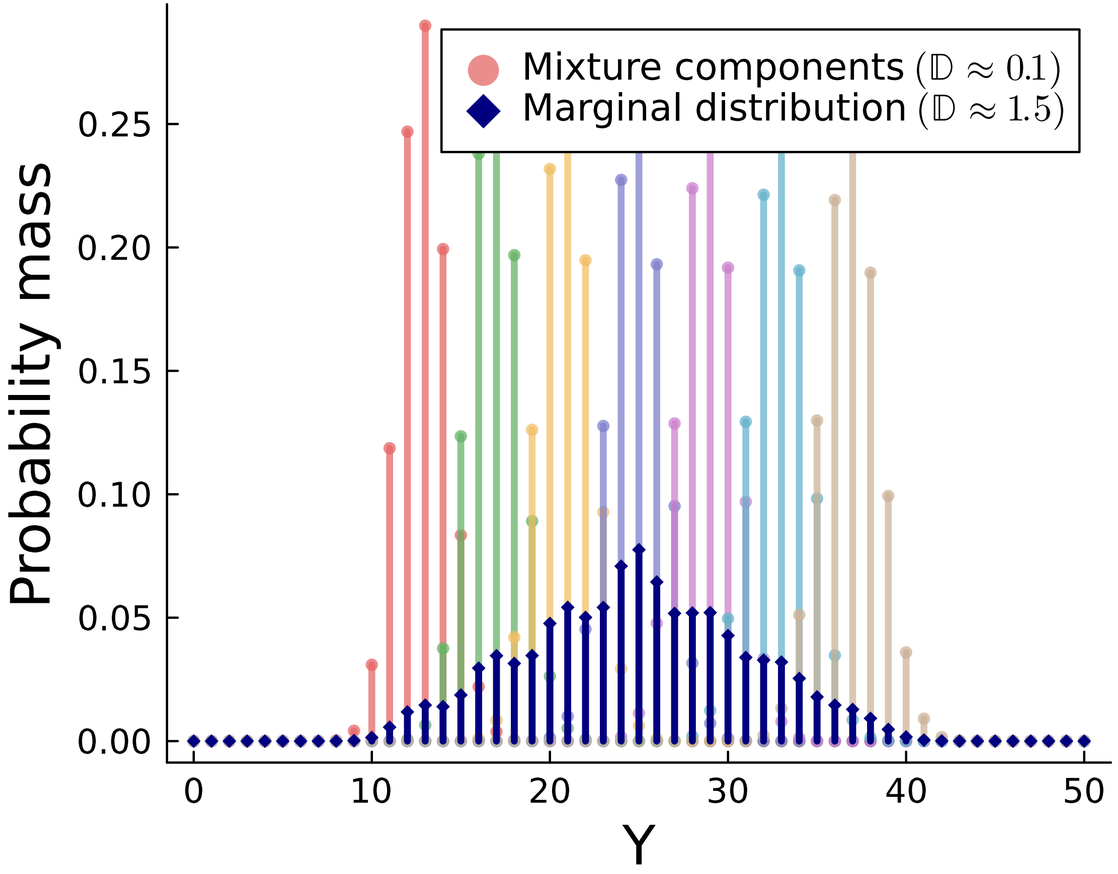
Modeling Latent Underdispersion with Discrete Order Statistics

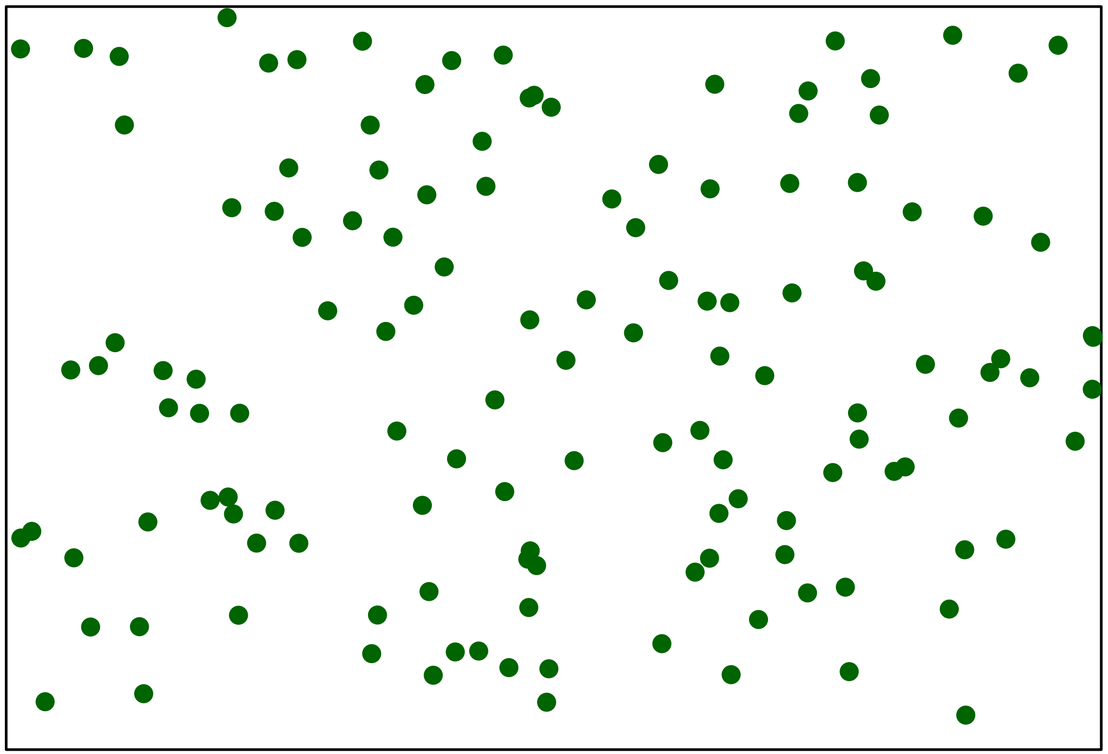


Jimmy Lederman, University of Chicago

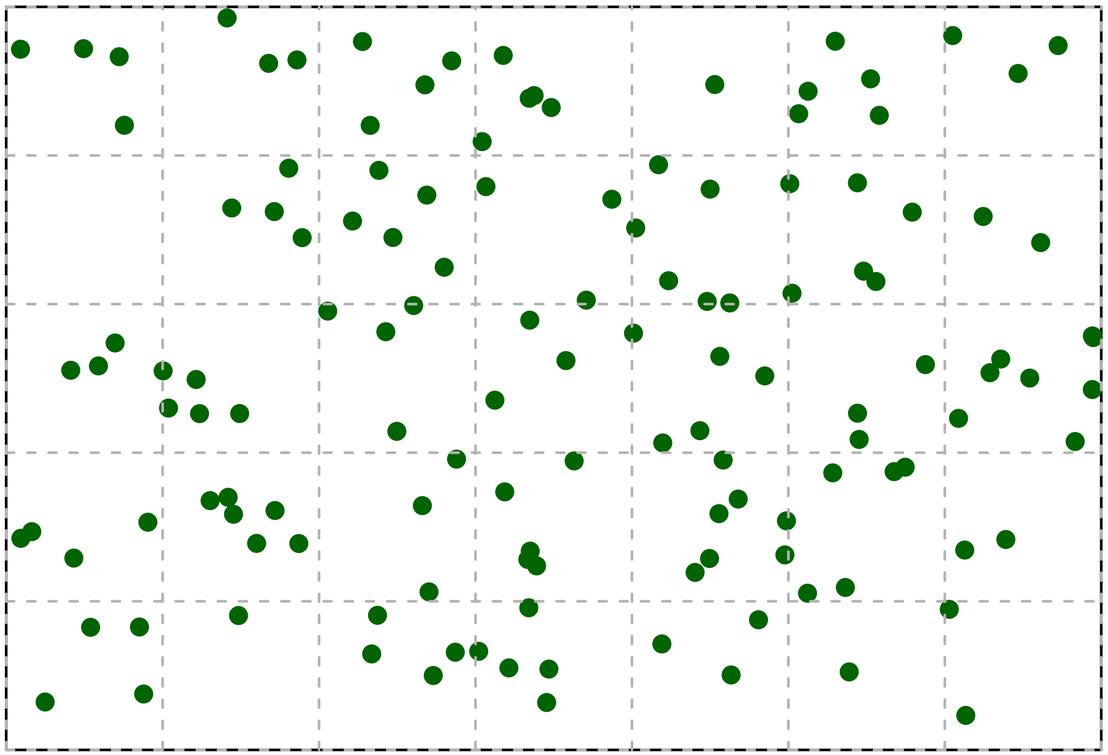
Work with Aaron Schein

Fast and Curious 2: MCMC in Action

The Poisson process

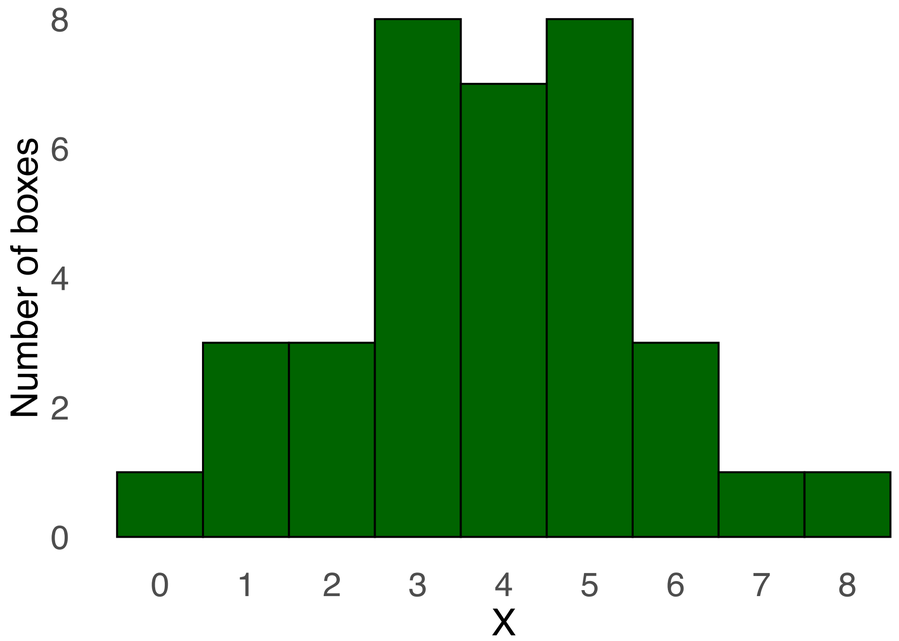


The Poisson process

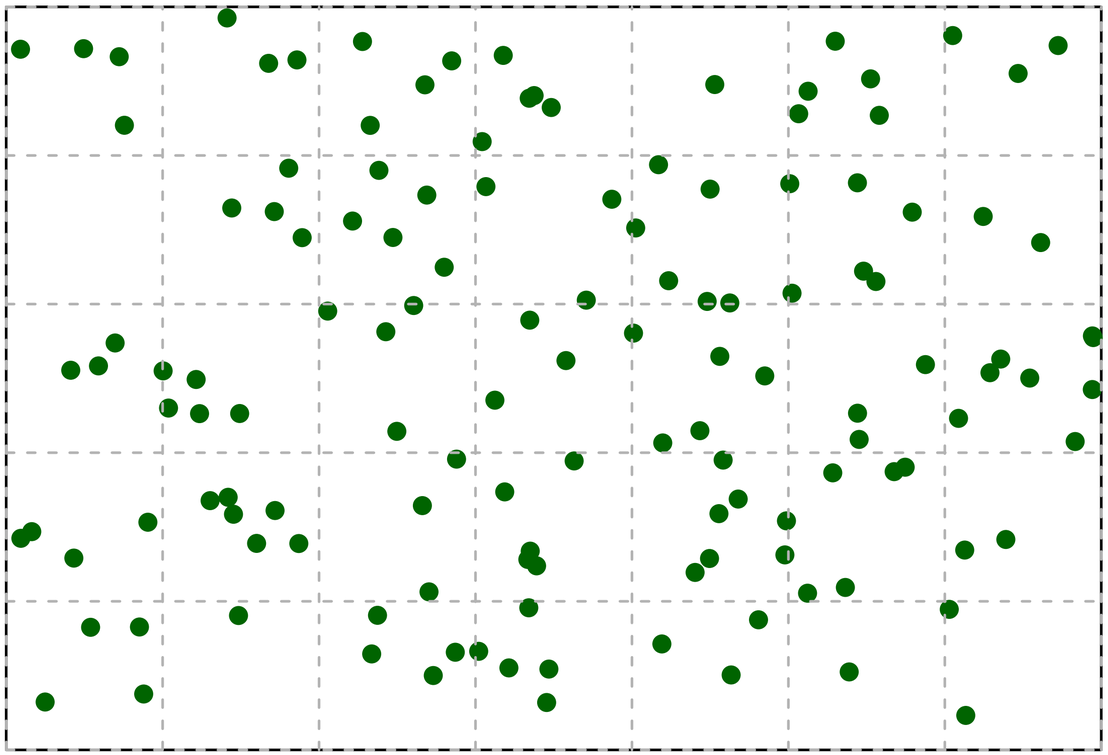


X = number of points in each box

$X \sim \text{Poisson}(\mu)$

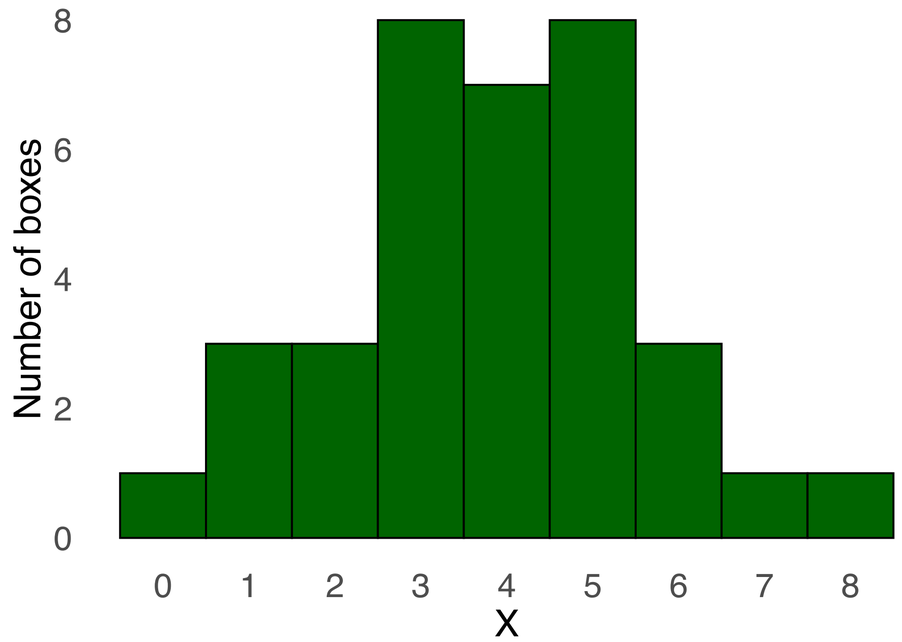


The Poisson process



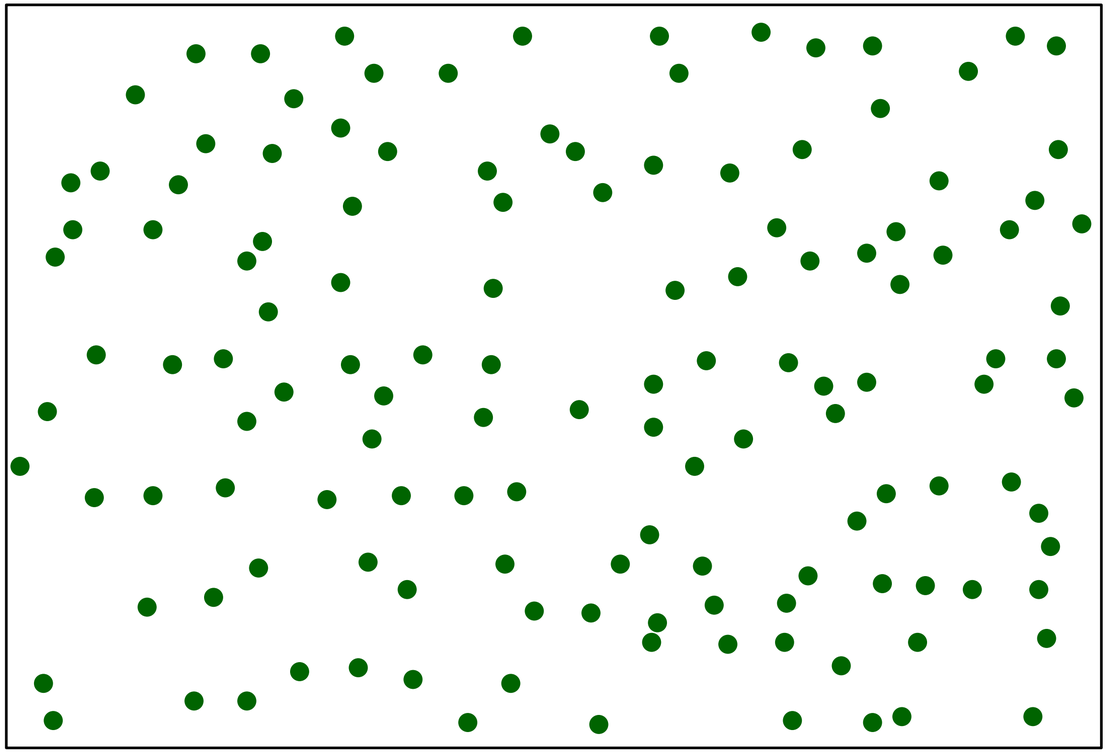
X = number of points in each box

$X \sim \text{Poisson}(\mu)$

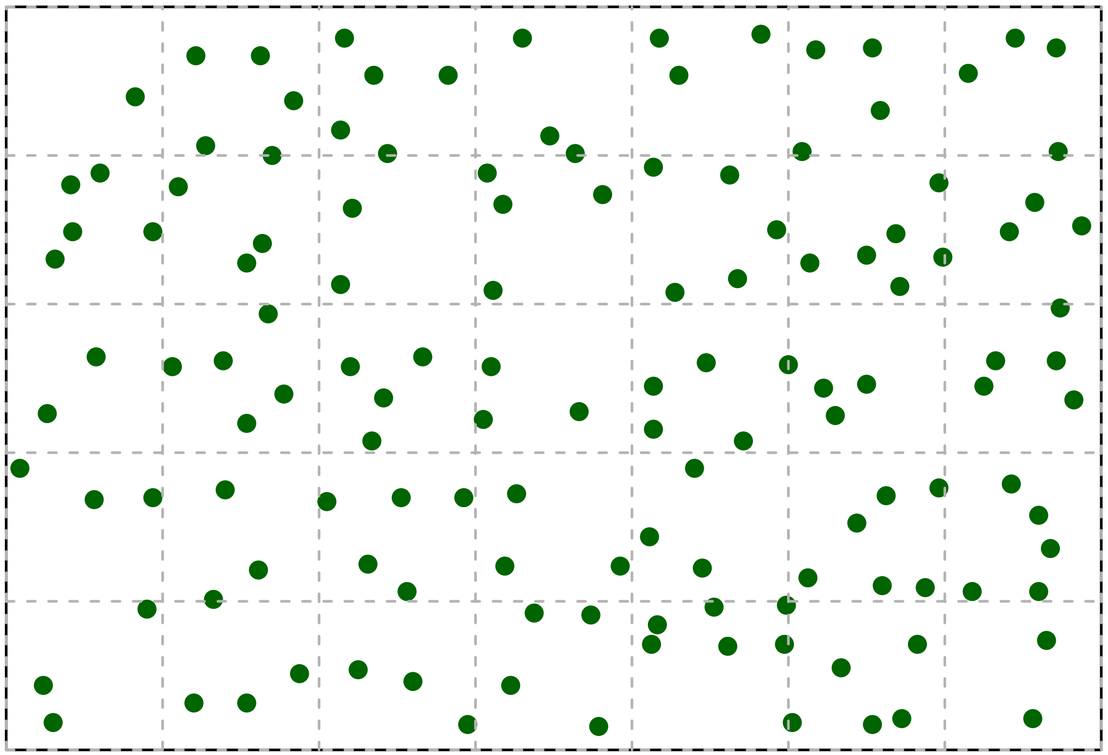


Dispersion $\longrightarrow$ $\mathbb{D}[X] = \dfrac{\mathbb{V}[X]}{\mathbb{E}[X]} = 1$

