Maximum Likelihood Methods
for the Social Sciences
POLS 510 · CSSS 510

Models of Count Data

Political Science *and* CSSS
University of Washington, Seattle

Christopher Adolph

Concepts for Count Data

Over the next few lectures, we will learn to model count data

bounded counts unbounded counts contaminated counts

We will also learn and apply new MLE concepts

reparameterization compound probability distributions

overdispersion mixed probability distributions

generalized linear models censored probability distributions

quasilikelihood truncated probability distributions

Watch for these concepts throughout the lecture

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

The number of voters who turned out to vote in each of N voting districts

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

The number of voters who turned out to vote in each of N voting districts

The number of people who fall into each of N cells of a contingency table

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

The number of voters who turned out to vote in each of N voting districts

The number of people who fall into each of N cells of a contingency table

Note that counts can be cross-sectional or time series

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

The number of voters who turned out to vote in each of N voting districts

The number of people who fall into each of N cells of a contingency table

Note that counts can be cross-sectional or time series

They can have clear upper bounds, vague upper bounds, or no upper bounds

Count Data

Event counts could include

The number of email messages sent by a person in each of N days

The number of attacks carried out in each of N months of a guerilla war

The number of people who got sick from the flu in each of N cities

The number of voters who turned out to vote in each of N voting districts

The number of people who fall into each of N cells of a contingency table

Note that counts can be cross-sectional or time series

They can have clear upper bounds, vague upper bounds, or no upper bounds

Generally,
counts are aggregates (sums) of events
whose individual-level data generating processes are unobserved

Counts with an upper bound

Consider a count p with an upper bound

For each of N students,
the number p of correctly answered items on a test with M questions

[Note the number of questions M has no upper bound]

Counts with an upper bound

Consider a count p with an upper bound

For each of N students,
the number p of correctly answered items on a test with M questions

[Note the number of questions M has no upper bound]

Another count p with an upper bound

For each of N voting districts,
the number of citizens p voting out of M registered to vote

[Note the number of citizens M has no upper bound]

For the counts in yellow we can use models for "grouped counts"

For counts in pink, these models won't work

Counts with an upper bound: binomial regression

Assumptions:

1. A count consists of a sum of M binary variables

Counts with an upper bound: binomial regression

Assumptions:

1. A count consists of a sum of M binary variables

2. Each of the binary variables is iid Bernoulli

Counts with an upper bound: binomial regression

Assumptions:

1. A count consists of a sum of M binary variables

2. Each of the binary variables is iid Bernoulli

These assumptions lead to the binomial distribution, a generalization of the Bernoulli

Counts with an upper bound: binomial regression

Assumptions:

1. A count consists of a sum of M binary variables

2. Each of the binary variables is iid Bernoulli

These assumptions lead to the binomial distribution, a generalization of the Bernoulli

If we have the binary variables, we could just use logit or probit

Counts with an upper bound: binomial regression

Assumptions:

1. A count consists of a sum of M binary variables

2. Each of the binary variables is iid Bernoulli

These assumptions lead to the binomial distribution, a generalization of the Bernoulli

If we have the binary variables, we could just use logit or probit

But if that is infeasible, because the binary variables are lost or very numerous,
we use binomial regression $\rightarrow$ just another kind of logit or probit

Binomial Likelihood

Start with the binomial distribution

$$\Pr(y_i|\pi_i, M_i) = \frac{M_i!}{y_i!(M_i - y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Binomial Likelihood

Start with the binomial distribution

$$\Pr(y_i|\pi_i, M_i) = \frac{M_i!}{y_i!(M_i - y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Note that $\mathbb{E}(y_i) = \mu_i = M_i\pi_i, \qquad \mathrm{Var}(y_i) = M_i\pi_i(1-\pi_i)$

Binomial Likelihood

Start with the binomial distribution

$$\Pr(y_i|\pi_i, M_i) = \frac{M_i!}{y_i!(M_i - y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Note that $\mathbb{E}(y_i) = \mu_i = M_i\pi_i, \qquad \mathrm{Var}(y_i) = M_i\pi_i(1-\pi_i)$

We can model the systematic component using either

- $M_i\pi_i$, the expected sum of 1s across the grouped events, or

Binomial Likelihood

Start with the binomial distribution

$$\Pr(y_i|\pi_i, M_i) = \frac{M_i!}{y_i!(M_i - y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Note that $\mathbb{E}(y_i) = \mu_i = M_i\pi_i, \qquad \mathrm{Var}(y_i) = M_i\pi_i(1-\pi_i)$

We can model the systematic component using either

- $M_i\pi_i$, the expected sum of 1s across the grouped events, or
- π_i, the probability of an individual event within the grouped events

Binomial Likelihood

Start with the binomial distribution

$$\Pr(y_i|\pi_i, M_i) = \frac{M_i!}{y_i!(M_i - y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Note that $\mathbb{E}(y_i) = \mu_i = M_i\pi_i, \qquad \mathrm{Var}(y_i) = M_i\pi_i(1-\pi_i)$

We can model the systematic component using either

- $M_i\pi_i$, the expected sum of 1s across the grouped events, or
- π_i, the probability of an individual event within the grouped events

We will use simply π, to keep interpretation simple and logit-like

Binomial Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \prod_{i=1}^{N} \frac{M_i!}{y_i!(M_i - y_i)!} \pi_i^{y_i}(1-\pi_i)^{M_i - y_i}$$

Binomial Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \prod_{i=1}^{N} \frac{M_i!}{y_i!(M_i-y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i-y_i}$$

