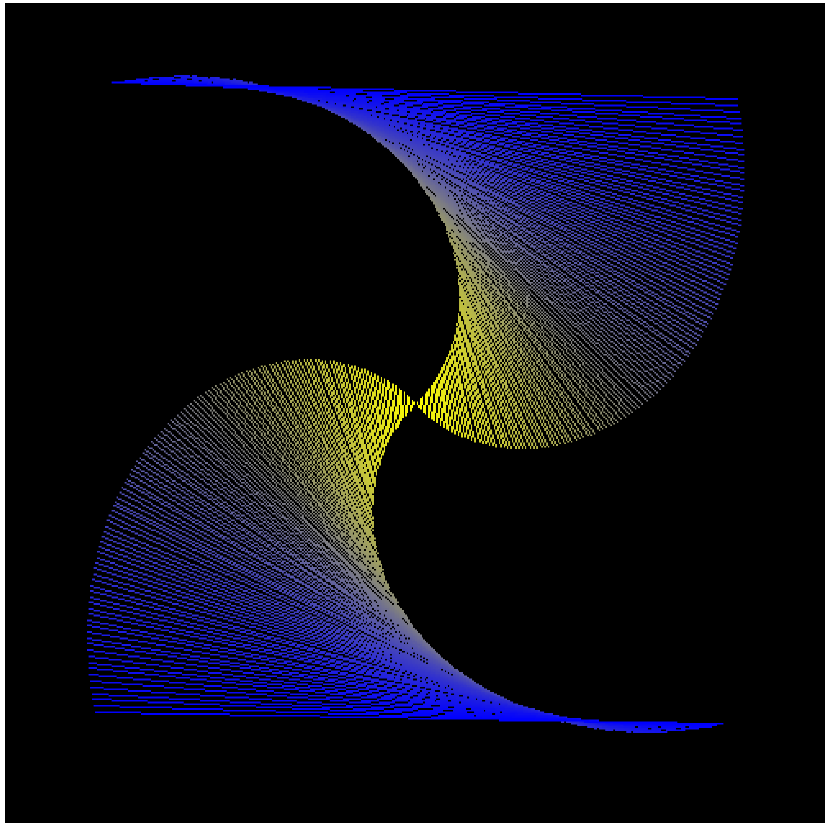
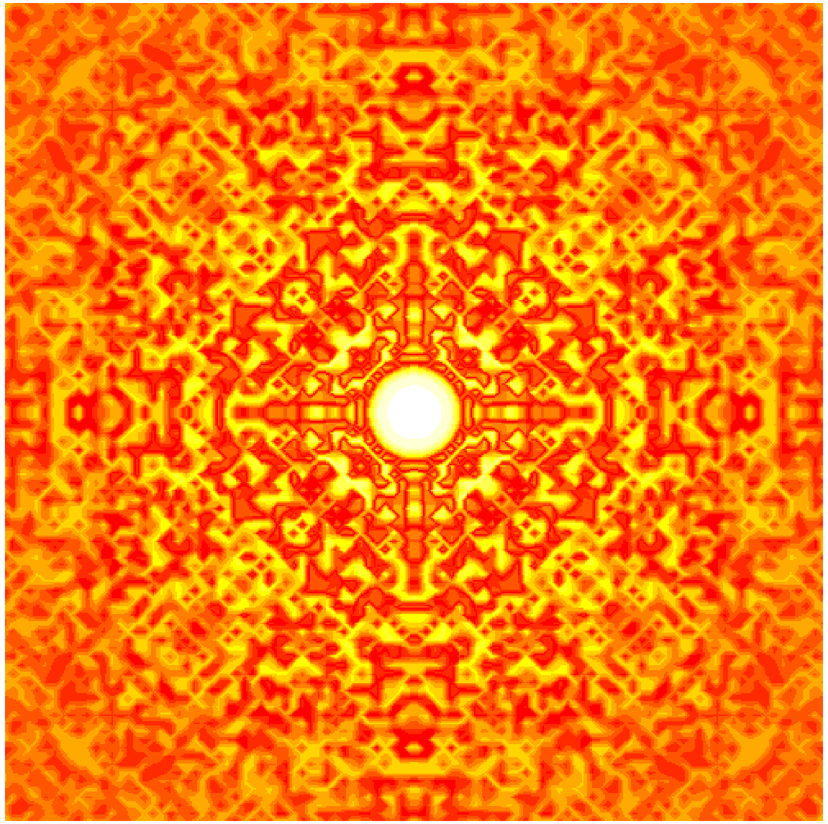
So you've got lots of zeros