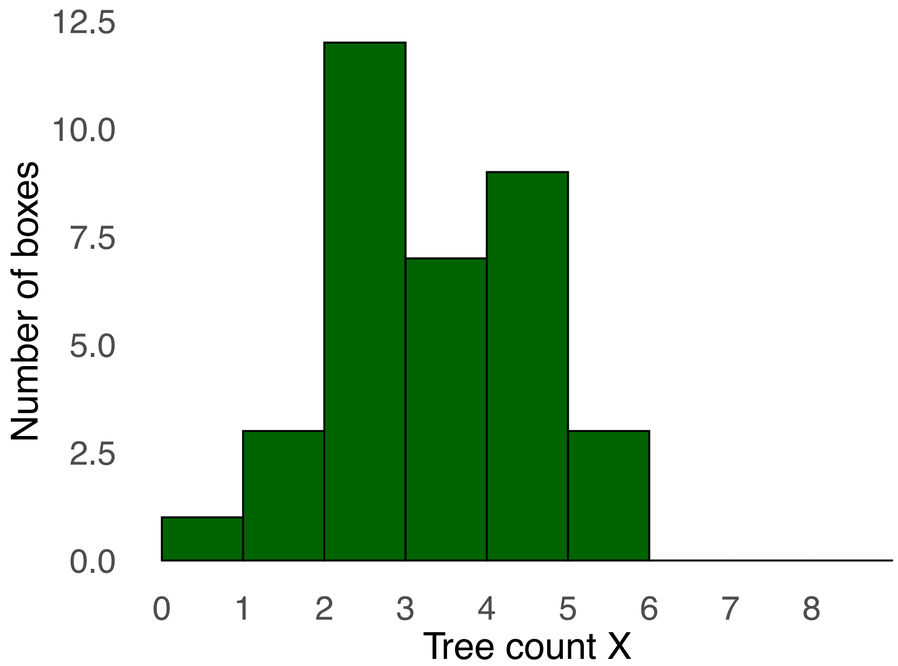
Underdispersion: Spruce trees



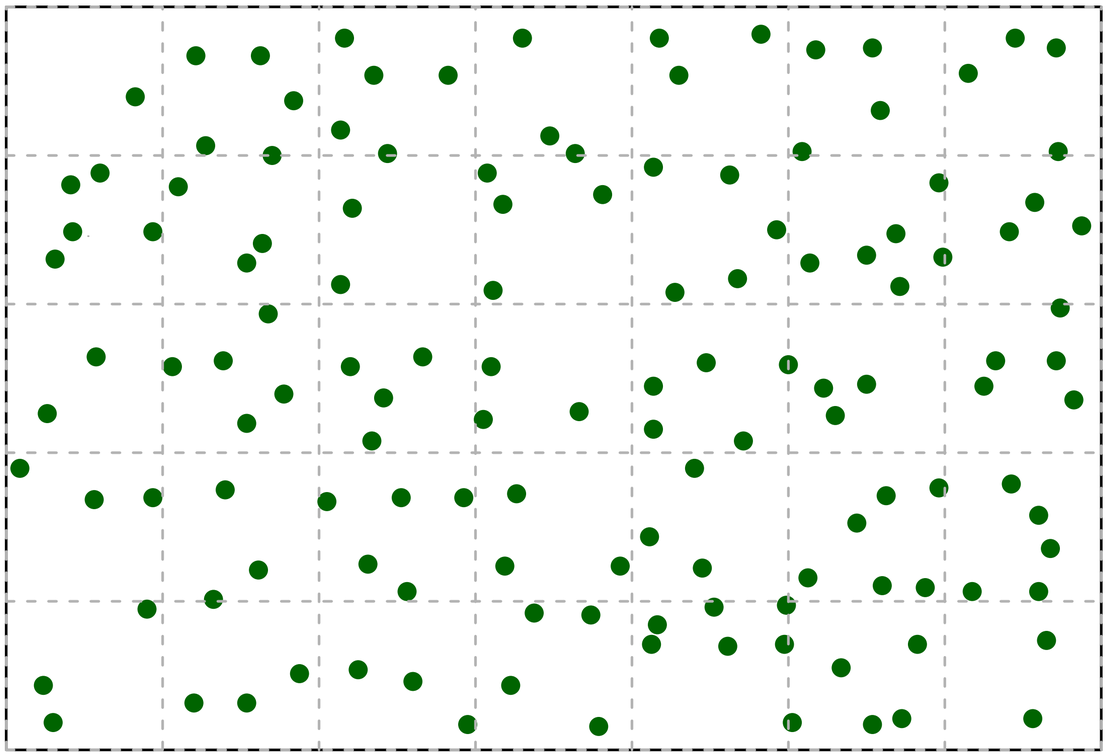
Underdispersion: Spruce trees



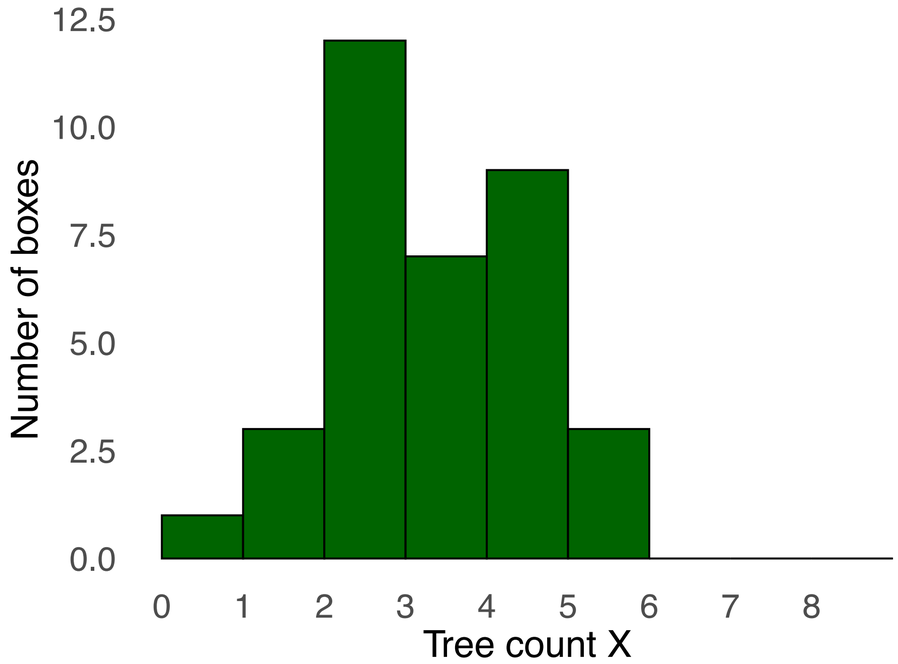
X = number of trees in each box



Underdispersion: Spruce trees

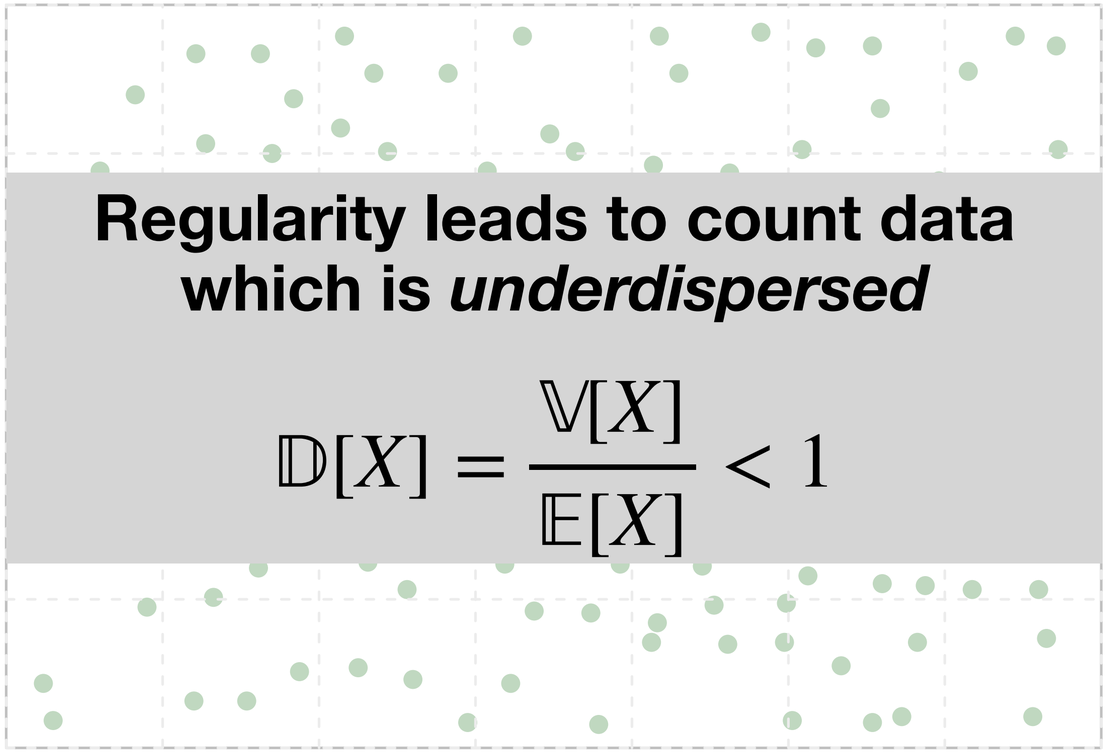


X = number of trees in each box



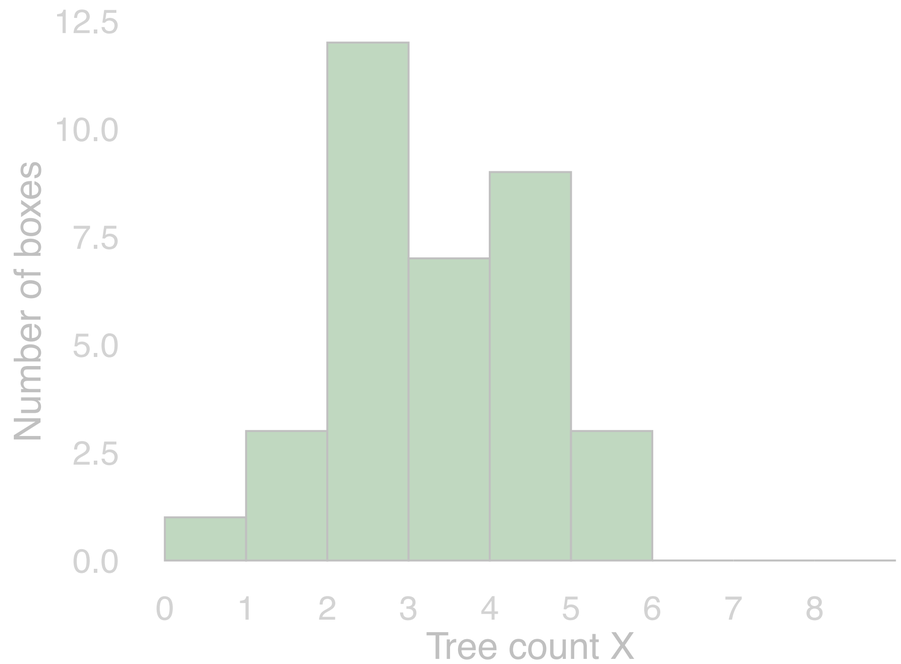
Dispersion $\longrightarrow$ $\mathbb{D}[X] = \dfrac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx .41$

Underdispersion: Spruce trees

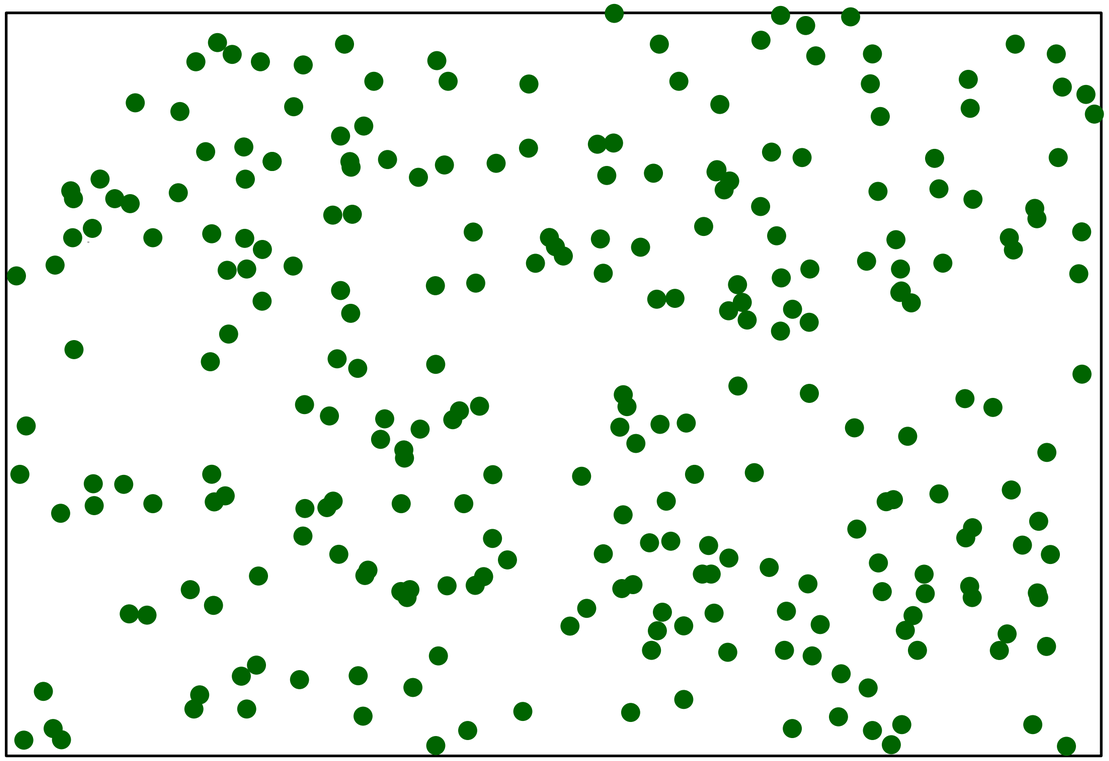


Regularity leads to count data which is *underdispersed*

$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} < 1$$

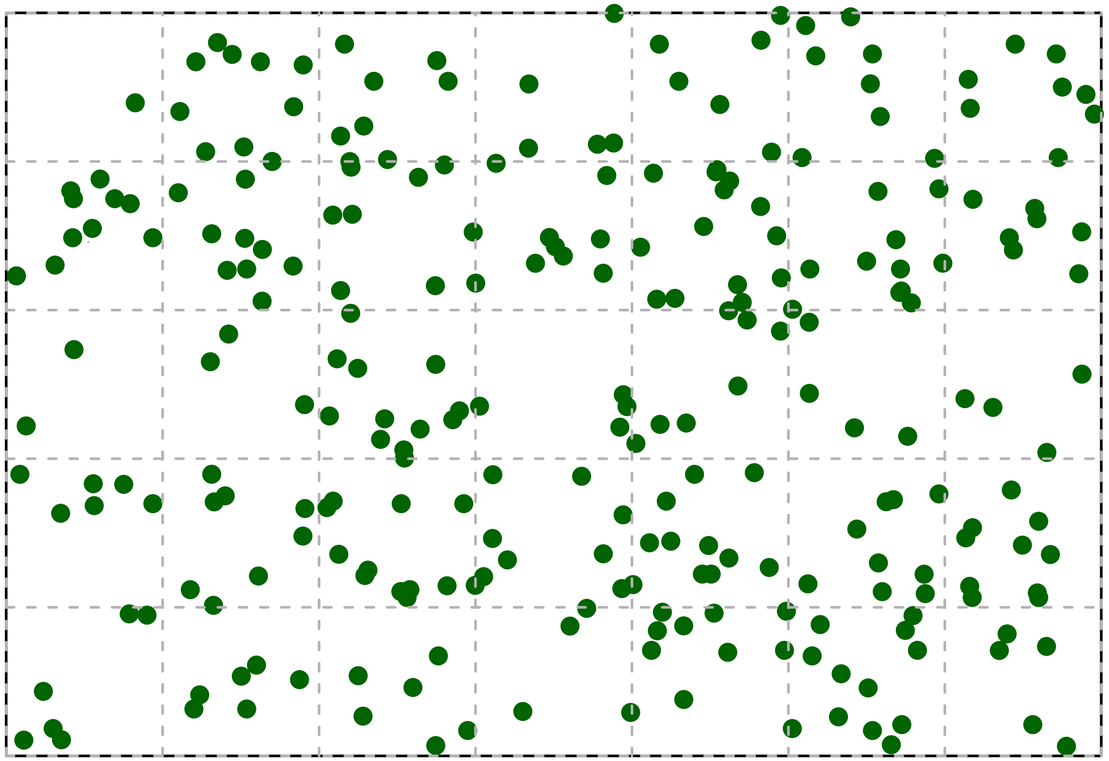


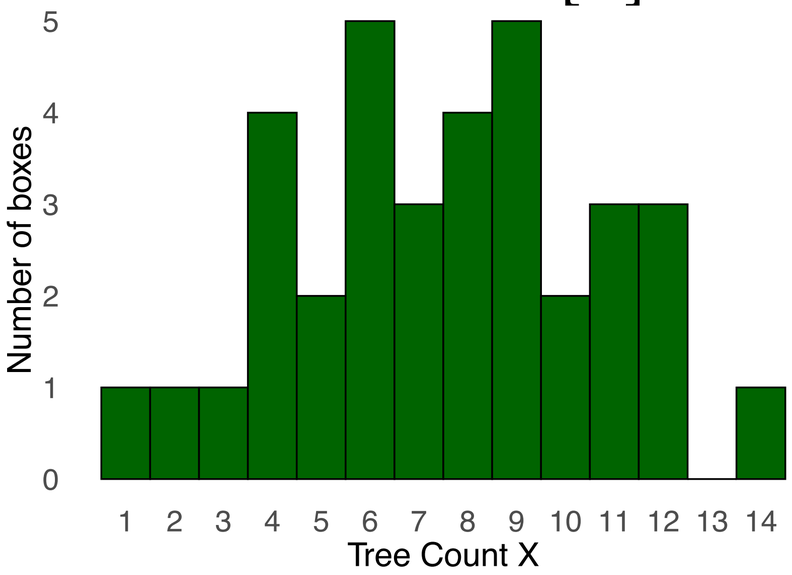
Conditional underdispersion



Conditional underdispersion

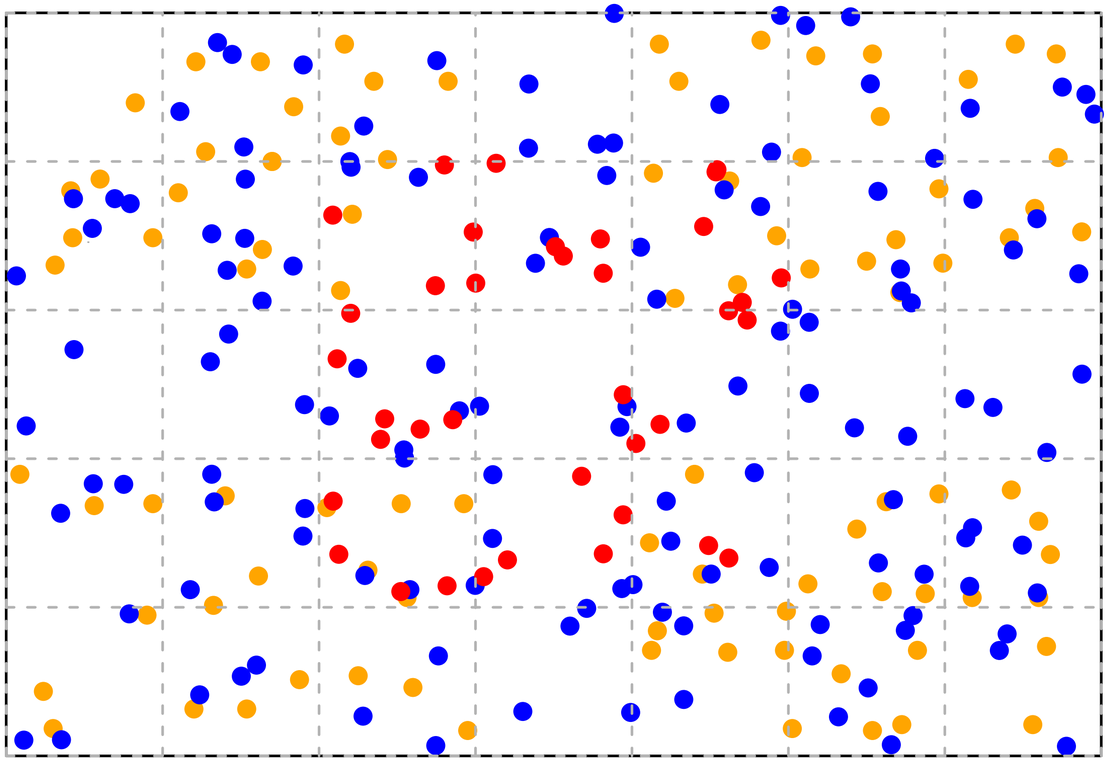
$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 1.29$$