Take logs

$$\log \mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \sum_{i=1}^{N}\left\{\log \frac{M_i!}{y_i!(M_i-y_i)!} + y_i \log \pi_i + (M_i-y_i)\log(1-\pi_i)\right\}$$

Binomial Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \prod_{i=1}^{N} \frac{M_i!}{y_i!(M_i-y_i)!}\,\pi_i^{y_i}(1-\pi_i)^{M_i-y_i}$$

Take logs

$$\log \mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \sum_{i=1}^{N}\left\{\log \frac{M_i!}{y_i!(M_i-y_i)!} + y_i \log \pi_i + (M_i-y_i)\log(1-\pi_i)\right\}$$

Reduce to sufficient statistics

$$\log \mathcal{L}(\boldsymbol{\pi}|\mathbf{y},\mathbf{M}) = \sum_{i=1}^{N}\{y_i \log \pi_i + (M_i-y_i)\log(1-\pi_i)\}$$

Binomial Likelihood

Assume π_i follows an inverse-logit

$$[1 + \exp(-x_i\beta)]^{-1}$$

Note: we could assume any functional form we like and get a different flavor of binomial regression; e.g., binomial probit, binomial cloglog, binomial scobit, etc.

Substitute for π_i and the systematic component

$$\log \mathcal{L}(\boldsymbol{\pi}|\mathbf{y}, \mathbf{M}) = \sum_{i=1}^{N} \{y_i \log [1 + \exp(-x_i\beta)] + (M_i - y_i) \log (1 - [1 + \exp(-x_i\beta)])\}$$

We could code this into `R` and estimate the binomial model using `optim()`

Or we could use `glm(model, data, family=binomial)`

Binomial Regression Interpretation

Binomial coefficients are (just like) logit coefficients –
assuming, of course, a logit specification for π

Could use similar tricks to interpret them:

- Exponentiate to see the effect of a unit change in x on the log odds
- Calculate expected probabilities, first differences,
  or relative risks for the underlying probability of success (now, the rate of success)
- Multiply expected probabilities (first differences) by a particular M
  to see the expected (change in) count given a hypothetical number of trials

Binomial Regression Goodness of Fit

Can use the same techniques to test GoF we used for logit,
or for GLMs more generally:

- Likelihood ratio tests, AIC, BIC
- Residual plots – more useful now if some cases have much larger M
- Actual vs Predicted plots, percent correctly predicted, *etc.*
- *New tricks:* mean absolute error, root mean squared error

MAE and RMSE are the most intuitive, useful, and widely used metrics for fit

Now that our outcome is less restricted to a narrow, discrete range,
they will be more useful

Binomial Regression: 2004 Washington Governor's Race

Recall our earlier binomial distribution example:
the number of voters who turned out in each of the 39 Washington counties in 2004

Our outcome variable has two parts

voters – the count of registered voters who turned out

non-voters – the count of registered voters who stayed home

Let's expand the example to include covariates; either might raise turnout

income – the median household income in the county in 2004

college – the % of residents over 25 with at least a college degree in 2005

The last covariate is only available for the 18 largest counties

I use multiple imputation to fill in the missings

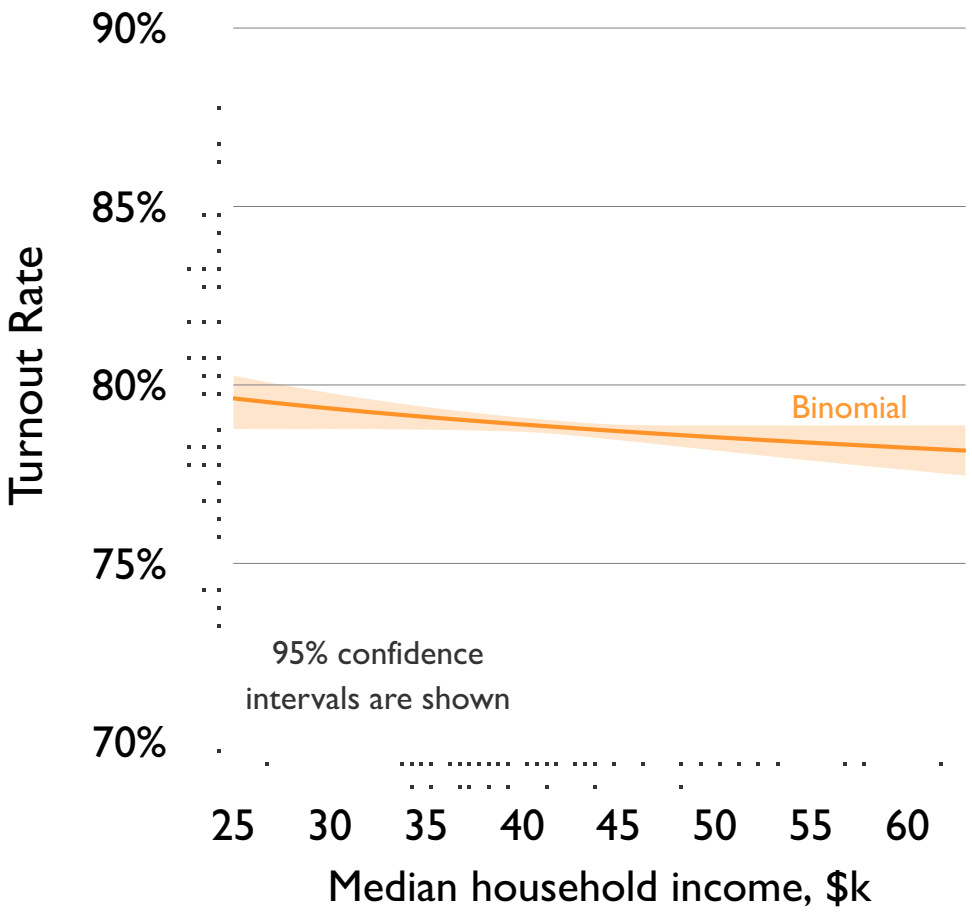
	Least Squares	Binomial
log Income	-0.05	-0.10
	(0.05)	(0.05)
College	0.27	0.82
	(0.12)	(0.09)
Intercept	1.28	2.17
	(0.55)	(0.49)
N	39	39
log $\mathcal{L}$	—	-7437
AIC	—	14881
In-sample Mean Absolute Error (null=3.25%)	2.84%	2.96%
5-fold cross-validated MAE (null=3.34%)	3.13%	3.28%