Jonathan Coop

Statistics Bootcamp

02.10.21



The plan and a disclaimer

- Think about why a dataset might have a lot of zeros and some different approaches to modeling with zeros.
- Analyze some of my data on tree seedling counts using generalized linear models.
- Analyze some of your data.

* I am not a statistician and everything I know is what I learned is through my own struggles with data, and lots of google searches.

Why might a dataset have lots of zeros?

Categorical (binomial vs. multinomial, ordinal (ranked) vs. nominal (unranked)), **discrete**, and **continuous** variables.

- We often use zeros vs. ones to represent two alternative outcomes of a binary outcome (e.g., 0 = absent, 1 = present). Here, a lot of zeros might be expected.
- Count data: typically, we end up zeros in our response variable when we are counting things that are not common or happen infrequently.
- Zero might be the median value of a continuous variable that spans a negative and positive range. E.g., Palmer Drought Severity, ENSO index – normal distribution?

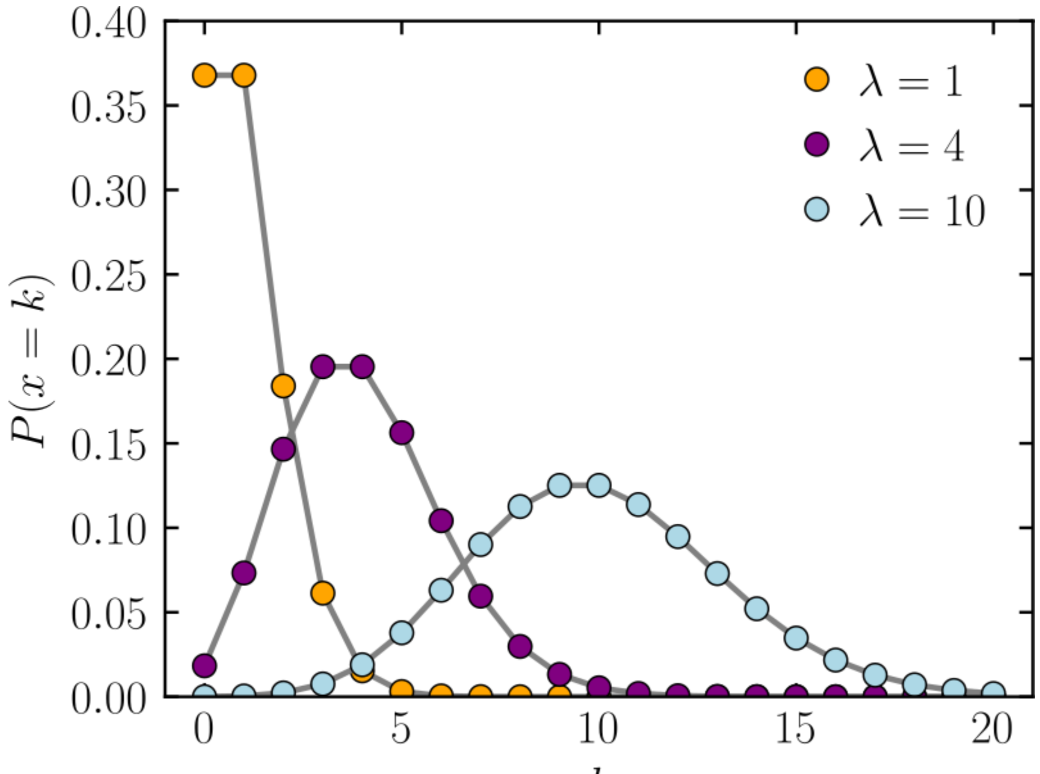
Count data with lots of zeros