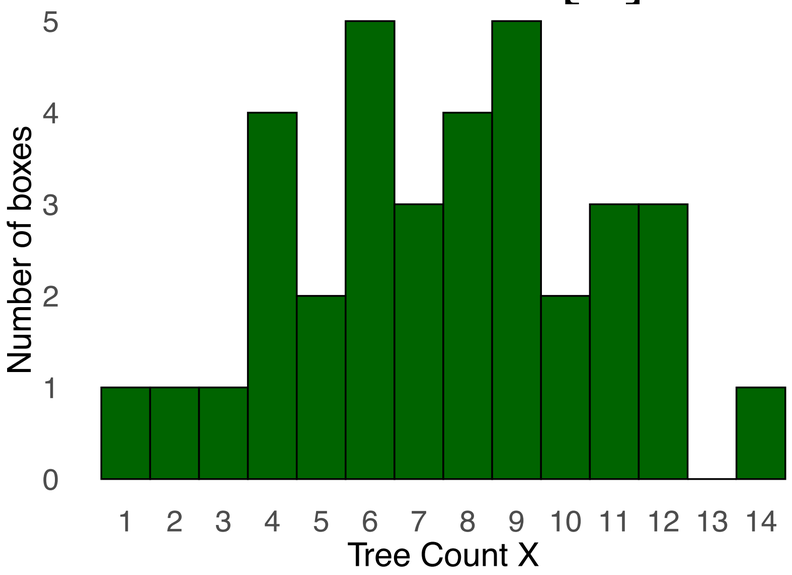




Conditional underdispersion

$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 1.29$$

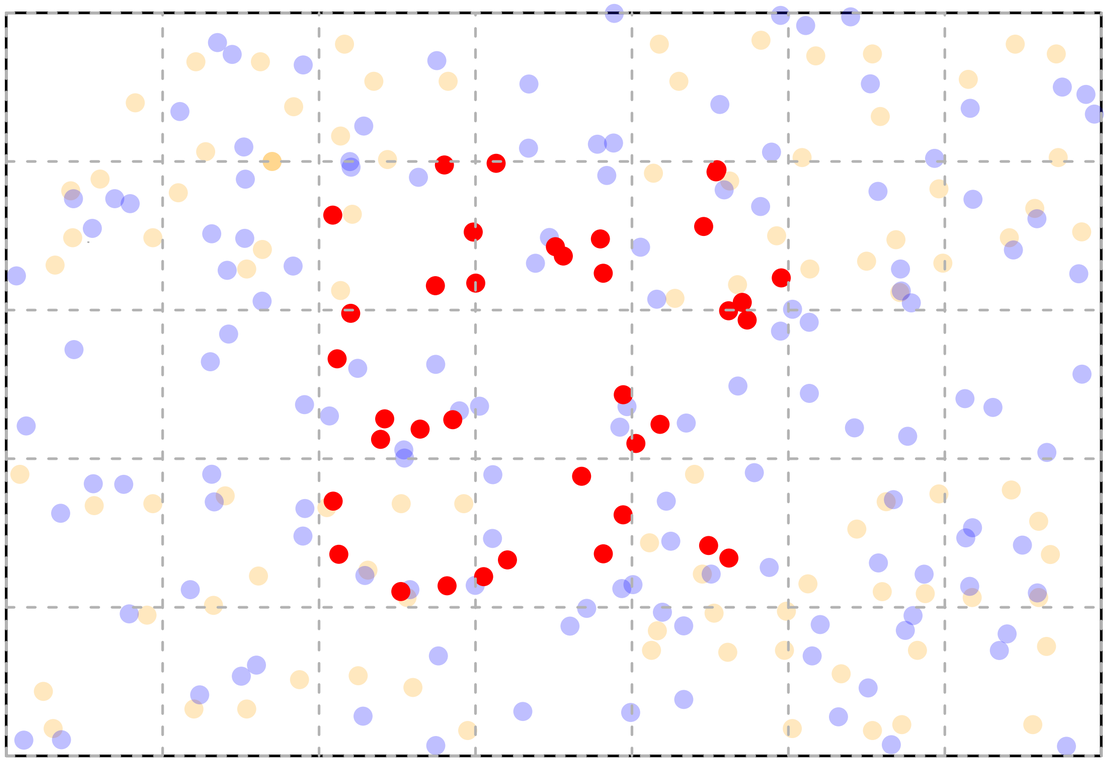


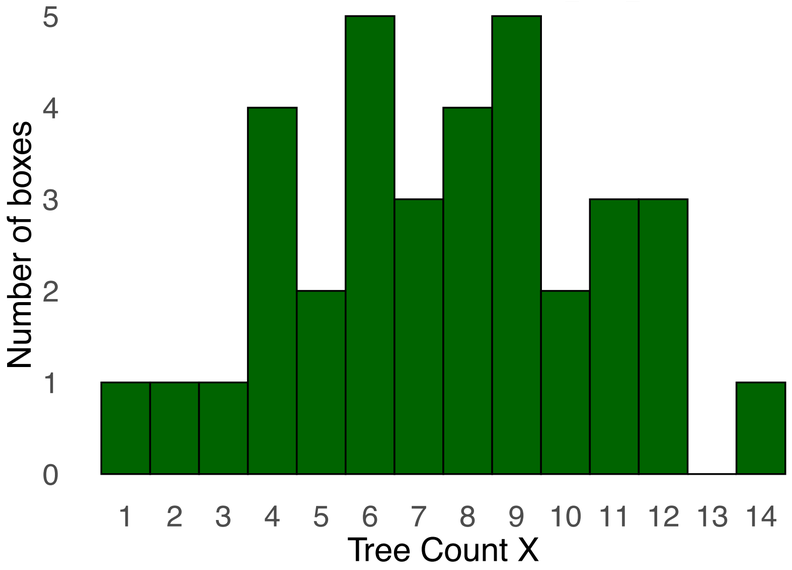


Color denotes the covariate species

Conditional underdispersion

$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 1.29$$



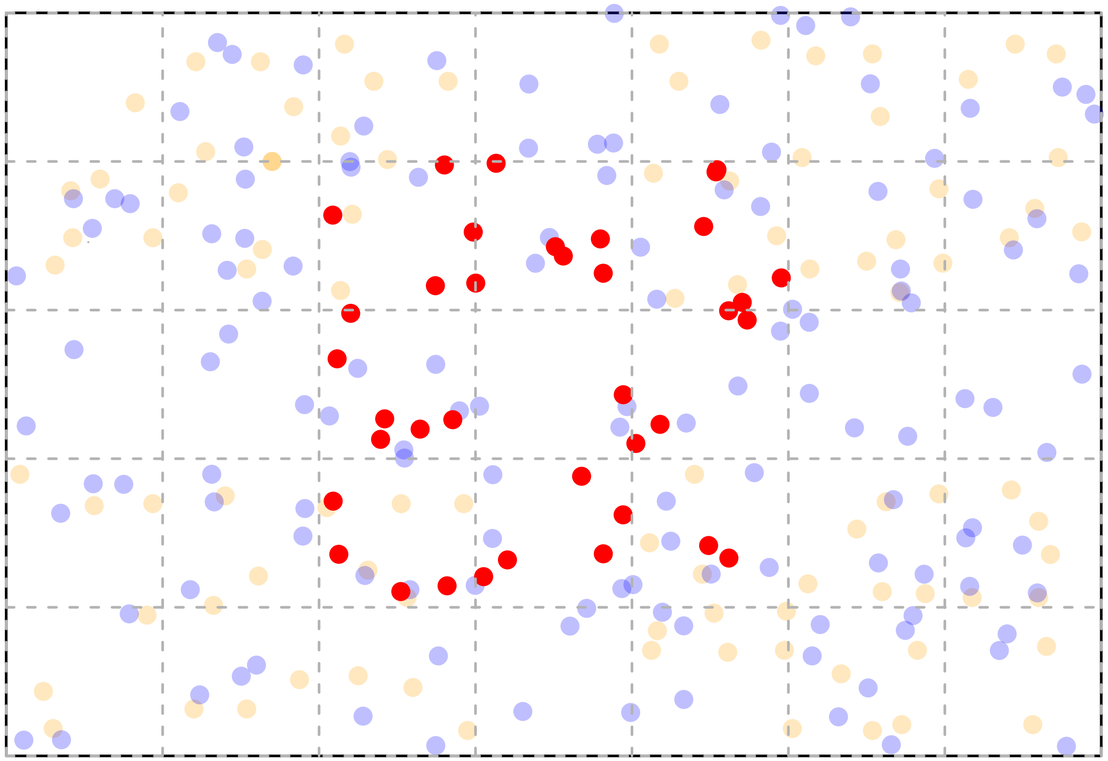


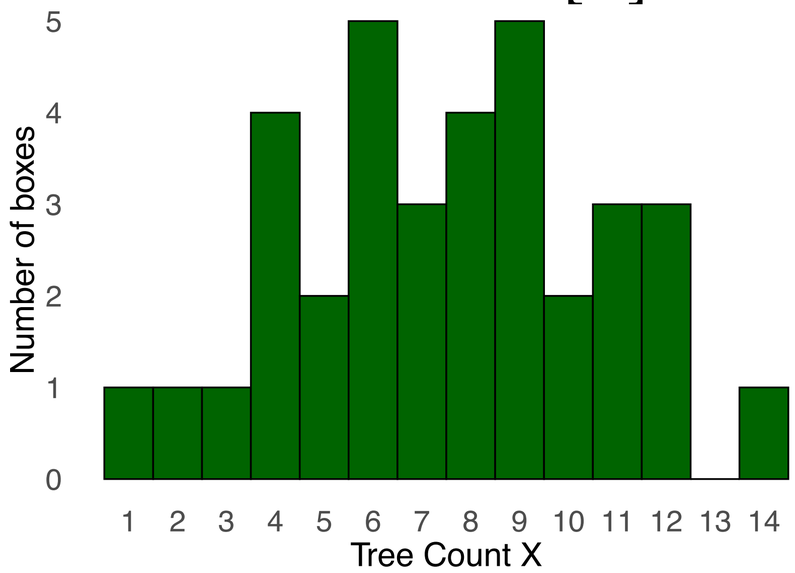
Color denotes the covariate species

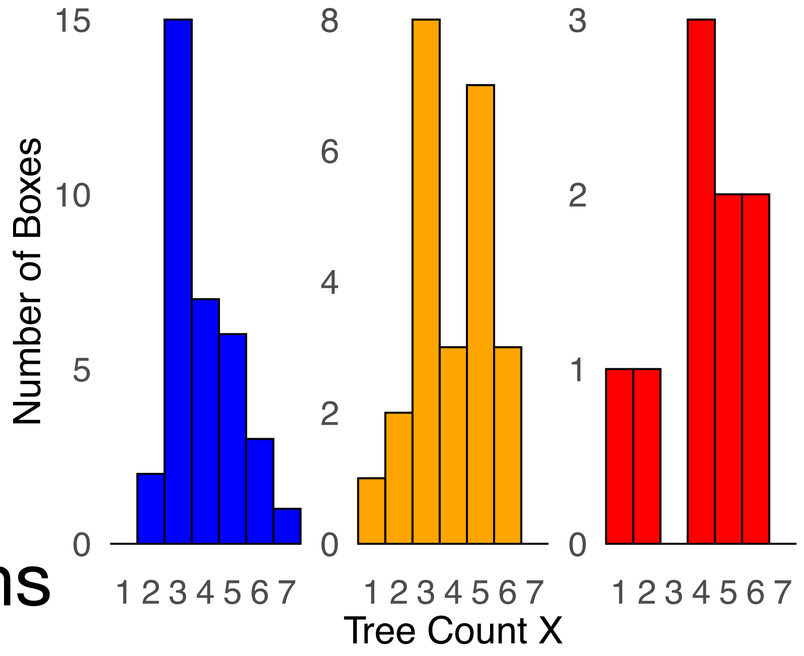
$\mathbb{D}[X \mid \text{Species} = \text{red}] \approx .70$

Conditional underdispersion

$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 1.29$$





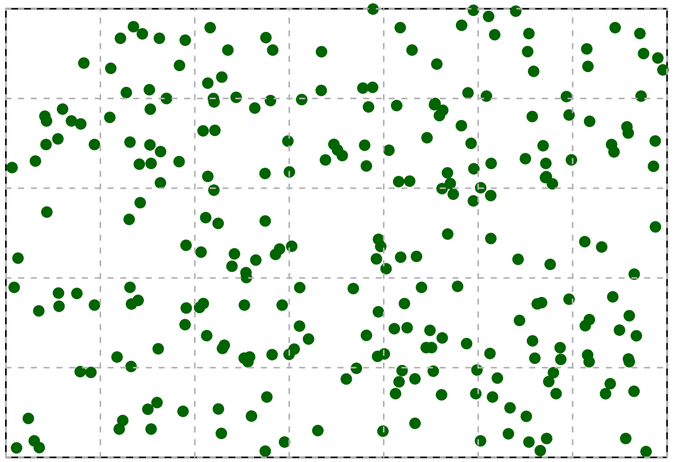


Color denotes the covariate species

$\mathbb{D}[X \mid \text{Species} = \text{red}] \approx .70$

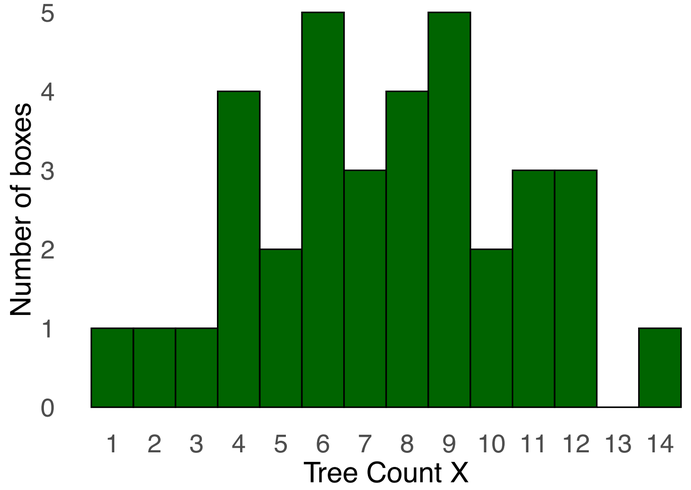
Latent structure can reveal more regular patterns