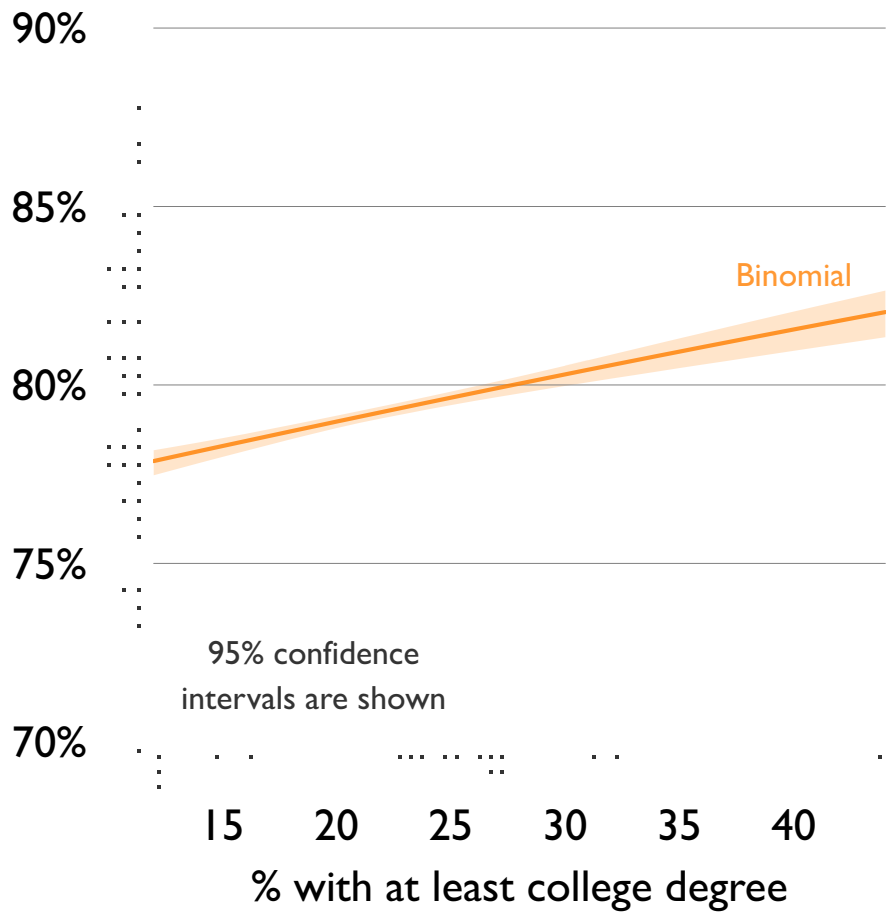
Results above combined across 100 imputations

Parameters are not directly comparable – linear vs. logit coefficients

Binomial regression finds a significant result for income – linear regression doesn't

