Collins

THAT darn DATA


NATURE NEEDS DATA SCIENCE
CODE FOR THE CAUSE


NATURE NEEDS DATA SCIENCE
CODE FOR THE CAUSE


NATURE NEEDS DATA SCIENCE
CODE FOR THE CAUSE