Underdispersion might be latent

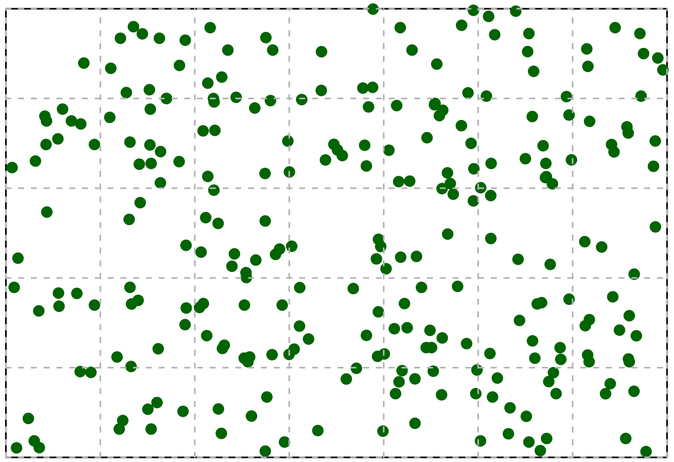


Marginal view

Likely overdispersed

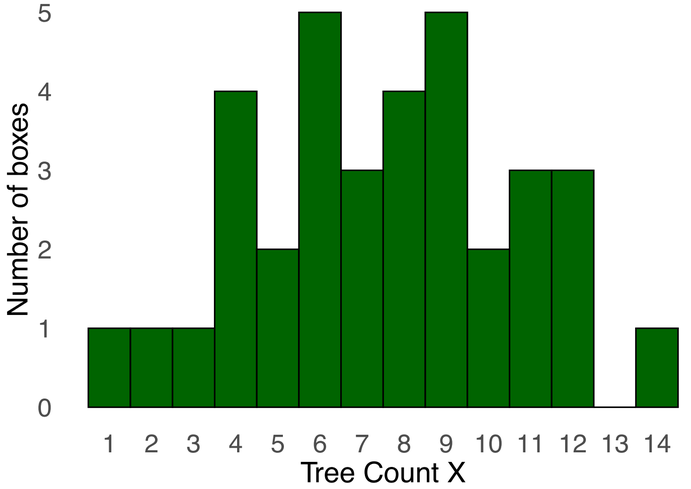


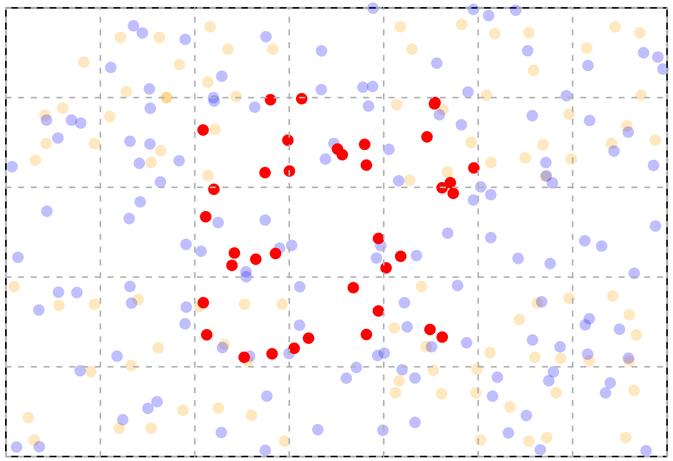
Underdispersion might be latent



Marginal view

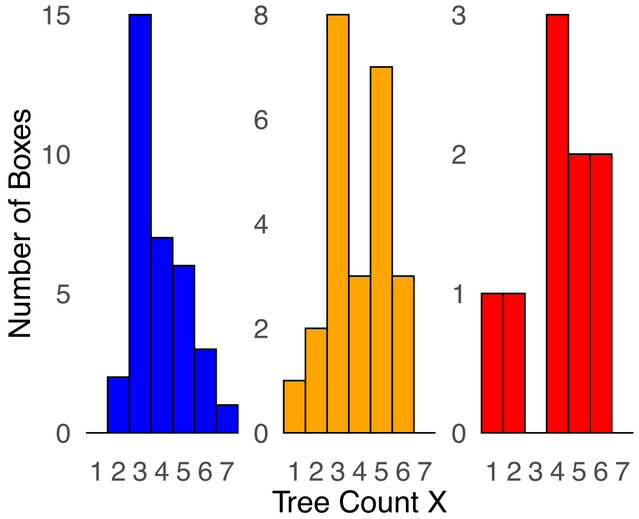
Likely overdispersed



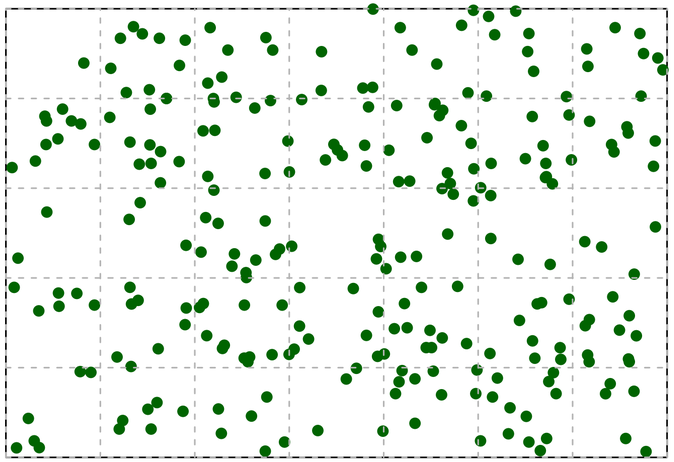


Conditional view

Potentially underdispersed

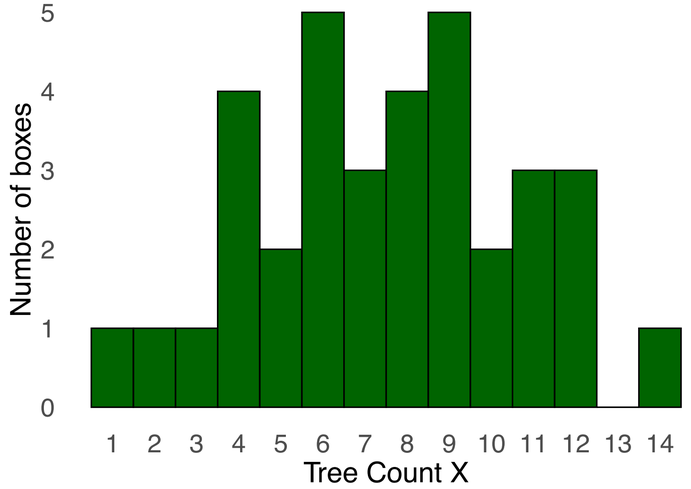


Underdispersion might be latent

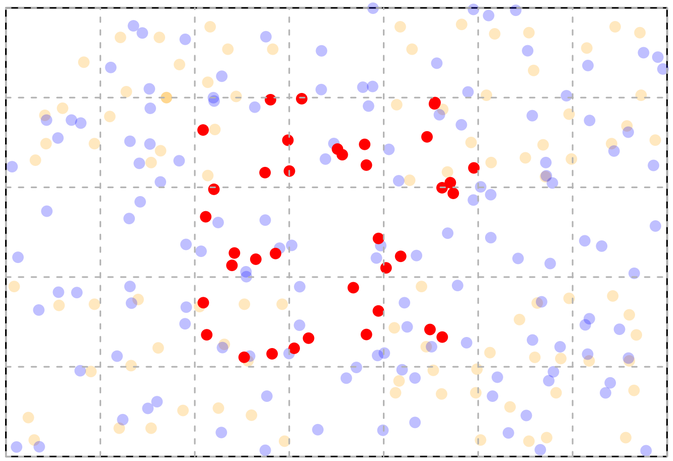


Marginal view

Likely overdispersed

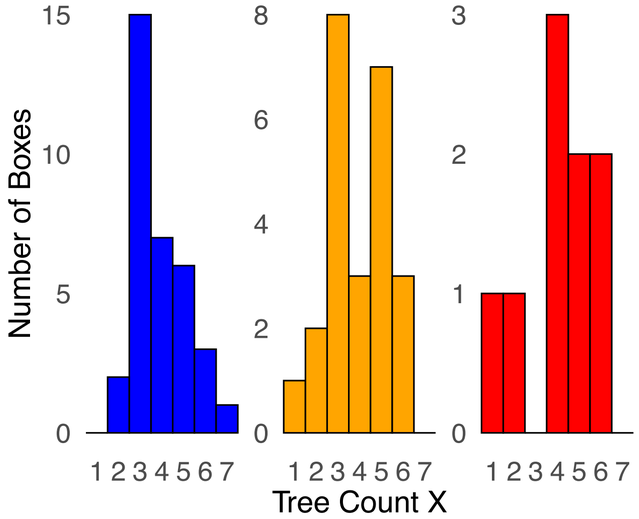


Goal: build **latent variable models** which allow for **conditional underdispersion**



Conditional view

Potentially underdispersed



Goal: latent variable models which allow for **conditional underdispersion**

Goal: latent variable models which allow for **conditional underdispersion**

Count distributions which allow for underdispersion include:

- Conway-Maxwell Poisson [Conway, 1961]
- Double Poisson [Efron, 1986]
- Gamma count distribution [Winklemann, 1995]
- Generalized Poisson [Consul and Famoye, 2006]

Goal: latent variable models which allow for **conditional underdispersion**

Count distributions which allow for underdispersion include:

- Conway-Maxwell Poisson [Conway, 1961]
- Double Poisson [Efron, 1986]
- Gamma count distribution [Winklemann, 1995]
- Generalized Poisson [Consul and Famoye, 2006]

Problem: these distributions lack closed-form conjugate priors

Goal: latent variable models which allow for **conditional underdispersion**

Count distributions which allow for underdispersion include:

- Conway-Maxwell Poisson [Conway, 1961]
- Double Poisson [Efron, 1986]
- Gamma count distribution [Winklemann, 1995]
- Generalized Poisson [Consul and Famoye, 2006]

Problem: these distributions lack closed-form conjugate priors

Our solution: build models around a distribution which

1. is capable of attaining underdispersion
2. can be written in terms of latent Poisson random variables

Poisson order statistics for underdispersed counts

Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Poisson order statistics for underdispersed counts

Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

order D
number of latent Poissons

Poisson order statistics for underdispersed counts

Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

order D
number of latent Poissons

Poisson order statistics for underdispersed counts

Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

rank r: which order statistic

order D number of latent Poissons

Poisson order statistics for underdispersed counts

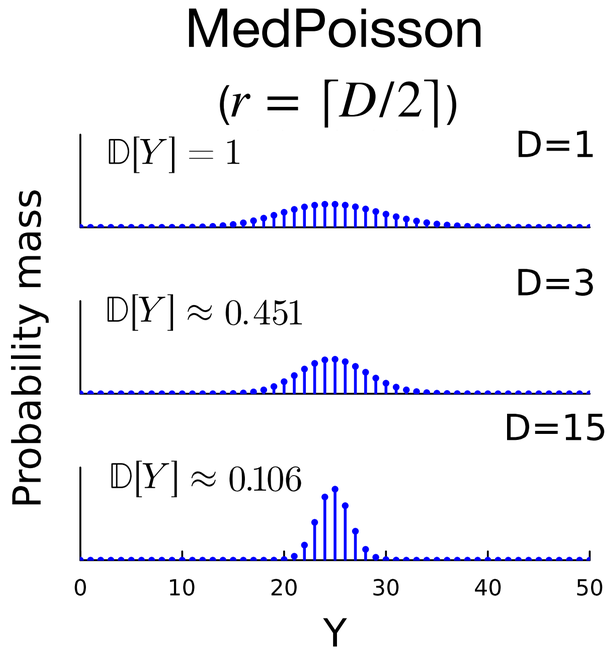
Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

rank r: which order statistic

order D number of latent Poissons



Poisson order statistics for underdispersed counts

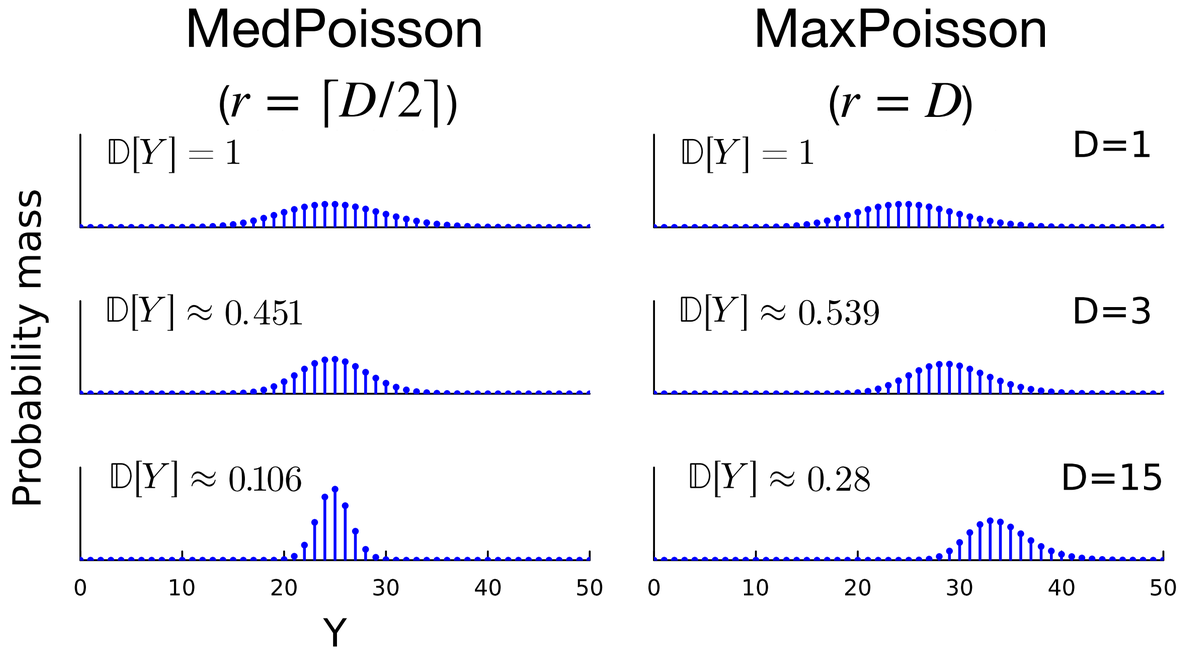
Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

rank r: which order statistic

order D number of latent Poissons



Poisson order statistics for underdispersed counts

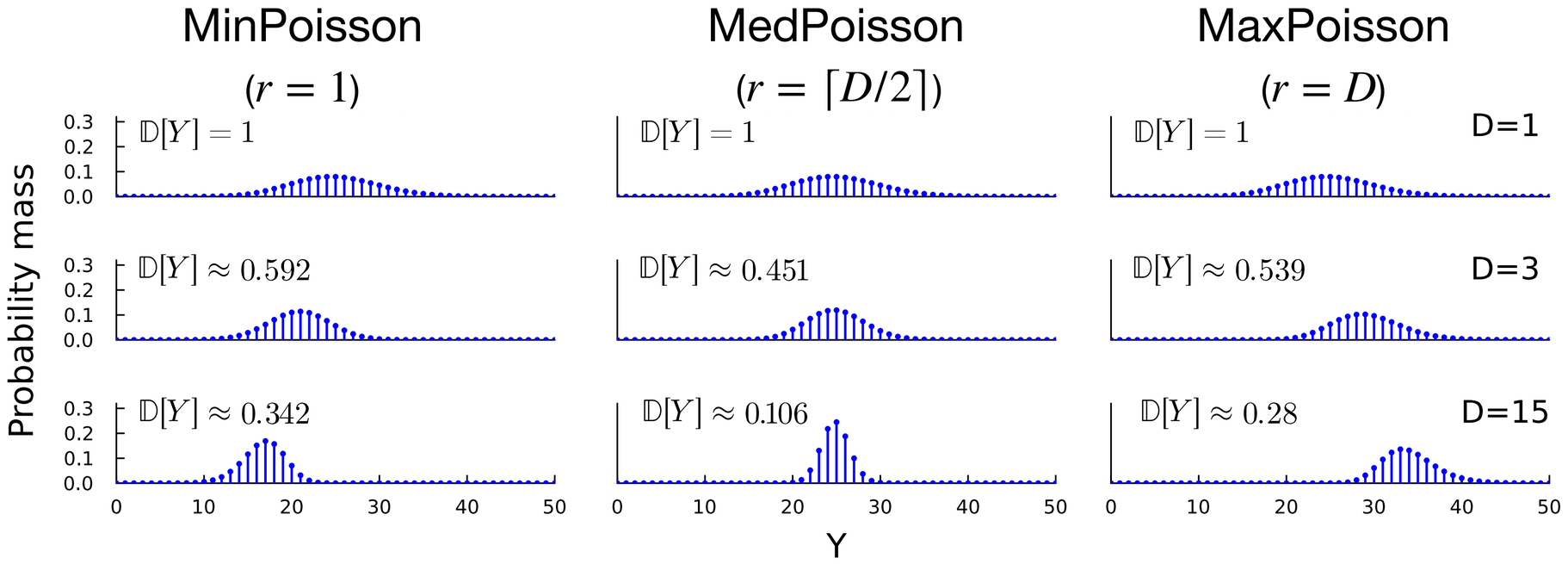
Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

rank r: which order statistic

order D number of latent Poissons



Poisson order statistics for underdispersed counts

Consider a count-valued datapoint $Y \in \mathbb{N}_0$ assumed to be a Poisson order statistic

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

latent rate μ

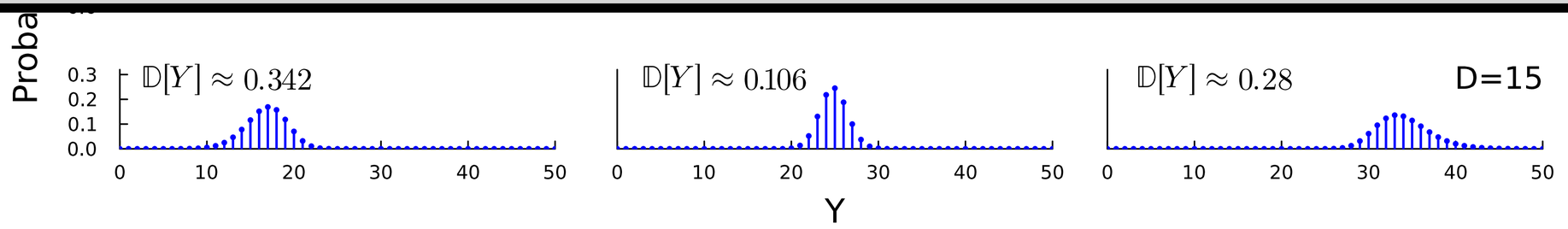
rank r: which order statistic

order D number of latent Poissons

MinPoisson MedPoisson MaxPoisson

Key fact [Badiella, 2023]: For any μ, r, and $D > 1$, the Poisson order statistic $Y \sim \text{Pois}_{\mu}^{(r,D)}$ is underdispersed:

$$\mathbb{D}[Y] = \frac{\mathbb{V}[Y]}{\mathbb{E}[Y]} < 1$$

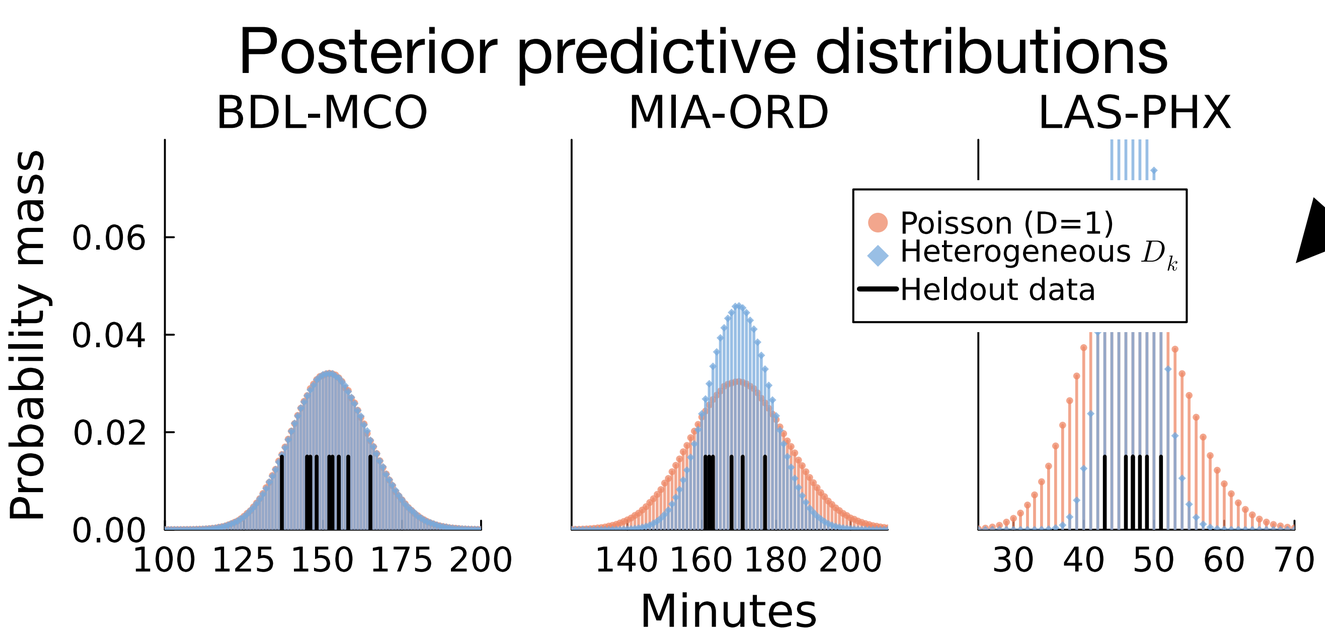


Talk outline

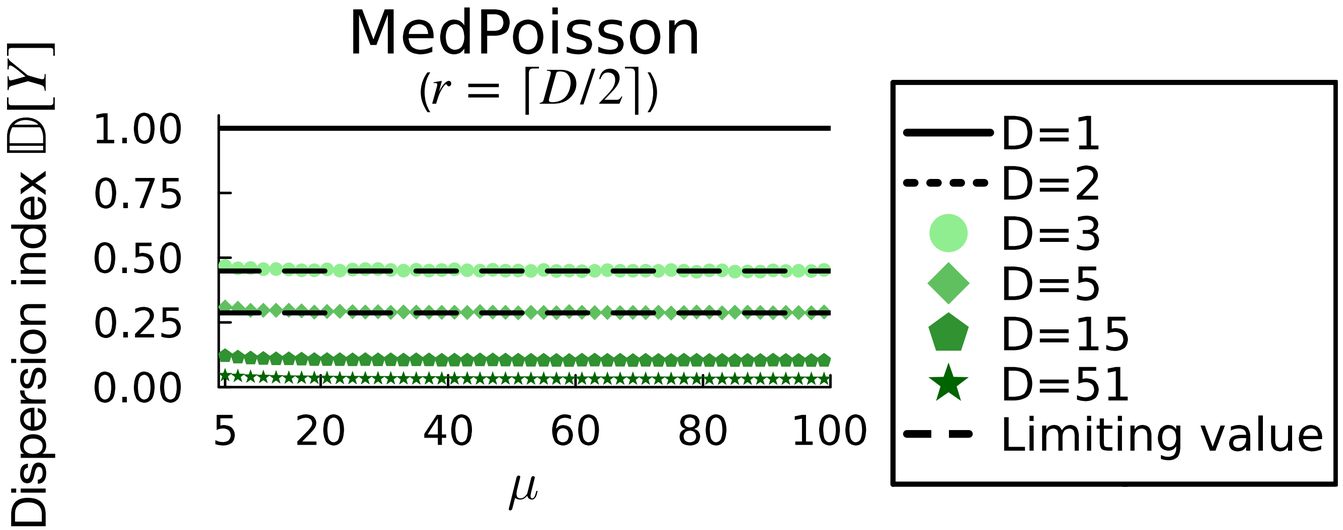
1. Dispersion properties of Poisson order statistics