Linear regression actually predicts better in cross-validation



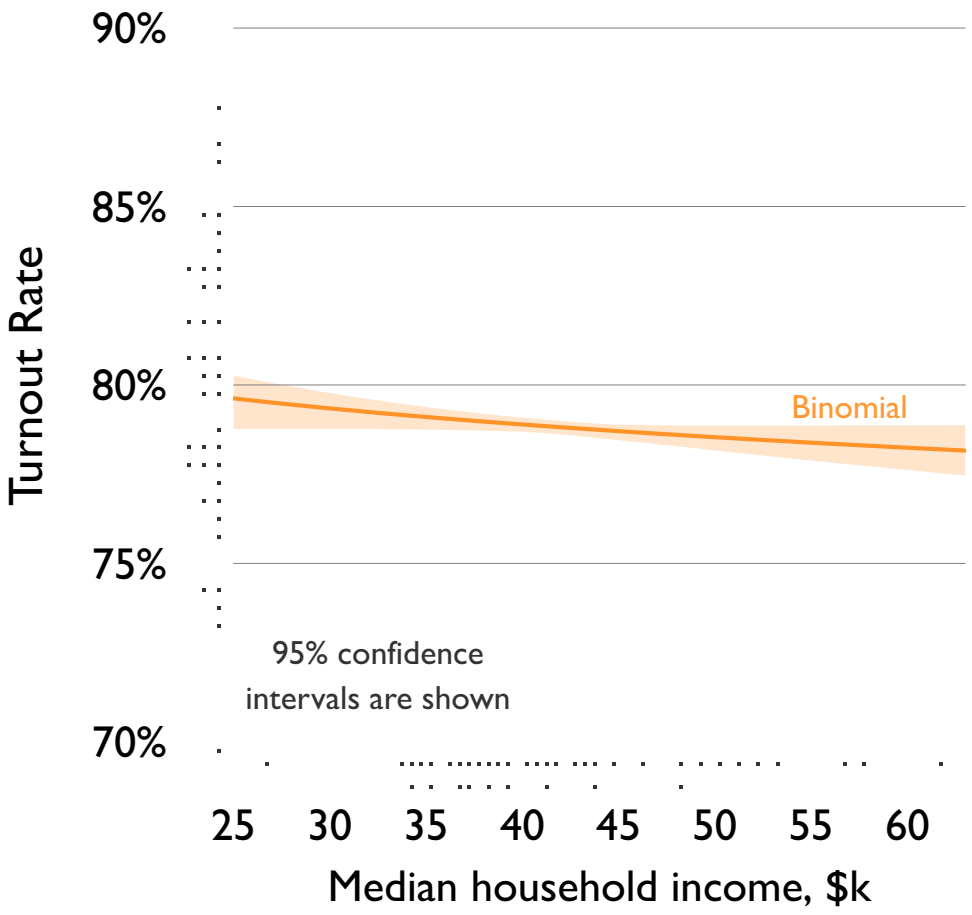


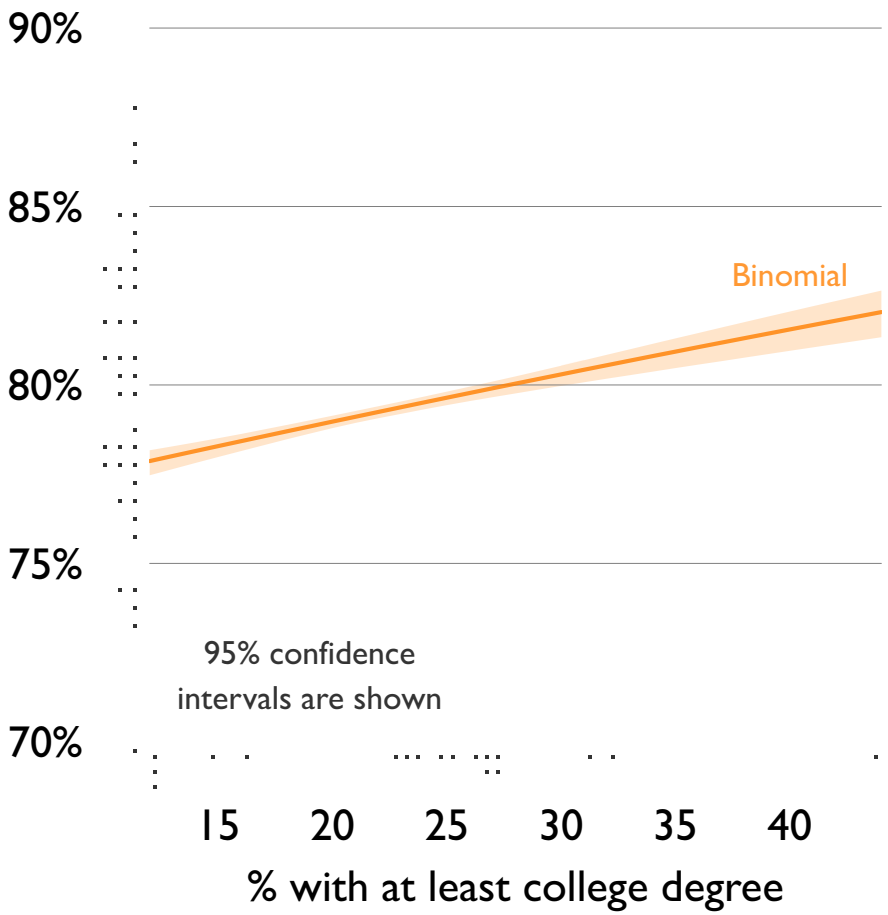
Interpretation of the model is just like logit and shows expected *rates* of voting

Use `logitsimev()`, `logitsimfd()`, and `logitsimrr()` as usual

What if we want the expected count for a county instead of the expected rate?