- For a range of statistical approaches, zeros can often be dealt with via non-parametric or zero-friendly alternatives (e.g., Mann-Whitney U, Wilcoxon signed rank sum, Kruskal Wallis, Fisher's exact test).
- Assuming we are interested in a linear model where y = f(x). For a normal distribution, a linear model might be y = mx + b.
- Count data are unlikely to be normally distributed, so we need to think about other types ("families") of distributions for generalized linear models.

Count data with lots of zeros

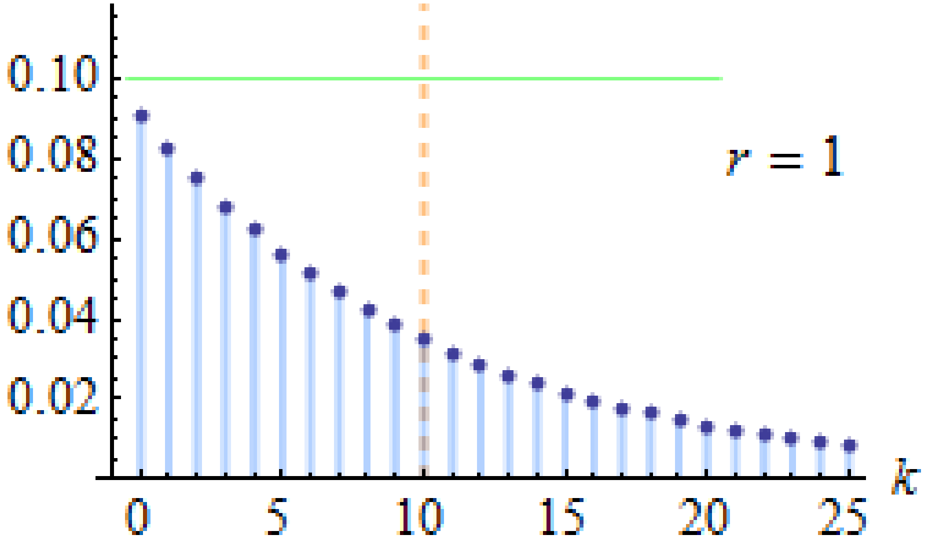
Poisson distribution: number of events occurring within a discrete interval of time or space

- Events are independent.
- λ is expected rate of occurrence (mean)
- Variance of a Poisson distribution is also λ.



Count data with lots of zeros

- Negative binomial distribution: number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures occurs.
- Similar to Poisson, but allows for overdispersion (variance not necessarily λ) – this would be expected in a dataset with many zeros.
- Mean = μ; variance = μ(1 + μ/Θ) -- overdispersion is modeled with its own parameter Θ.