2. A data augmentation strategy for inference of the parameters of any discrete order statistic **(not only for the Poisson)**

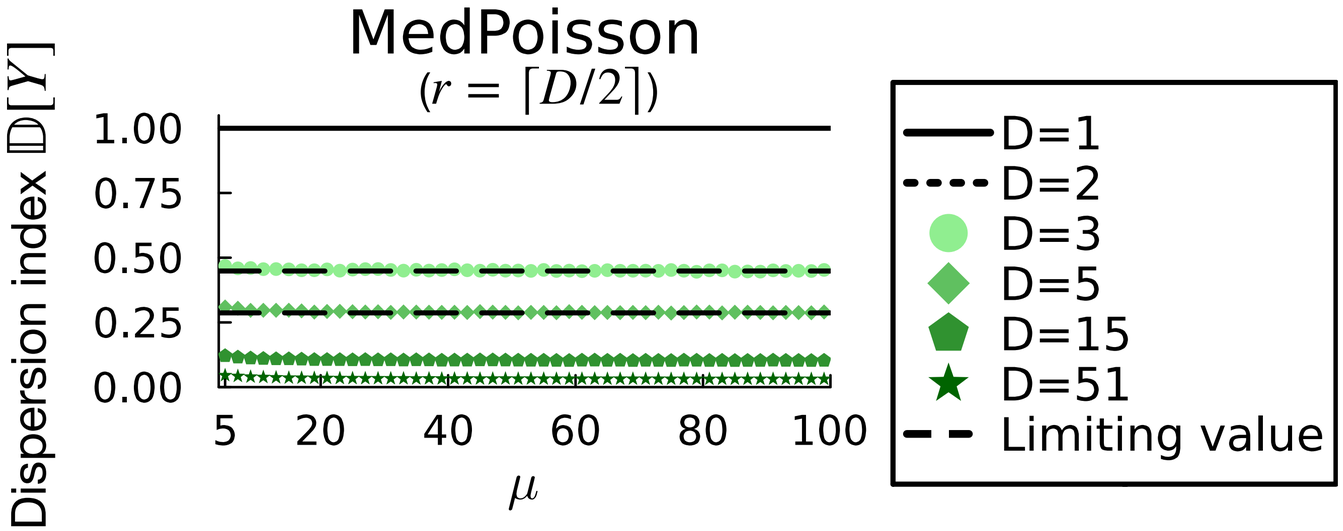
3. Applications where we build and fit hierarchical models for conditionally underdispersed count data, yielding **more precise probabilistic predictions**



Dispersion of Poisson order statistics



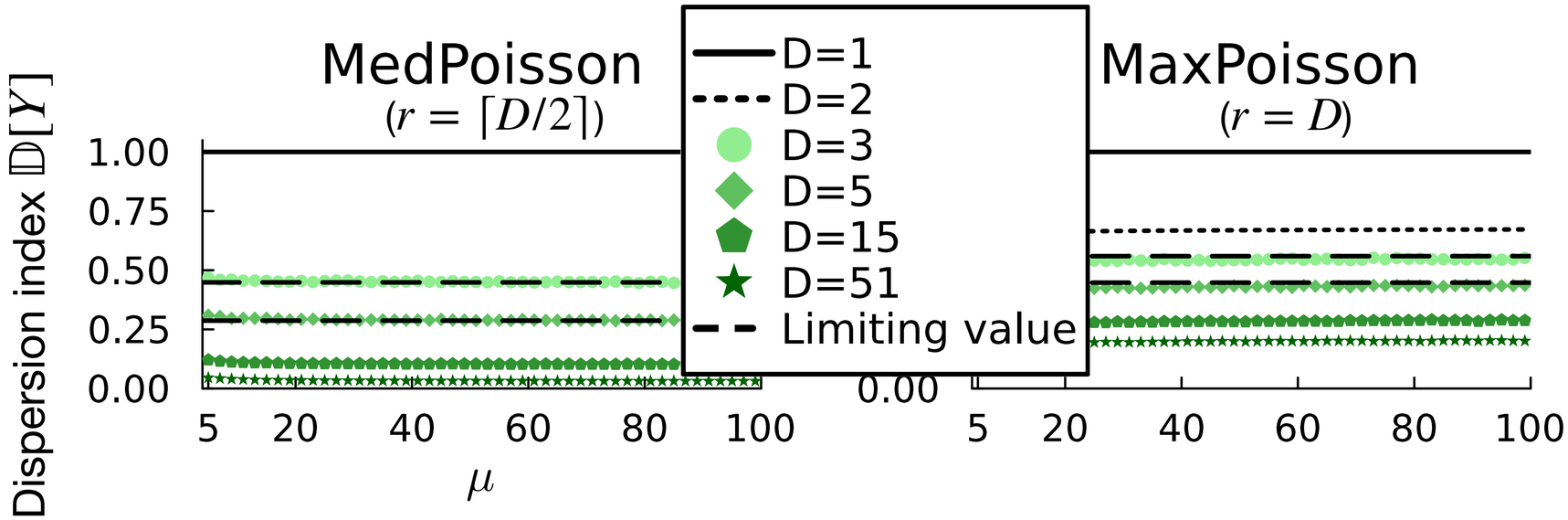
Dispersion of Poisson order statistics



The dispersion $\mathbb{D}_{Y \sim \text{Pois}_{\mu}^{(r,D)}}[Y]$:

1. is stable across large values of μ for each rank r and order D
2. decreases as the order D increases

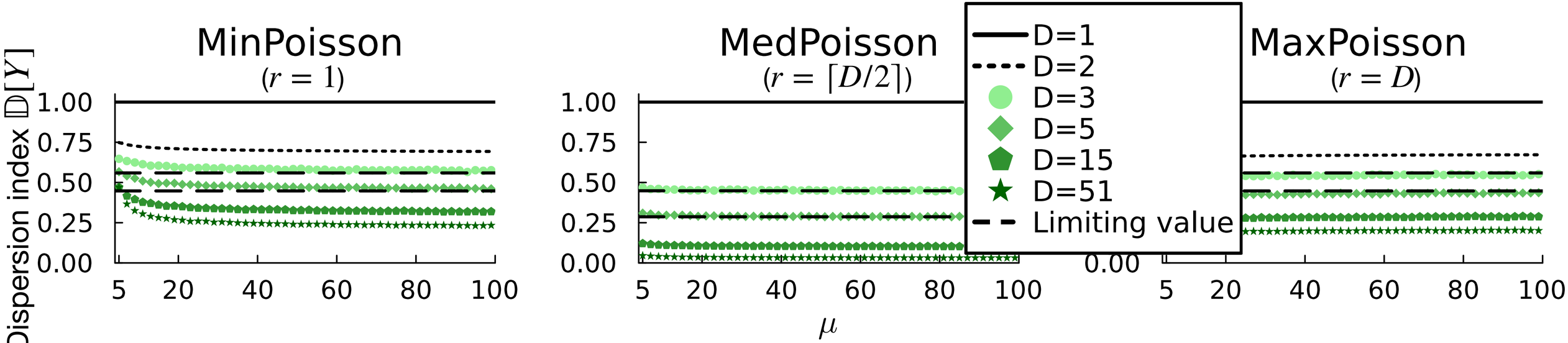
Dispersion of Poisson order statistics



The dispersion $\mathbb{D}_{Y \sim \text{Pois}_\mu^{(r,D)}}[Y]$:

1. is stable across large values of μ for each rank r and order D
2. decreases as the order D increases

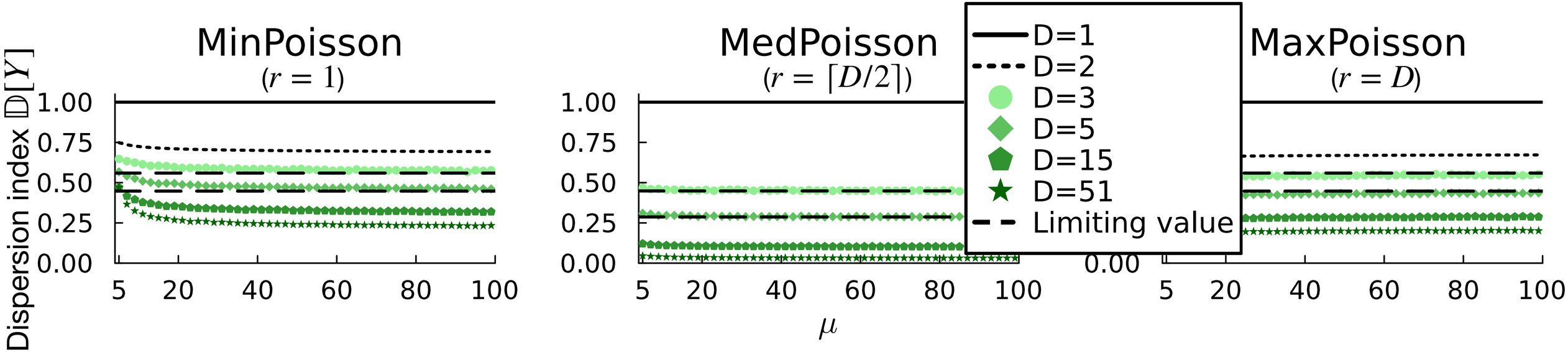
Dispersion of Poisson order statistics



The dispersion $\mathbb{D}_{Y \sim \text{Pois}_{\mu}^{(r,D)}}[Y]$:

1. is stable across large values of μ for each rank r and order D
2. decreases as the order D increases

Dispersion of Poisson order statistics



The dispersion $\mathbb{D}_{Y \sim \text{Pois}_{\mu}^{(r,D)}}[Y]$:

1. is stable across large values of μ for each rank r and order D
2. decreases as the order D increases

We can use the order D as a "pseudo-index" to control the level of dispersion when building models

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

First: $$P_{\mu}(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y) \propto 1(Y = Z^{(r,D)}) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Second: $$P(\mu \mid \mathbf{Z}_{1:D}) \propto g(\mu) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Data augmentation

First:
$$P_{\mu}(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y) \propto 1(Y = Z^{(r,D)}) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Second:
$$P(\mu \mid \mathbf{Z}_{1:D}) \propto g(\mu) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Data augmentation

First: $$P_{\mu}(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y) \propto 1(Y = Z^{(r,D)}) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Parameter inference

Second: $$P(\mu \mid \mathbf{Z}_{1:D}) \propto g(\mu) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Data augmentation → difficult

First: $$P_{\mu}(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y) \propto 1(Y = Z^{(r,D)}) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Parameter inference → easy

Second: $$P(\mu \mid \mathbf{Z}_{1:D}) \propto g(\mu) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

If $\mathbf{Z}_{1:D}$ were continuous

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

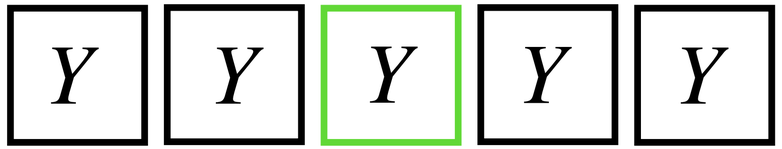
$$Z^{(3,5)} = Y$$

$\mathbf{Z}_{1:D}$ is **discrete**

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Example arrangements:

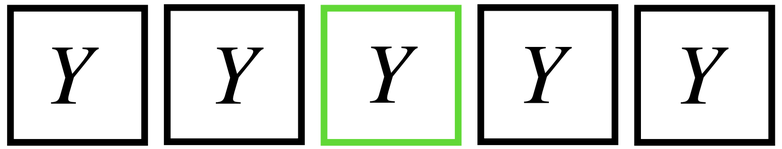


Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Only the support of each Z_d matters to satisfy the constraint $Z^{(r,D)} = Y$

Example arrangements:



Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Only the support of each Z_d matters to satisfy the constraint $Z^{(r,D)} = Y$

Introduce a categorical random variable for each Z_d denoted $C_d \in \{<Y, =Y, >Y\}$ which determines the support of Z_d

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Only the support of each Z_d matters to satisfy the constraint $Z^{(r,D)} = Y$

Introduce a categorical random variable for each Z_d denoted $C_d \in \{<Y, =Y, >Y\}$ which determines the support of Z_d

1. $\mathbf{Z}_{1:D}$ are conditionally independent of $Z^{(r,D)} = Y$ given $\mathbf{C}_{1:D}$

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Only the support of each Z_d matters to satisfy the constraint $Z^{(r,D)} = Y$

Introduce a categorical random variable for each Z_d denoted $C_d \in \{<Y, =Y, >Y\}$ which determines the support of Z_d

1. $\mathbf{Z}_{1:D}$ are conditionally independent of $Z^{(r,D)} = Y$ given $\mathbf{C}_{1:D}$

$$P_\mu(\mathbf{Z}_{1:D} \mid \mathbf{C}_{1:D}, Z^{(r,D)} = Y) = P_\mu(\mathbf{Z}_{1:D} \mid \mathbf{C}_{1:D}) = \prod_{d=1}^{D} \underset{\mathbb{Z}_{C_d}}{\text{trunc}}\,\text{Pois}_\mu(Z_d)$$

Data Augmentation: Sampling from $P_\mu(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Only the support of each Z_d matters to satisfy the constraint $Z^{(r,D)} = Y$

Introduce a categorical random variable for each Z_d denoted $C_d \in \{<Y, =Y, >Y\}$ which determines the support of Z_d

1. $\mathbf{Z}_{1:D}$ are conditionally independent of $Z^{(r,D)} = Y$ given $\mathbf{C}_{1:D}$

$$P_\mu(\mathbf{Z}_{1:D} \mid \mathbf{C}_{1:D}, Z^{(r,D)} = Y) = P_\mu(\mathbf{Z}_{1:D} \mid \mathbf{C}_{1:D}) = \prod_{d=1}^{D} \underset{\mathbb{Z}_{C_d}}{\text{trunc}}\,\text{Pois}_\mu(Z_d)$$

2. Given C_d, sample each Z_d as a truncated Poisson, which can be made efficient

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$C_1 = <Y$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$C_1 = <Y$$

$$C_2 = <Y$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$C_1 = {<}Y \quad C_3 = {=}Y$

$C_2 = {<}Y$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$\mathbf{C}_d \,|\, \mathbf{C}_{1:d-1} \sim \text{Categorical}(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)})$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$\mathbf{C}_d \mid \mathbf{C}_{1:d-1} \sim \text{Categorical}(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)})$$

We can calculate these probabilities efficiently in closed-form

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$\mathbf{C}_d \mid \mathbf{C}_{1:d-1} \sim \text{Categorical}(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)})$$

$$n_d^{(<Y)} = \sum_{s=1}^{d-1} 1\{C_s = <Y\}$$

$$n_d^{(=Y)} = \sum_{s=1}^{d-1} 1\{C_s = =Y\},$$

$$n_d^{(>Y)} = \sum_{s=1}^{d-1} 1\{C_s = >Y\}$$

We can calculate these probabilities efficiently in closed-form

Main idea: The past counts for each group are **sufficient statistics**

$$P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}) = P_\mu(Z^{(r,D)} = Y \mid n_d^{(<Y)}, n_d^{(=Y)}, n_d^{(>Y)})$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

Strategy: We sample the support $\mathbf{C}_{1:D}$ sequentially from $P_\mu(C_d \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$$\mathbf{C}_d \mid \mathbf{C}_{1:d-1} \sim \text{Categorical}(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)})$$

$$n_d^{(<Y)} = \sum_{s=1}^{d-1} 1\{C_s = <Y\}$$

$$n_d^{(=Y)} = \sum_{s=1}^{d-1} 1\{C_s = =Y\},$$

$$n_d^{(>Y)} = \sum_{s=1}^{d-1} 1\{C_s = >Y\}$$

We can calculate these probabilities efficiently in closed-form

Main idea: The past counts for each group are **sufficient statistics**

$$P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}) = P_\mu(Z^{(r,D)} = Y \mid n_d^{(<Y)}, n_d^{(=Y)}, n_d^{(>Y)})$$

By Bayes rule, for $c \in \{<Y, =Y, >Y\}$:

$$p_d^{(c)} := P_\mu(C_d = c \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y) = \frac{P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}, C_d = c)}{P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1})} P_\mu(C_d = c)$$

Sampling the support $\mathbf{C}_{1:D}$ from $P_\mu(\mathbf{C}_{1:D} \mid Z^{(r,D)} = Y)$

$$Z^{(3,5)} = Y$$

$\mathbf{C}_d \mid \mathbf{C}_{1:d-1} \sim \text{Categorical}(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)})$

Takeaway: we can sample the support $\mathbf{C}_{1:D}$ in closed-form,

Therefore, we can sample the latent $\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y$ as desired

$$P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}) = P_\mu(Z^{(r,D)} = Y \mid n_d^{(<Y)}, n_d^{(=Y)}, n_d^{(>Y)})$$

$$p_d^{(c)} := P_\mu(C_d = c \mid \mathbf{C}_{1:d-1}, Z^{(r,D)} = Y) = \frac{P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}, C_d = c)}{P_\mu(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1})} P_\mu(C_d = c)$$

Inference for discrete order statistics via data augmentation

Our strategy revolves around inferring the latent $\mathbf{Z}_{1:D}$

$$Y \sim \text{Pois}_{\mu}^{(r,D)} \iff Y = Z^{(r,D)} \text{ where } Z_1, \ldots, Z_D \sim \text{Pois}(\mu)$$

Data augmentation → difficult

First: $$P_{\mu}(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y) \propto 1(Y = Z^{(r,D)}) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Parameter inference → easy

Second: $$P(\mu \mid \mathbf{Z}_{1:D}) \propto g(\mu) \prod_{d=1}^{D} \text{Pois}_{\mu}(Z_d)$$

Case study: predicting flight times

Each observation $Y_i \in \mathbb{N}_0$ is the number of minutes for flight i between departure and arrival.

$$Y_i \overset{\text{ind.}}{\sim} \text{MedPois}_{\mu_i}^{(D_{\text{route}[i]})} \text{ where } \mu_i \overset{\text{def}}{=} a_{\text{orig}[i]} + b_{\text{dest}[i]} + c_{\text{route}[i]} \, \text{dist}_{\text{route}[i]}$$

Case study: predicting flight times

Each observation $Y_i \in \mathbb{N}_0$ is the number of minutes for flight i between departure and arrival.

$$Y_i \overset{\text{ind.}}{\sim} \text{MedPois}_{\mu_i}^{(D_{\text{route}[i]})} \text{ where } \mu_i \overset{\text{def}}{=} a_{\text{orig}[i]} + b_{\text{dest}[i]} + c_{\text{route}[i]} \, \text{dist}_{\text{route}[i]}$$

↓ heterogeneous dispersion for each route k

$D_k \sim \text{OddBinomial}(D_{\text{max}}, \rho)$ such that $D_k \in \{1,3,5,\ldots,D_{\text{max}}\}$

Case study: predicting flight times

Each observation $Y_i \in \mathbb{N}_0$ is the number of minutes for flight i between departure and arrival.