Just multiply the output of `logitsimev()` by M_{hyp} *still an MLE*

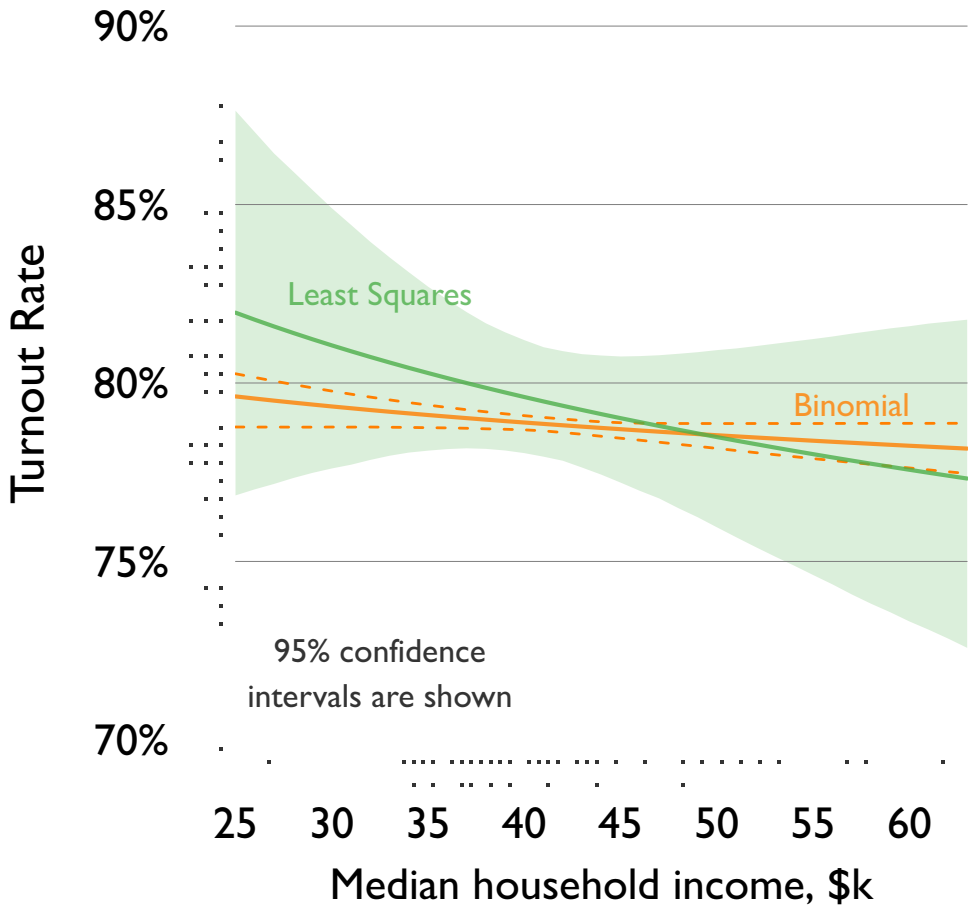


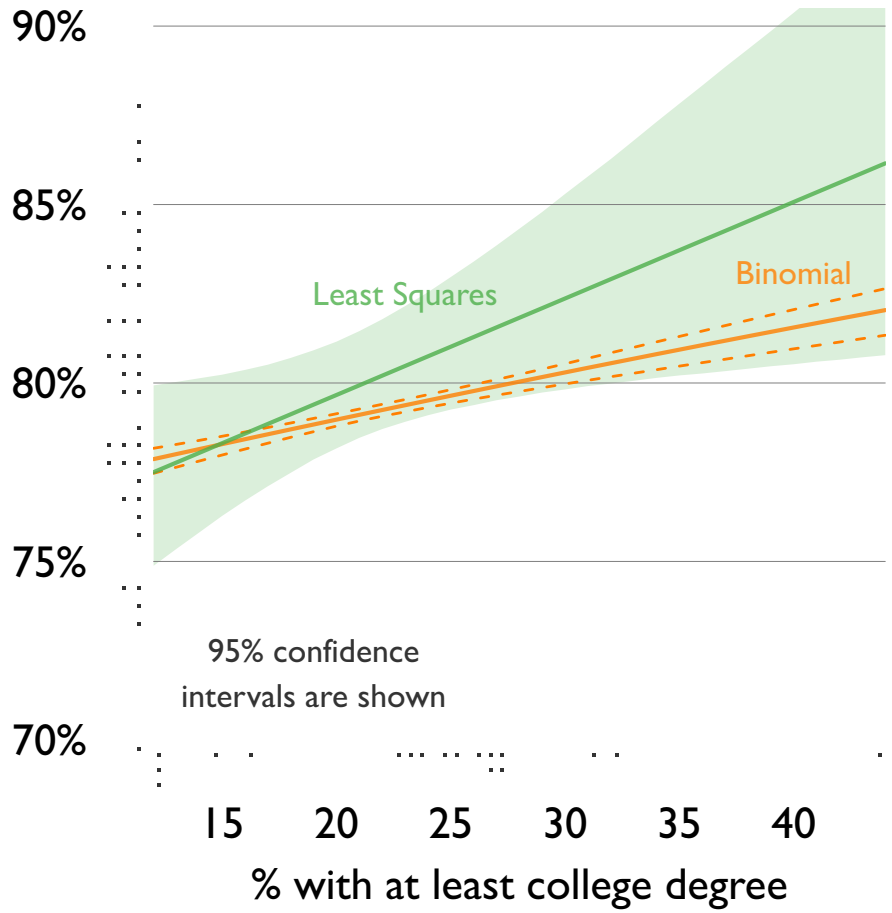


High income are slightly but significantly *less* likely to vote (!)

Also a moderate and very significant link between education and turnout

Are precise results reasonable for a model with 39 observations & imputation?