Count data with lots of zeros

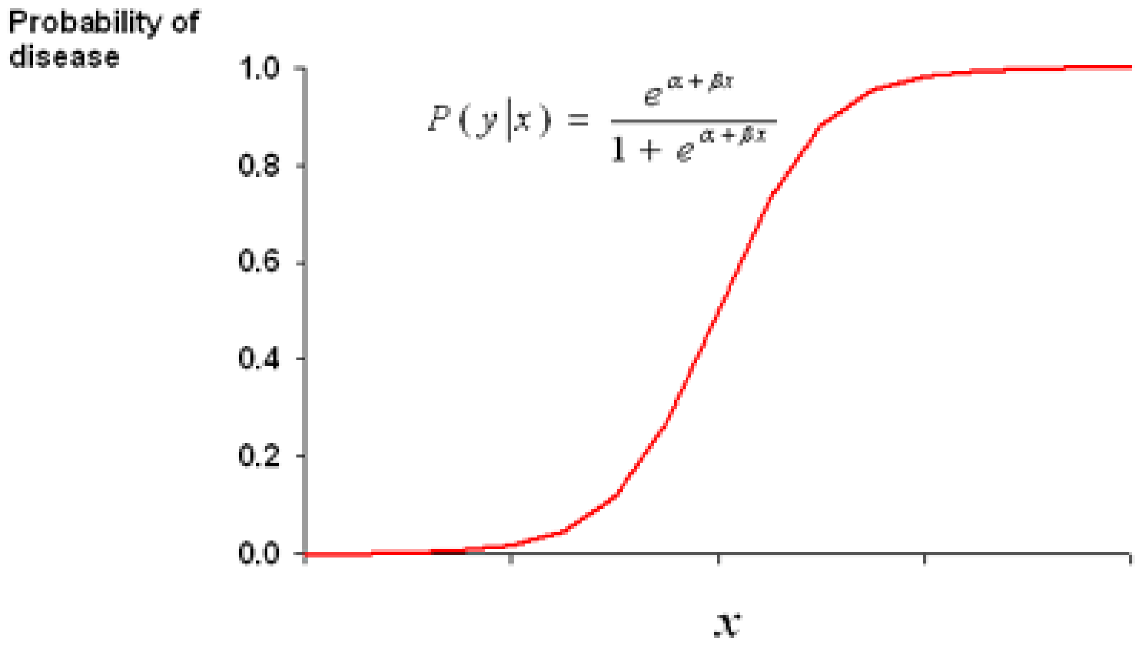
- Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) models.
- For use where zero-inflated, and zero-inflated & overdispersed distributions.
- Basically employs two models at the same time:

1. Model zeros vs. non-zero observations.
2. Model non-zero observations (Poisson or NB).

What kinds of processes might give rise to zero-inflated distributions?

Count data with lots of zeros

- Binomial logistic model – two exclusive categories that can be represented as 0 or 1.
- Sometimes it is just easier to collapse a complicated distribution into a simple one, e.g. presence/absence, or occurrence > some critical minimum threshold.



Tree seedlings in post-fire landscapes

- We counted tree seedlings in 686, 100-m^2 plots in 12 burns in the W US.
- To what extent do fire refugia (unburned patches of forest) promote post-fire forest recovery?

esa ECOSPHERE

Contributions of fire refugia to resilient ponderosa pine and dry mixed-conifer forest landscapes

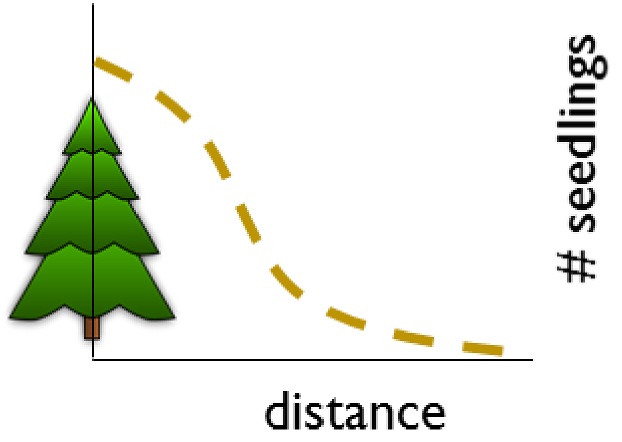
JONATHAN D. COOP,[1,]† TIMOTHY J. DELORY,[2] WILLIAM M. DOWNING,[3] SANDRA L. HAIRE,[4] MEG A. KRAWCHUK,[3] CAROL MILLER,[5] MARC-ANDRÉ PARISIEN,[6] AND RYAN B. WALKER[1]