$$Y_i \overset{\text{ind.}}{\sim} \text{MedPois}_{\mu_i}^{(D_{\text{route}[i]})} \text{ where } \mu_i \overset{\text{def}}{=} a_{\text{orig}[i]} + b_{\text{dest}[i]} + c_{\text{route}[i]} \, \text{dist}_{\text{route}[i]}$$

heterogeneous dispersion for each route k

$D_k \sim \text{OddBinomial}(D_{\text{max}}, \rho)$ such that $D_k \in \{1,3,5,\ldots,D_{\text{max}}\}$

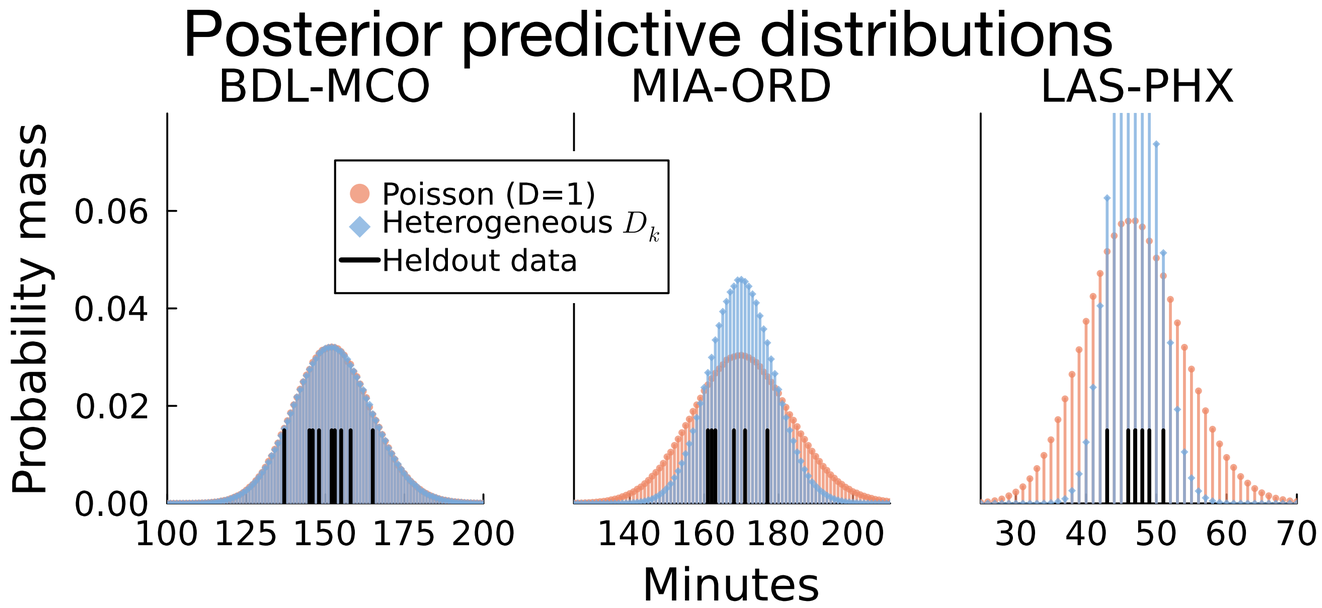
Case study: predicting flight times

Each observation $Y_i \in \mathbb{N}_0$ is the number of minutes for flight i between departure and arrival.

$$Y_i \overset{\text{ind.}}{\sim} \text{MedPois}_{\mu_i}^{(D_{\text{route}[i]})} \text{ where } \mu_i \overset{\text{def}}{=} a_{\text{orig}[i]} + b_{\text{dest}[i]} + c_{\text{route}[i]}\,\text{dist}_{\text{route}[i]}$$

heterogeneous dispersion for each route k

$D_k \sim \text{OddBinomial}(D_{\text{max}}, \rho)$ such that $D_k \in \{1,3,5,\ldots,D_{\text{max}}\}$



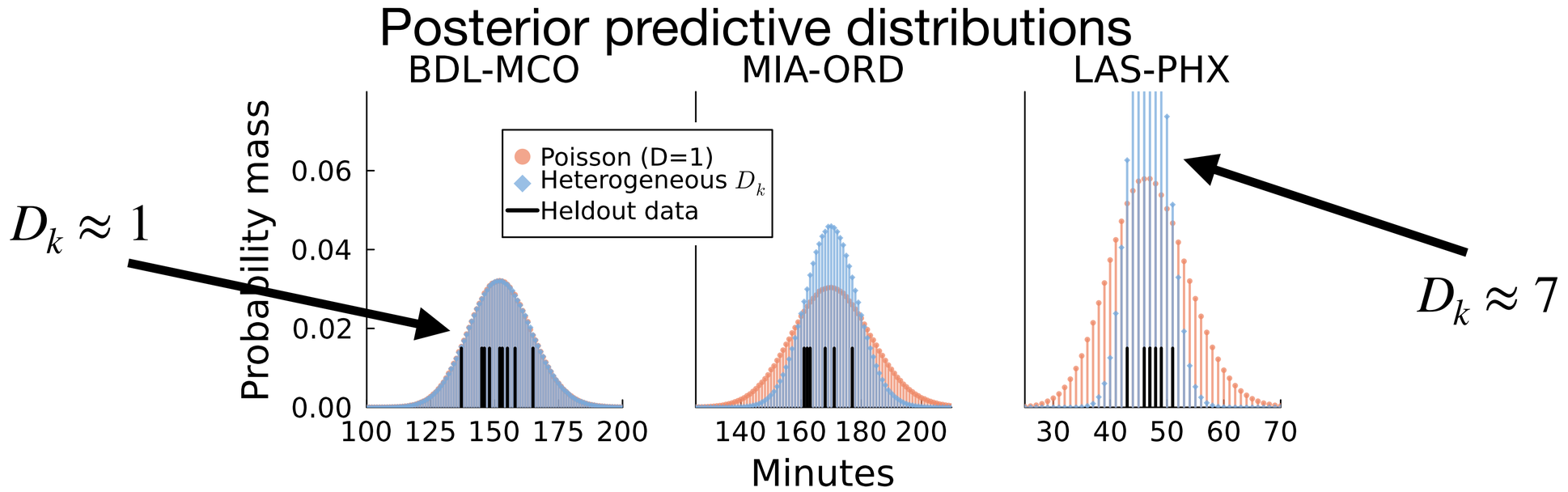
Case study: predicting flight times

Each observation $Y_i \in \mathbb{N}_0$ is the number of minutes for flight i between departure and arrival.

$$Y_i \overset{\text{ind.}}{\sim} \text{MedPois}_{\mu_i}^{(D_{\text{route}[i]})} \text{ where } \mu_i \overset{\text{def}}{=} a_{\text{orig}[i]} + b_{\text{dest}[i]} + c_{\text{route}[i]}\,\text{dist}_{\text{route}[i]}$$

heterogeneous dispersion for each route k

$D_k \sim \text{OddBinomial}(D_{\text{max}}, \rho)$ such that $D_k \in \{1,3,5,\ldots,D_{\text{max}}\}$



Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \text{ where } \mu_{i,t} := Y_{i,t-1} + \log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)$$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)$$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}^{(D_{i,t})}_{\mu_{i,t}} \quad \text{where} \quad \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \quad \text{where} \quad \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

heterogeneous dispersion

Case study: forecasting COVID-19 cases

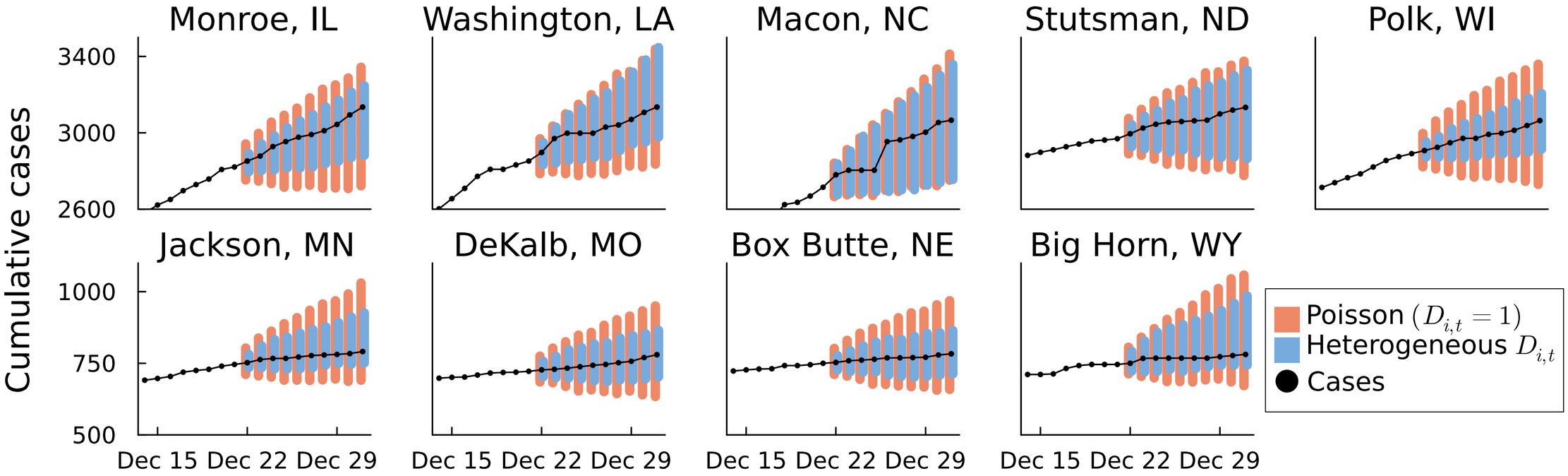
Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}^{(D_{i,t})}_{\mu_{i,t}} \quad \text{where} \quad \mu_{i,t} := Y_{i,t-1} + \log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)$$

heterogeneous dispersion

past case count

latent growth rate



More precise probabilistic forecasts

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}^{(D_{i,t})}_{\mu_{i,t}} \quad \text{where} \quad \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \quad \text{where} \quad \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k} \phi_{k,t}\right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \quad \text{where} \quad \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \log(\text{pop}_i) \underbrace{\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k} \phi_{k,t} \right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)} \phi_{k,t-1},\ b^{(\phi)})$

Data augmentation from [Acharya et al. 2015]

Case study: forecasting COVID-19 cases

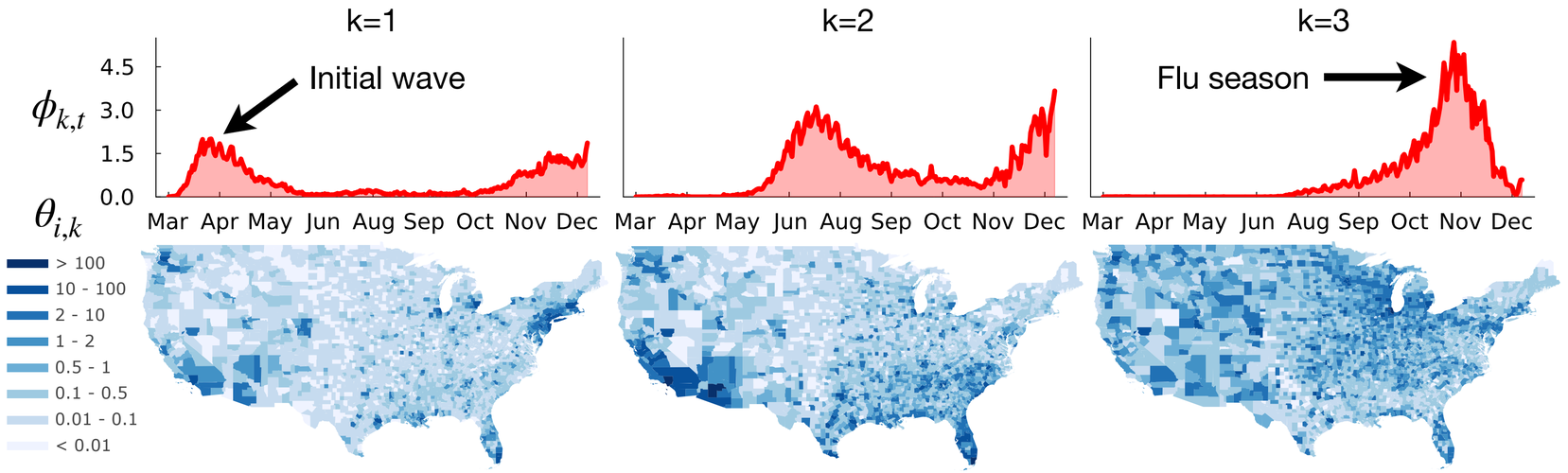
Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \quad \text{where} \quad \mu_{i,t} := Y_{i,t-1} + \log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)$$

past case count — latent growth rate

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

Data augmentation from [Acharya et al. 2015]



Smooth, interpretable structure

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

Data augmentation from [Acharya et al. 2015]

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \log(\text{pop}_i) \underbrace{\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k} \phi_{k,t} \right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)} \phi_{k,t-1},\ b^{(\phi)})$

Data augmentation from [Acharya et al. 2015]

For heterogeneous dispersion for each data point:

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}^{(D_{i,t})}_{\mu_{i,t}} \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

Data augmentation from [Acharya et al. 2015]

For heterogeneous dispersion for each data point:

$$D_{i,t} \sim \text{OddBinomial}\left(D_{\text{max}}, \rho_{i,t}\right) \text{ where } \rho_{i,t} := \text{logit}^{-1}\left(\sum_{q=1}^{Q} \beta_{i,q}\tau_{q,t}\right)$$

where $\beta_{i,q} \sim \mathcal{N}(0,1)$ and $\tau_{q,t} \sim \mathcal{N}\left(a^{(\tau)} + b^{(\tau)}\tau_{q,t-1},\ 1/\lambda_t^{(\tau)}\right)$

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}^{(D_{i,t})} \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \underbrace{\log(\text{pop}_i)\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

→ Data augmentation from [Acharya et al. 2015]

For heterogeneous dispersion for each data point:

$$D_{i,t} \sim \text{OddBinomial}\left(D_{\max}, \rho_{i,t}\right) \text{ where } \rho_{i,t} := \text{logit}^{-1}\left(\sum_{q=1}^{Q} \beta_{i,q}\tau_{q,t}\right)$$

where $\beta_{i,q} \sim \mathcal{N}(0,1)$ and $\tau_{q,t} \sim \mathcal{N}\left(a^{(\tau)} + b^{(\tau)}\tau_{q,t-1},\ 1/\lambda_t^{(\tau)}\right)$ → Data augmentation from [Polson et al. 2013]

Case study: forecasting COVID-19 cases

Each $Y_{i,t}$ is the cumulative number of COVID-19 cases at time t in county i.

$$Y_{i,t} \sim \text{MedPois}_{\mu_{i,t}}(D_{i,t}) \text{ where } \mu_{i,t} := \underbrace{Y_{i,t-1}}_{\text{past case count}} + \log(\text{pop}_i)\underbrace{\left(\varepsilon + \alpha \sum_{k=1}^{K} \theta_{i,k}\phi_{k,t}\right)}_{\text{latent growth rate}}$$

To forecast, $\phi_{k,t}$ evolves over time: $\phi_{k,t} \sim \Gamma(a^{(\phi)} + b^{(\phi)}\phi_{k,t-1},\ b^{(\phi)})$

→ Data augmentation from [Acharya et al. 2015]

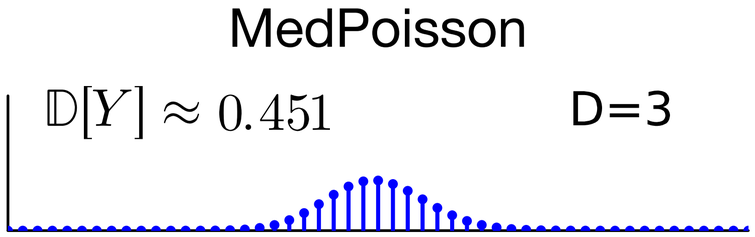
For heterogeneous dispersion for each data point:

$$D_{i,t} \sim \text{OddBinomial}\left(D_{\text{max}}, \rho_{i,t}\right) \text{ where } \rho_{i,t} := \text{logit}^{-1}\left(\sum_{q=1}^{Q} \beta_{i,q}\tau_{q,t}\right)$$

where $\beta_{i,q} \sim \mathcal{N}(0,1)$ and $\tau_{q,t} \sim \mathcal{N}\left(a^{(\tau)} + b^{(\tau)}\tau_{q,t-1},\ 1/\lambda_t^{(\tau)}\right)$ → Data augmentation from [Polson et al. 2013]

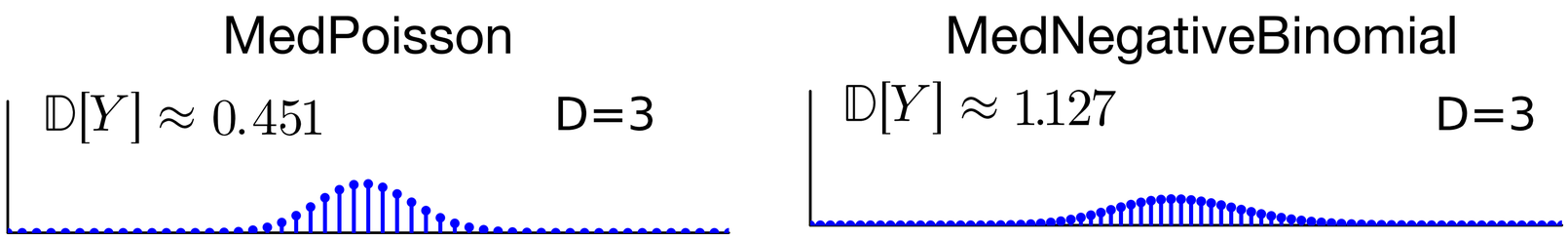
General tools enable modularity and increasing complexity

Summary



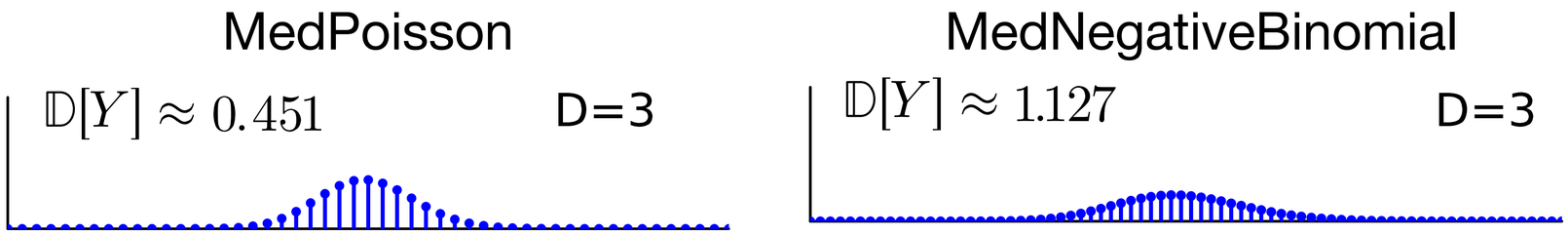
- We study **Poisson order statistics** as a tool for **modeling underdispersion** in probabilistic models of count data

Summary



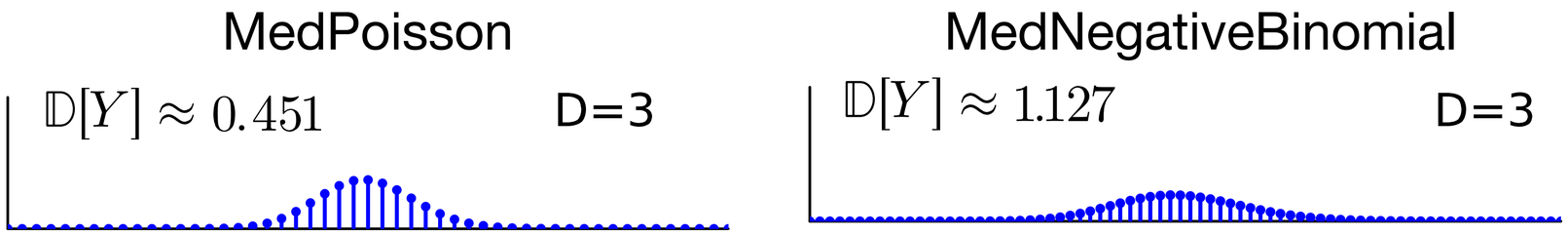
- We study **Poisson order statistics** as a tool for **modeling underdispersion** in probabilistic models of count data
- We develop a data augmentation for inference for **any** discrete order statistic