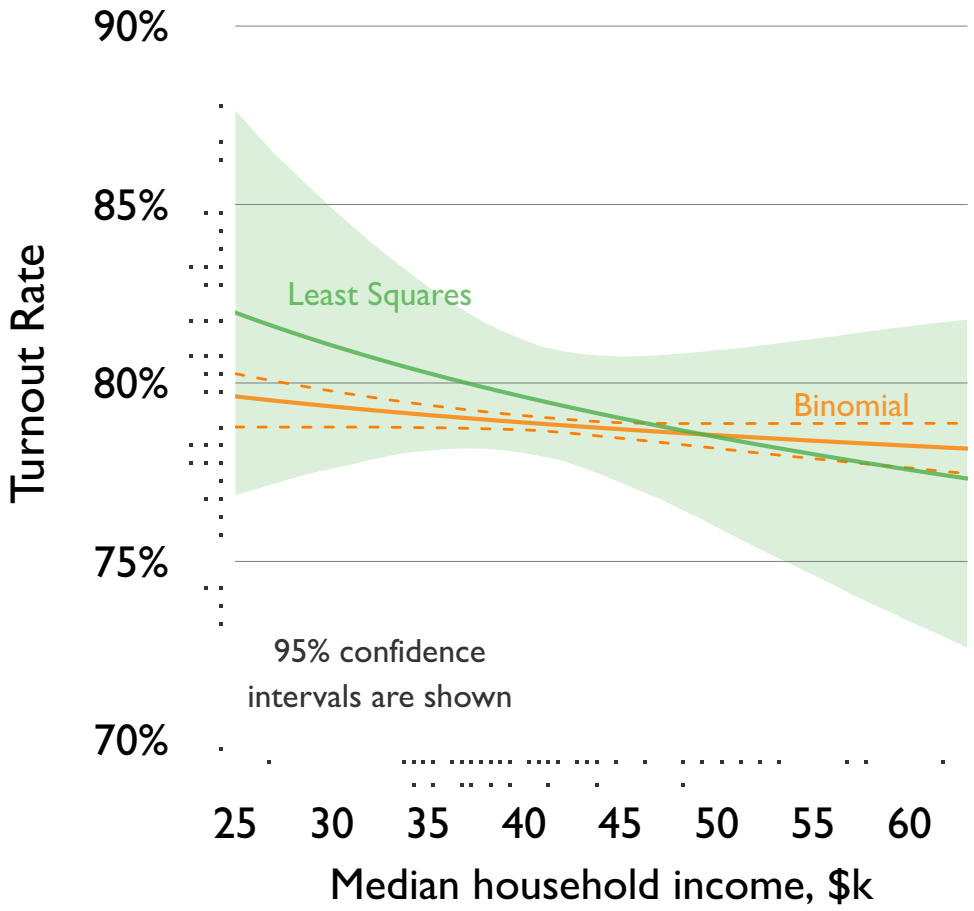


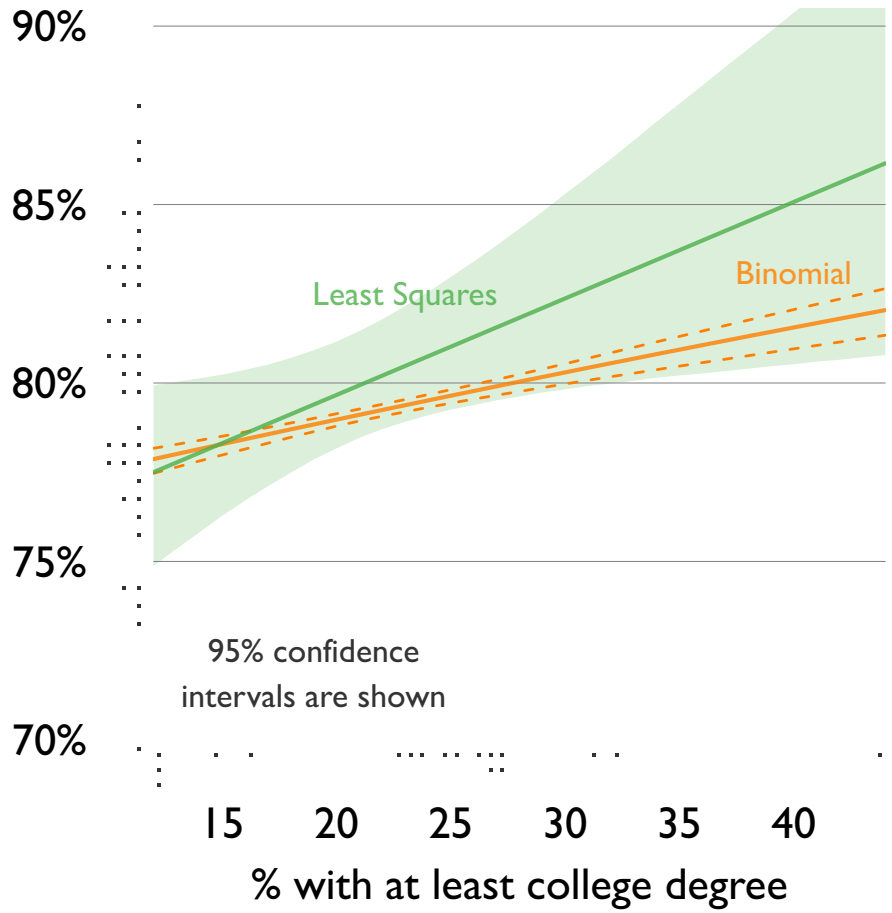


Linear regression has *much* wider CIs: income no longer remotely significant

Binomial regression assumes each individual is an independent trial, so in a sense there are *thousands* of observations

Linear regression treats each county as just 1 case despite aggregating many individual choices