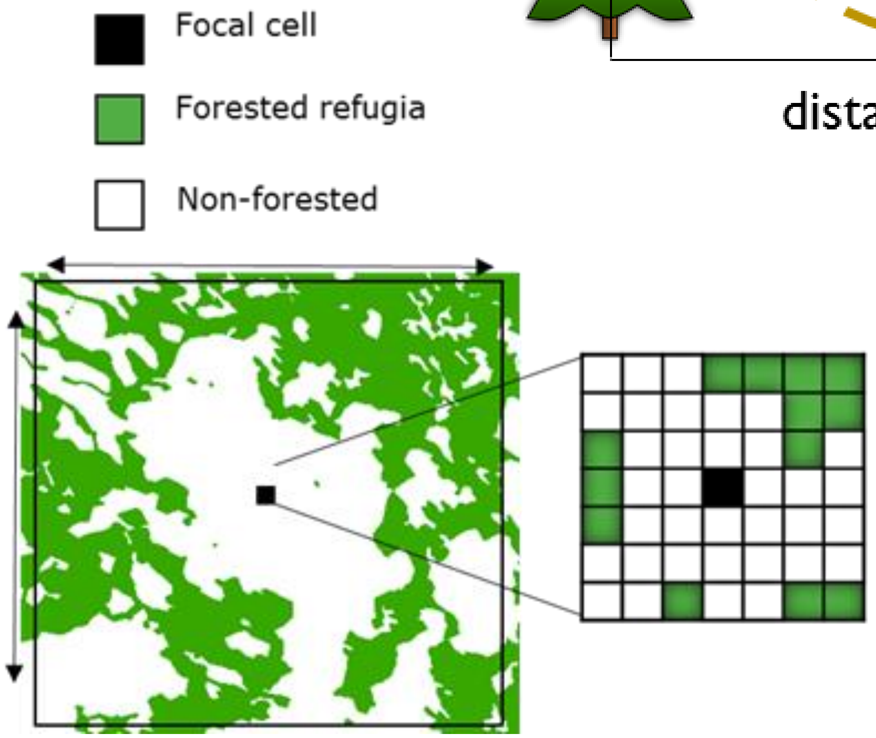
[1]School of Environment and Sustainability, Western Colorado University, Gunnison, Colorado 81231 USA
[2]Department of Biology, Utah State University, Logan, Utah 84322 USA
[3]Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon 97331 USA
[4]Haire Laboratory for Landscape Ecology, Belfast, Maine 04915 USA
[5]Aldo Leopold Wilderness Research Institute, Rocky Mountain Research Station, USDA Forest Service, Missoula, Montana 59801 USA
[6]Northern Forestry Centre, Canadian Forest Service, Natural Resources Canada, Edmonton, Alberta, Canada

Citation: Coop, J. D., T. J. DeLory, W. M. Downing, S. L. Haire, M. A. Krawchuk, C. Miller, M.-A. Parisien, and R. B. Walker. 2019. Contributions of fire refugia to resilient ponderosa pine and dry mixed-conifer forest landscapes. Ecosphere 10(7):e02809. 10.1002/ecs2.2809

SW_seedlings.csv

- Tree seedling counts in 368 plots from 8 SW burns along gradients of refugia proximity and abundance.
- Columns 1-5 are plot identifiers, coordinates.
- 6-10: elevation, topography.
- 11-15: measures of landscape pattern, burn severity. Distance is meters to surviving tree seed source. DWD is a composite measure of refugia proximity and abundance. D2WD is the same thing, but calculated with distances squared.





Distance-weighted density (m^{-1})

$$DWD = \sum_{i=1}^{N} 1/(d_i + 1)$$

SW_seedlings.csv

- 16-18: vegetation cover
- 19-41: measures of post-fire climate from ClimateNA (http://climatena.ca)
- 42-44: synthetic measures of climate (PCA of 19-41)
- 45-49: seedling counts for ponderosa pine (PIPO), Douglas-fire (PSME), white fir (ABCO), aspen (POTR), and combined piñon and juniper (PJ).

Questions:

- What kind of distribution do seedling counts follow?
- Which is the best model: poisson, negative binomial, or zero-inflated negative binomial? With or without a random effect term?
- **Which is a better predictor of PIPO seedling abundance: distance from seed source, DWD, or D2WD – or some combination?**