Summary

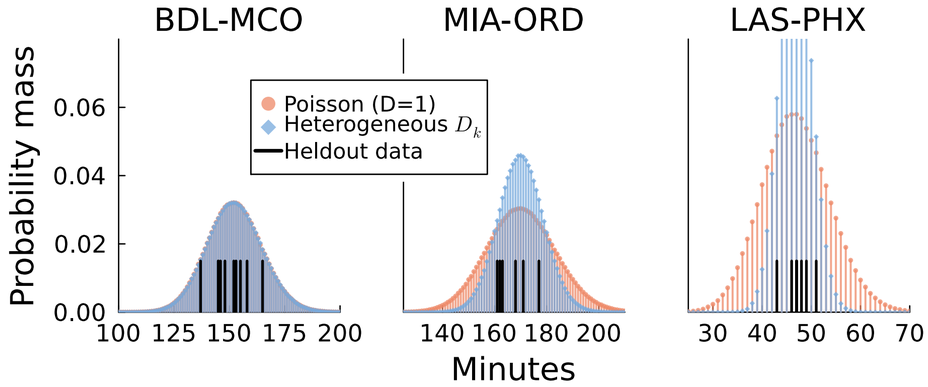


- We study **Poisson order statistics** as a tool for **modeling underdispersion** in probabilistic models of count data
- We develop a data augmentation for inference for **any** discrete order statistic
  - These tools might be useful for building Bayesian methods to model other phenomena, like extreme values and constrained parameter spaces

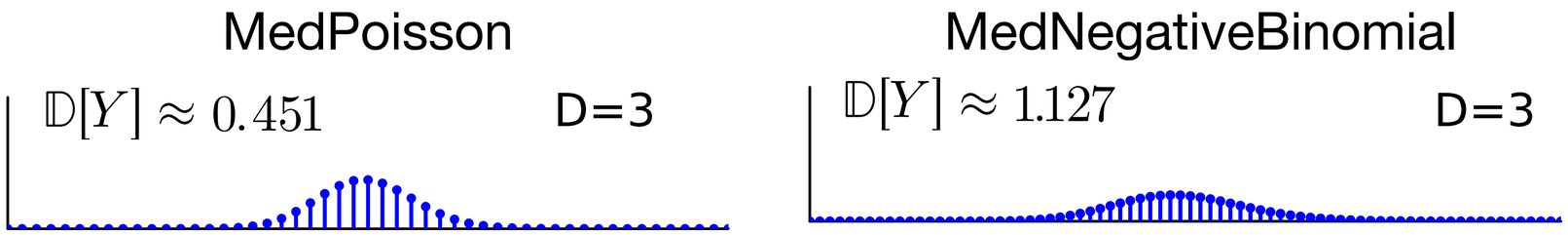
Summary



- We study **Poisson order statistics** as a tool for **modeling underdispersion** in probabilistic models of count data
- We develop a data augmentation for inference for **any** discrete order statistic
  - These tools might be useful for building Bayesian methods to model other phenomena, like extreme values and constrained parameter spaces
- Using these tools enables **more precise probabilistic predictions** for count data

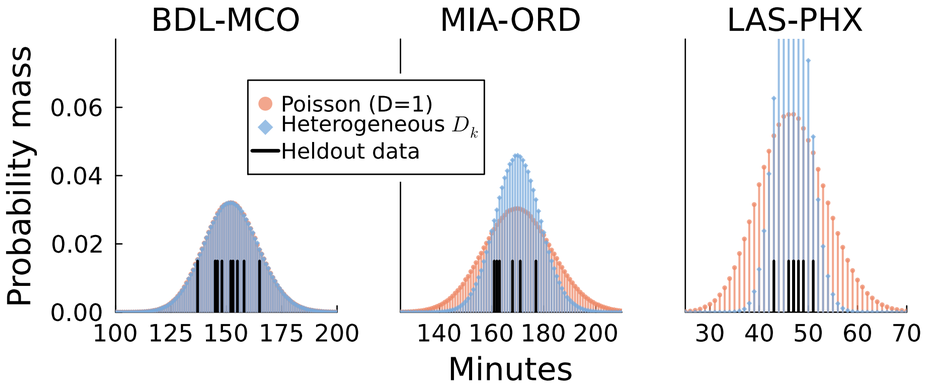


Summary



- We study **Poisson order statistics** as a tool for **modeling underdispersion** in probabilistic models of count data
- We develop a data augmentation for inference for **any** discrete order statistic
  - These tools might be useful for building Bayesian methods to model other phenomena, like extreme values and constrained parameter spaces
- Using these tools enables **more precise probabilistic predictions** for count data

Thank you!



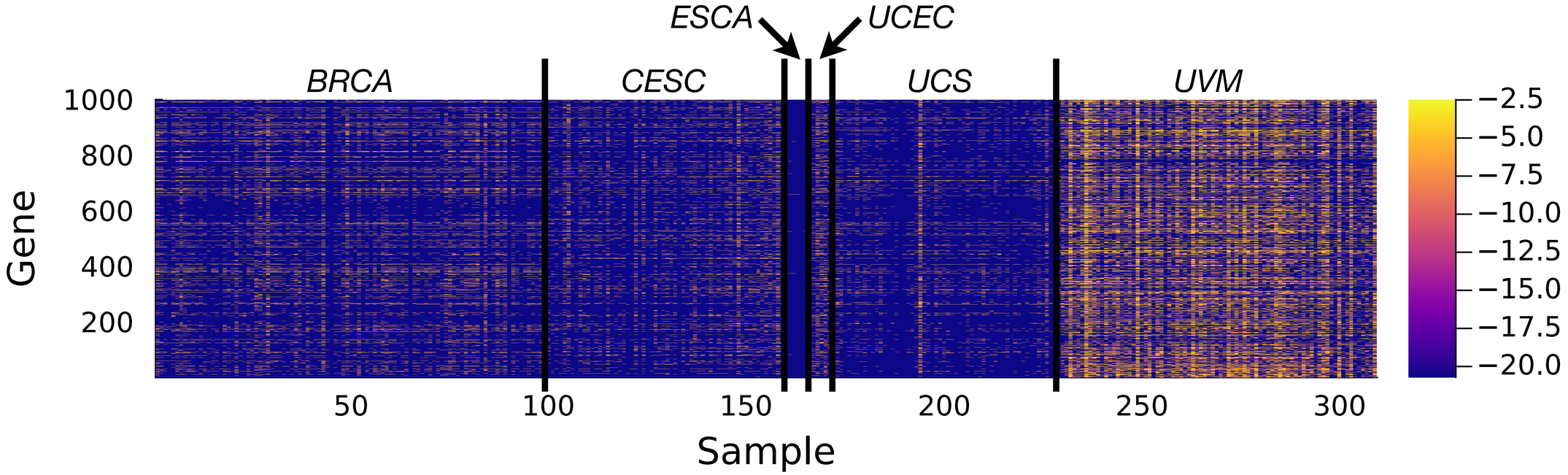
Case study: RNA-sequencing data

Is there evidence of conditional underdispersion in RNA-sequencing?

$Y_{i,j}$ is the read-count of gene j for subject i

$$Y_{i,j} \sim \text{MedNB}^{(D_{i,j})}_{\alpha_{i,j},\, p_j} \quad \text{where } \alpha_{i,j} := \sum_{k=1}^{K} \theta_{i,k}\phi_{k,j}$$

Posterior log-probability of underdispersion for each sample-gene pair



Exact form for support probabilities

For $c \in \{<Y, =Y, >Y\}$:

$$p_d^{(c)} = \frac{P_\theta(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}, C_d = c)}{P_\theta(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1})} P_\theta(C_d = c)$$

$$n_1 = \sum_{s=1}^{d-1} 1\{C_s = <Y\}$$

$$n_2 = \sum_{s=1}^{d-1} 1\{C_s = =Y\},$$

$$n_3 = \sum_{s=1}^{d-1} 1\{C_s = >Y\}$$

$$P_\theta(Z^{(r,D)} = Y \mid \mathbf{C}_{1:d-1}) = \begin{cases} 0 & \text{if } n_1 \geq r \text{ or } n_3 \geq D - r + 1 \\ F_\theta^{(r-n_1-n_2, D-d+1)}(Y) - F_\theta^{(r-n_1, D-d+1)}(Y-1) & \text{else if } n_2 < \min(r - n_1, D - n_3 - r + 1) \\ 1 - F_\theta^{(r-n_1, D-d+1)}(Y-1) & \text{else if } r - n_1 \leq n_2 < D - n_3 - r + 1 \\ F_\theta^{(r-n_1-n_2, D-d+1)}(Y) & \text{else if } D - n_3 - r + 1 \leq n_2 < r - n_1 \\ 1 & \text{otherwise} \end{cases}$$

Data augmentation

For a random variable $Y \sim f_\theta$, we can update θ with Poisson data augmentation if

1. We can represent Y as being generated from a latent Poisson:

$$Y \mid Z \sim h(Y \mid Z) \text{ where } Z \sim \text{Poisson}(\mu)$$

2. We can sample from the conditional distribution in closed-form

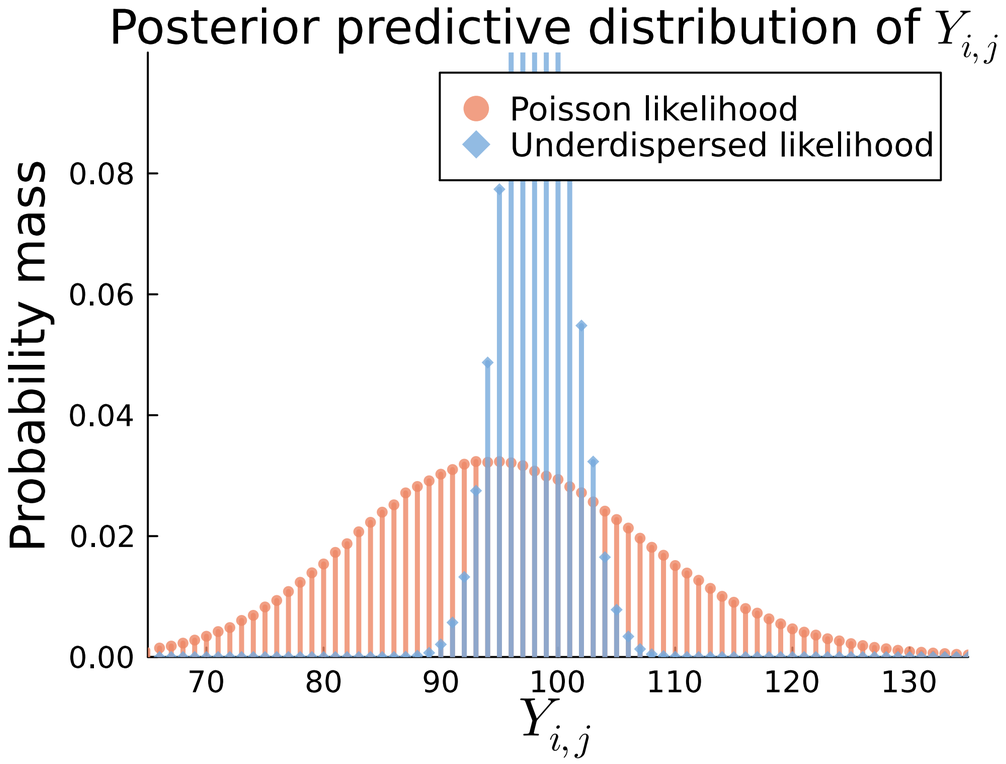
$$Z \mid Y \sim g(Z \mid Y, \theta)$$

Once we sample $Z \sim g(Z \mid Y, \theta)$, inference can proceed via Z, which is Poisson

For example, if $Y \sim$ NegativeBinomial(r,p), then if we sample $Z \mid Y \sim \text{CRT}(Y, r)$ then Z is marginally Poisson $Z \sim \text{Poisson}\left(r \frac{1}{1-p}\right)$

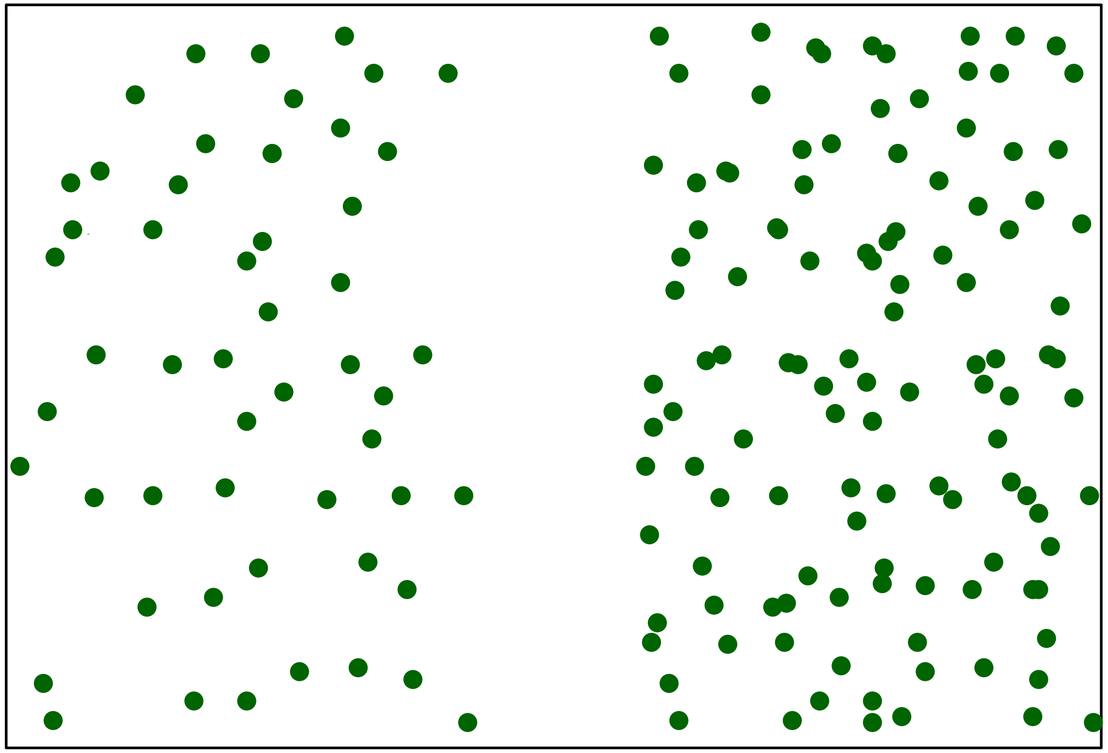
Why model underdispersion?

101	105	100	102	99
79	83	82	79	84
100	101	$Y_{i,j}$	102	98
80	81	78	82	84
99	103	98	100	100



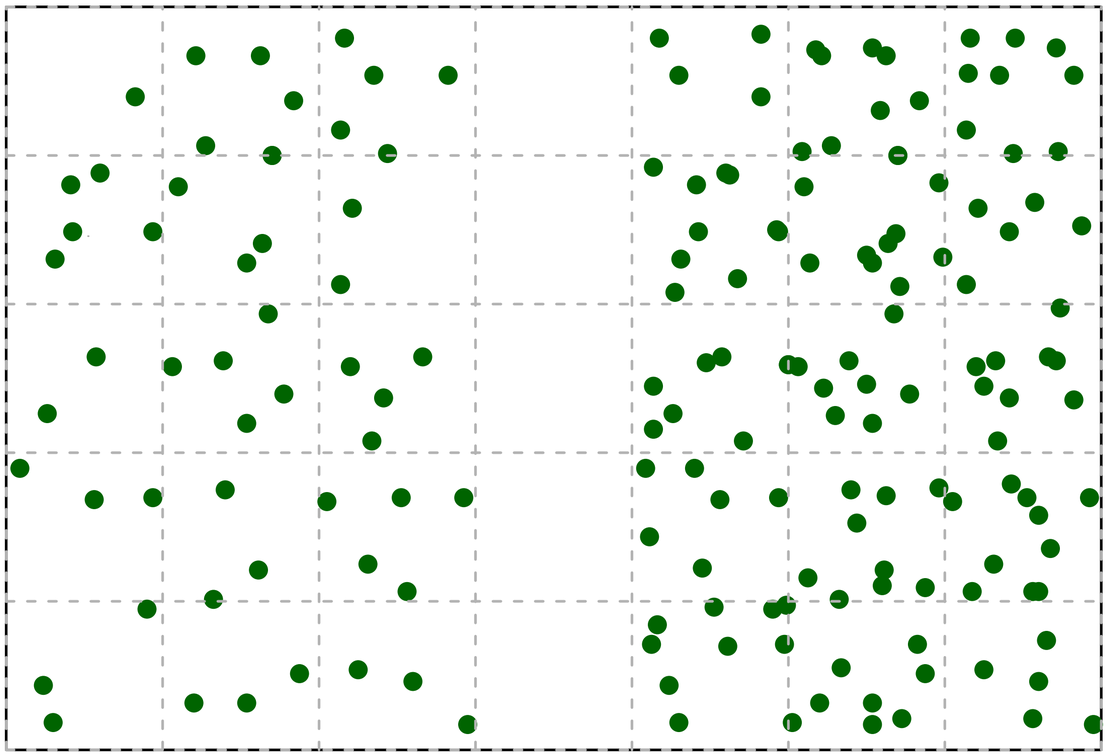
Capturing underdispersion allows us to make more precise probabilistic predictions than a Poisson likelihood would allow