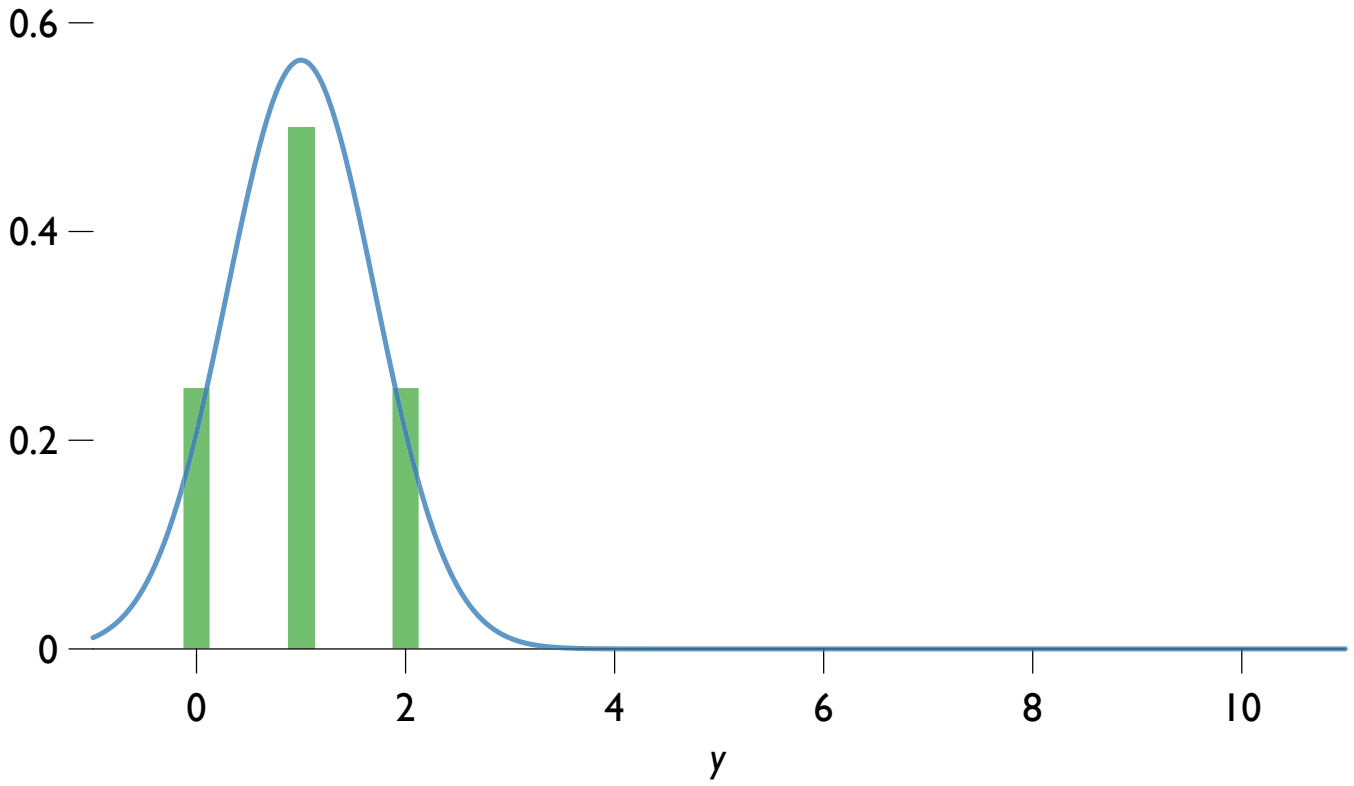




Problems whether we use least squares or the Binomial MLE

1. Linear regression weights counties with different populations equally– we might try to fix with population weights, but lack a principled way to do so

2. Binomial regression will be *overconfident* if individual turnout decisions are influenced by unmeasured common factors *or* other people's turnout



Binomial(π = 0.5, M = 2)

E(y) = $M\pi$ = 1

var(y) = $M\pi(1-\pi)$ = 0.5

Normal(μ = 1, σ^2 = 0.5)