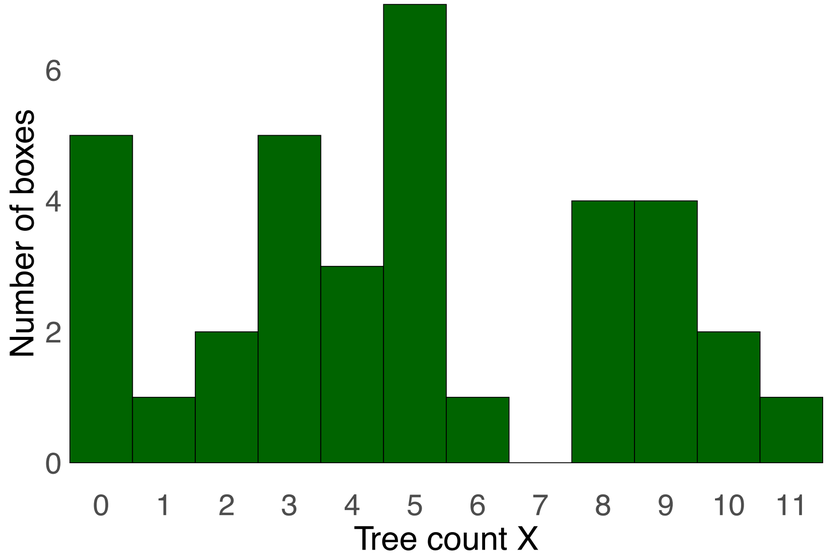
Conditional underdispersion



Conditional underdispersion

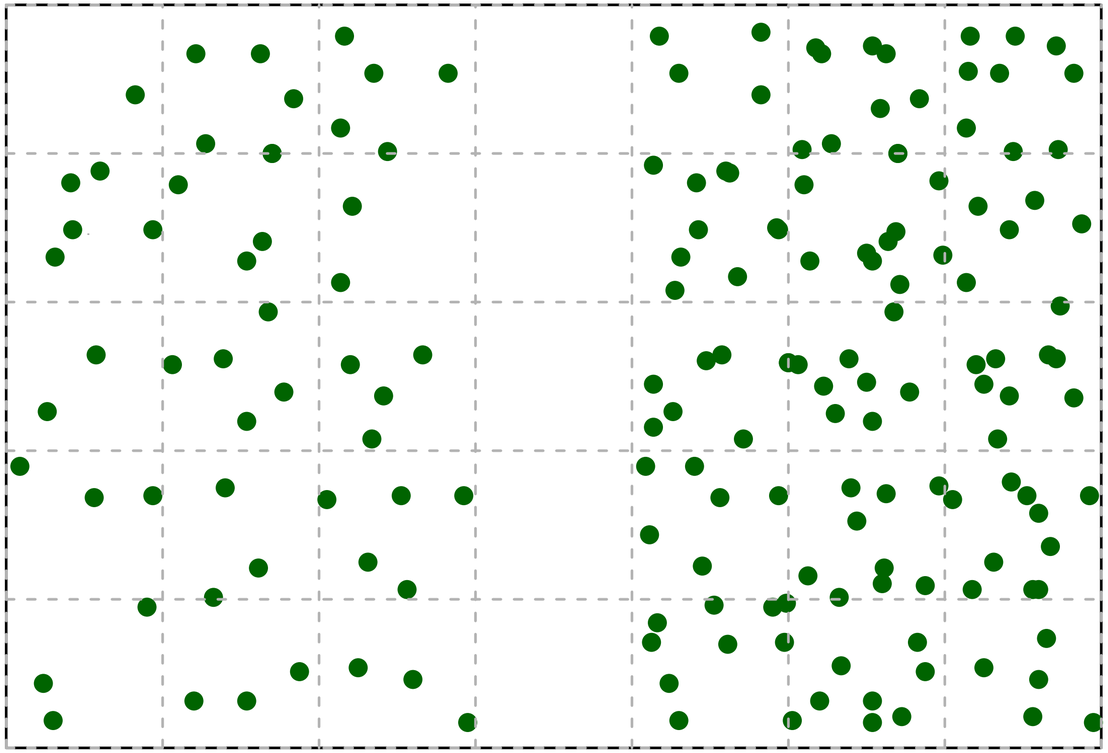
$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 2.20$$

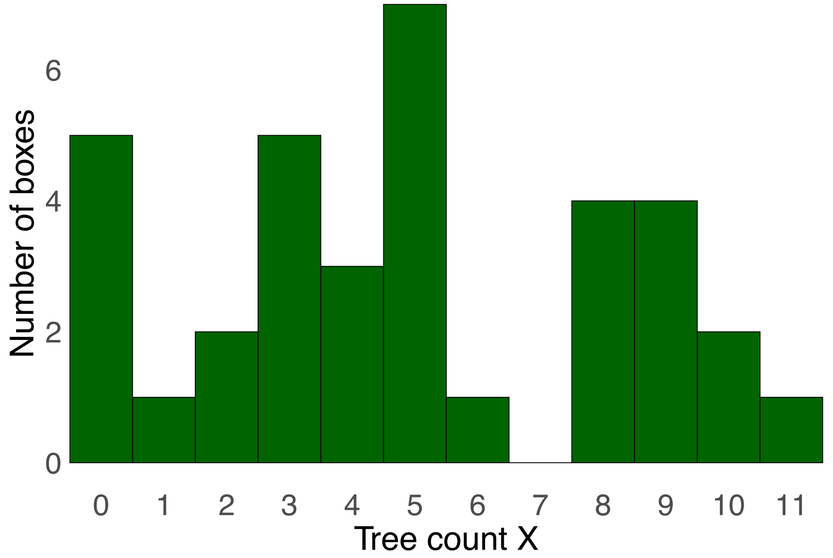




Conditional underdispersion

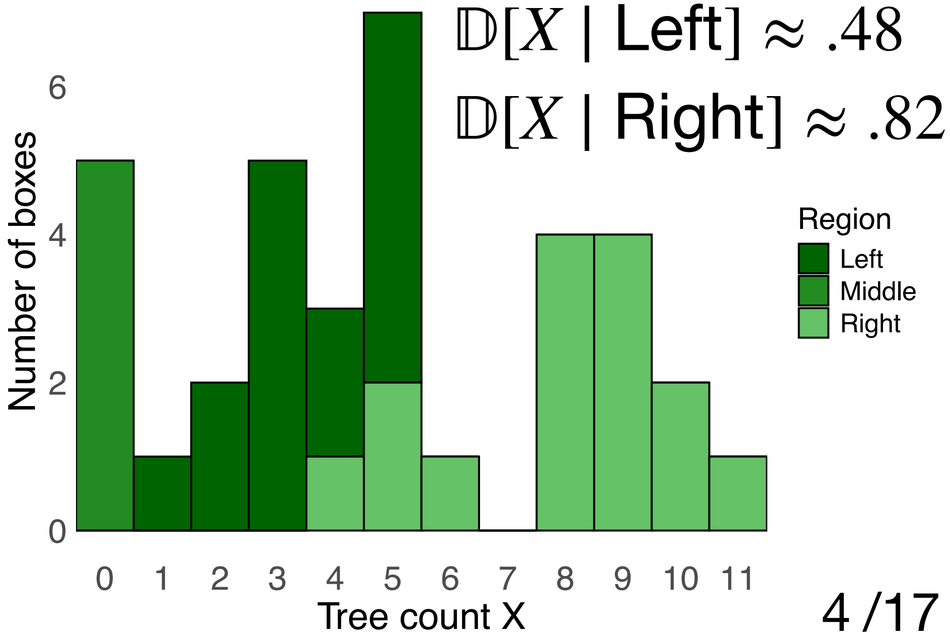
$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 2.20$$





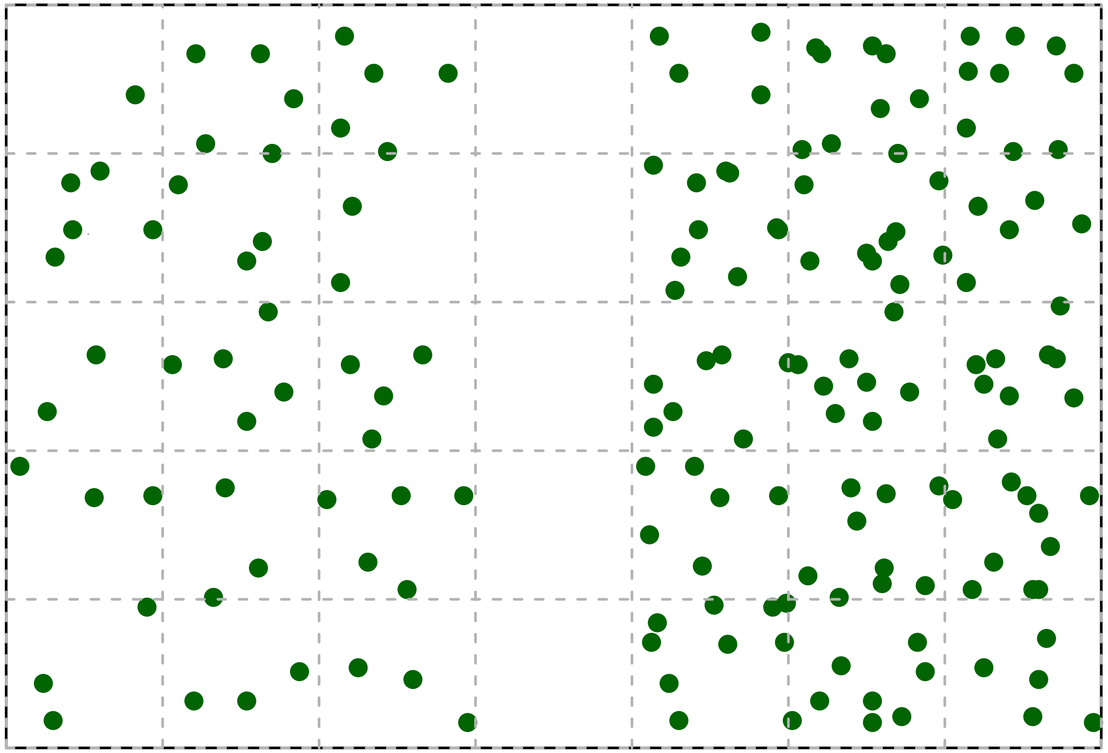
$$\mathbb{D}[X \mid \text{Left}] \approx .48$$

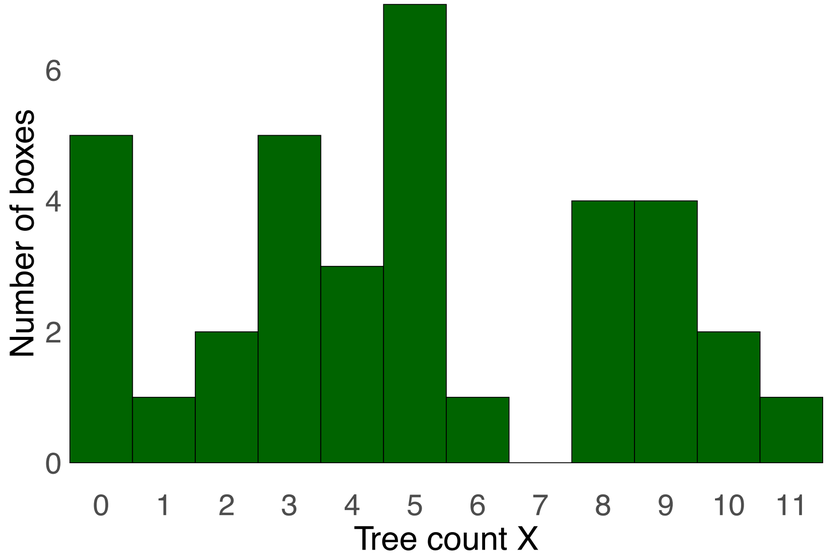
$$\mathbb{D}[X \mid \text{Right}] \approx .82$$



Conditional underdispersion

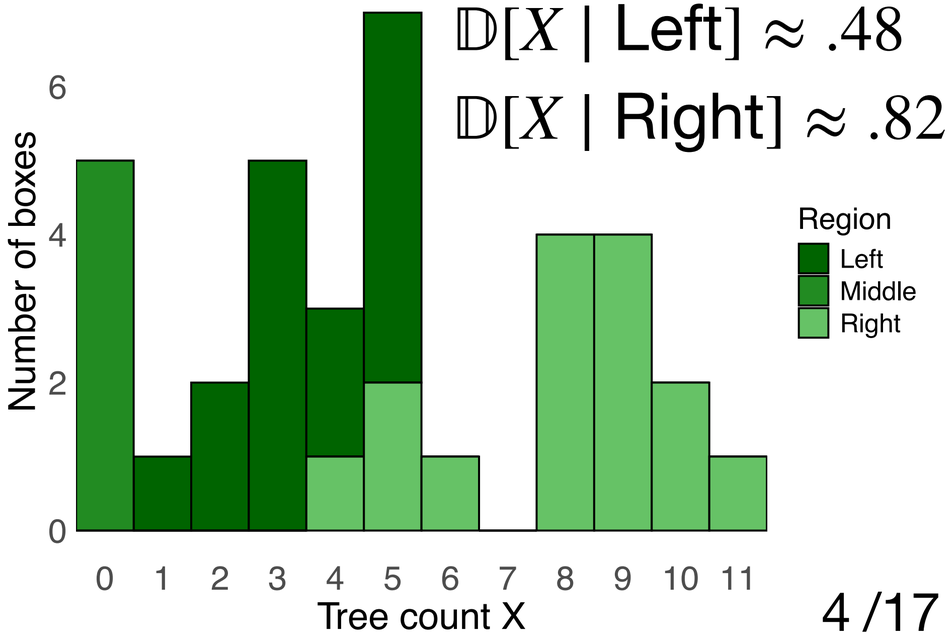
$$\mathbb{D}[X] = \frac{\mathbb{V}[X]}{\mathbb{E}[X]} \approx 2.20$$



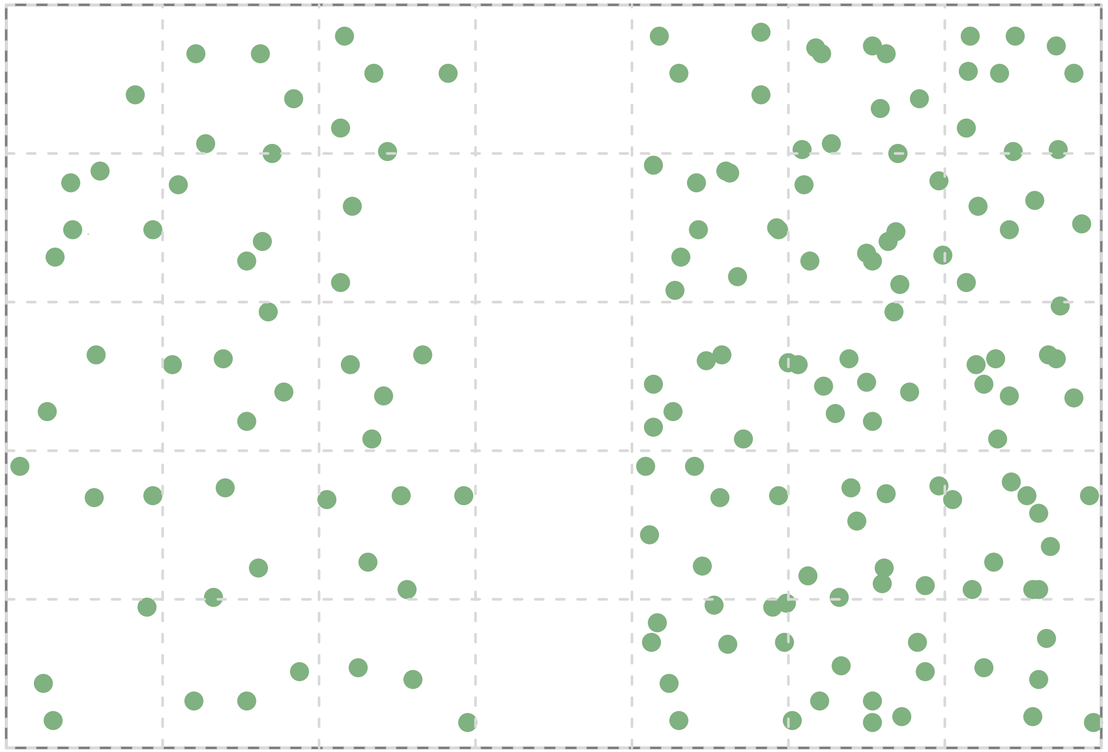


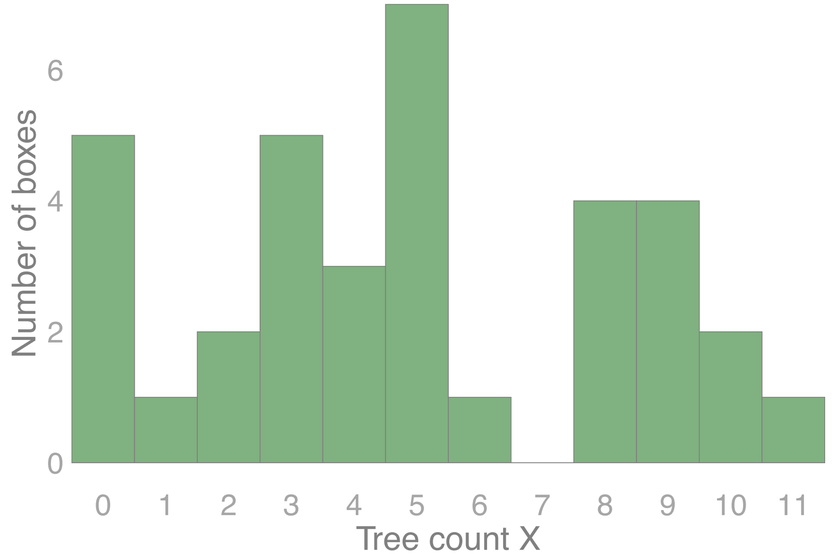
$$\mathbb{D}[X \mid \text{Left}] \approx .48$$

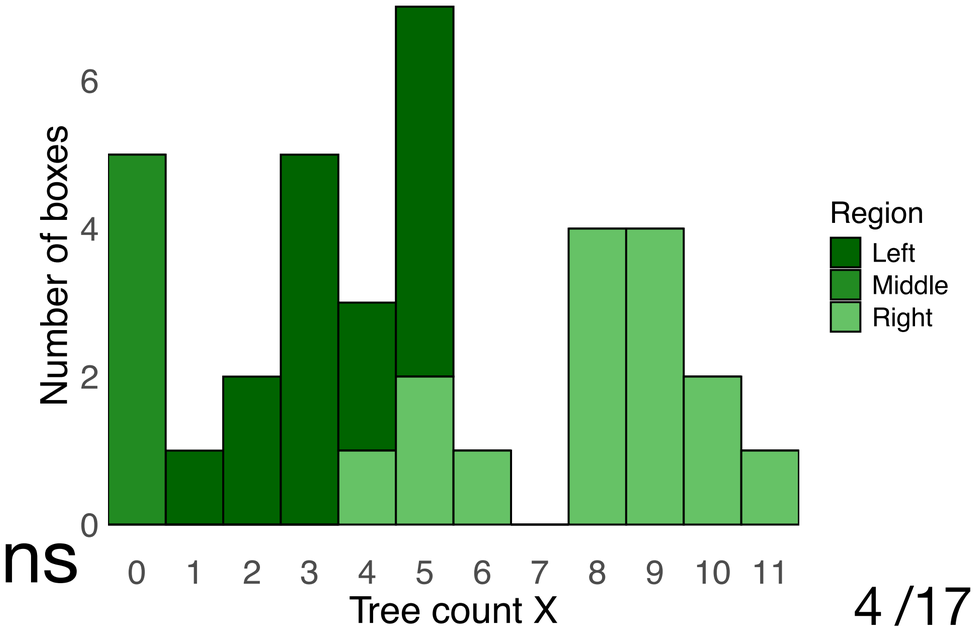
$$\mathbb{D}[X \mid \text{Right}] \approx .82$$



Marginal overdispersion can mask conditional underdispersion







Latent structure can reveal more regular patterns

Inference Summary

First:

$$\mathbf{Z}_{1:D} \sim P_\theta(\mathbf{Z}_{1:D} \mid Z^{(r,D)} = Y)$$

Second:

$$P(\theta \mid \mathbf{Z}_{1:D}) \propto g(\theta) \prod_{d=1}^{D} f_\theta(Z_d)$$

Algorithm B Exact simulation of $\mathbf{Z}_{1:D} \sim P_\theta(\mathbf{Z}_{1:D} \mid Z^{(r,D)}=Y)$

Input: observation $Y \in \mathbb{Z}$, order $D \in \mathbb{N}$, rank $r \in [D]$, parent distribution f_θ

Initialize: $N_0^{(<Y)} = N_0^{(=Y)} = N_0^{(>Y)} = 0$

for $d = 1 \dots D$ **do**

  Compute $[p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)}]$ // as defined in Equation (6)

  Sample $C_d \sim \text{Categorical}\left(p_d^{(<Y)}, p_d^{(=Y)}, p_d^{(>Y)}\right)$ // where $C_d \in \{<Y, =Y, >Y\}$

  $N_d^{(c)} \leftarrow N_{d-1}^{(c)} + \mathbb{1}\{C_d = c\}$ for $c \in \{<Y, =Y, >Y\}$ // update sufficient stats

  Sample $Z_d \sim \underset{\mathbb{Z}_{C_d}}{\text{trunc}}\, f_\theta$ // as given in Theorem 5.1

  // (Optional) assess/execute the break conditions in Theorem 5.2

Output: $\{Z_1, \dots, Z_D\}$

Using this data augmentation scheme, we can build Bayesian models with Poisson order statistic distributions