E(y) = μ = 1

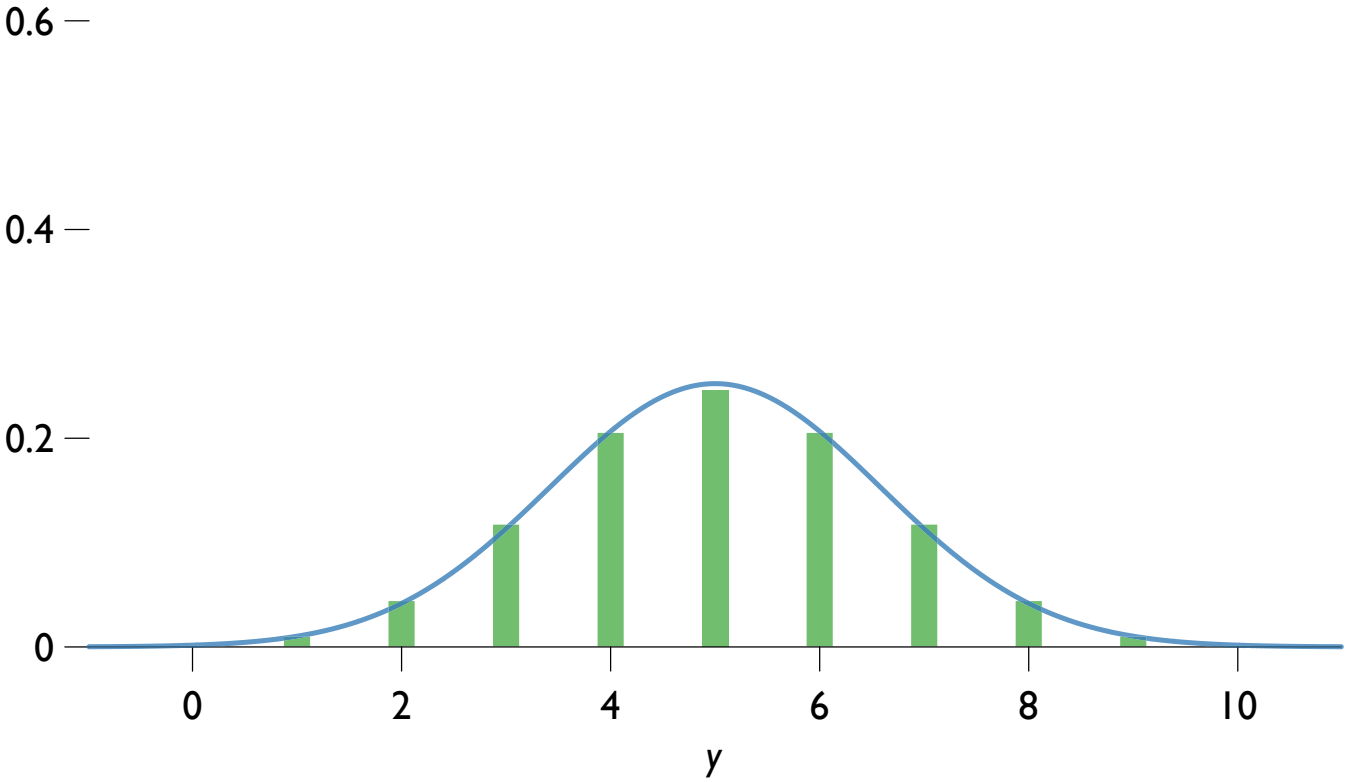
var(y) = σ^2 = 0.5

Why does Binomial regression tend to produce overconfident estimates?

Recall the Binomial represent the probability of sums for iid Bernoulli trials

Above is the pdf for a Binomial($\pi = 0.5$, $M = 2$) and its Normal approximation

When we match the moments of the Normal to the Binomial,
the pdfs are similar, except the Binomial is discrete



Binomial($\pi = 0.5$, $M = 10$)

$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

Normal($\mu = 5$, $\sigma^2 = 2.5$)

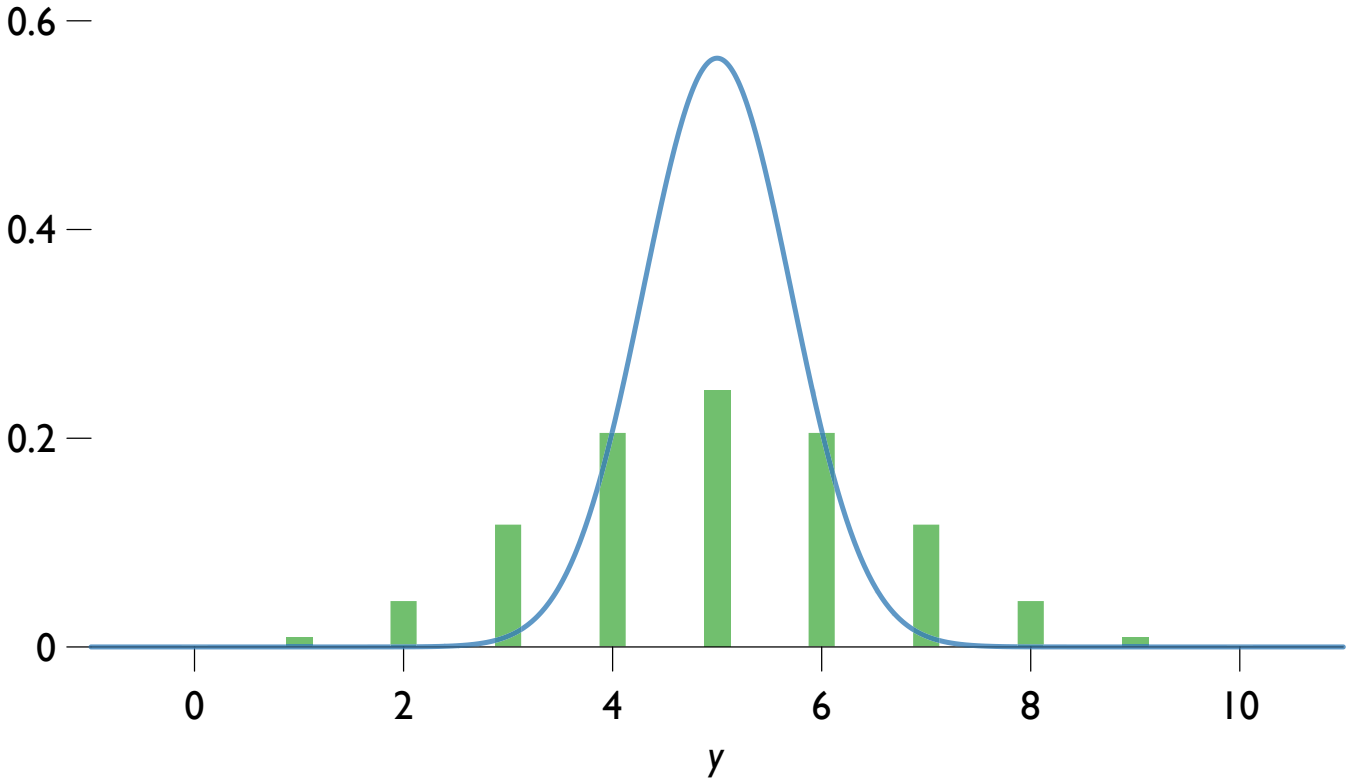
$E(y) = \mu = 5$

$\text{var}(y) = \sigma^2 = 2.5$

We can likewise match a Binomial with $M = 10$ trials to an appropriate Normal

Note the Normal has two free parameters, μ and σ,
and the Binomial only one, π

Why? Given iid trials, fixing the probability of success in 1 of the M trials determines the overall mean *and* variance of the sum of successful trials



Binomial($\pi = 0.5$, $M = 10$)

$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

Normal($\mu = 5$, $\sigma^2 = 0.5$)

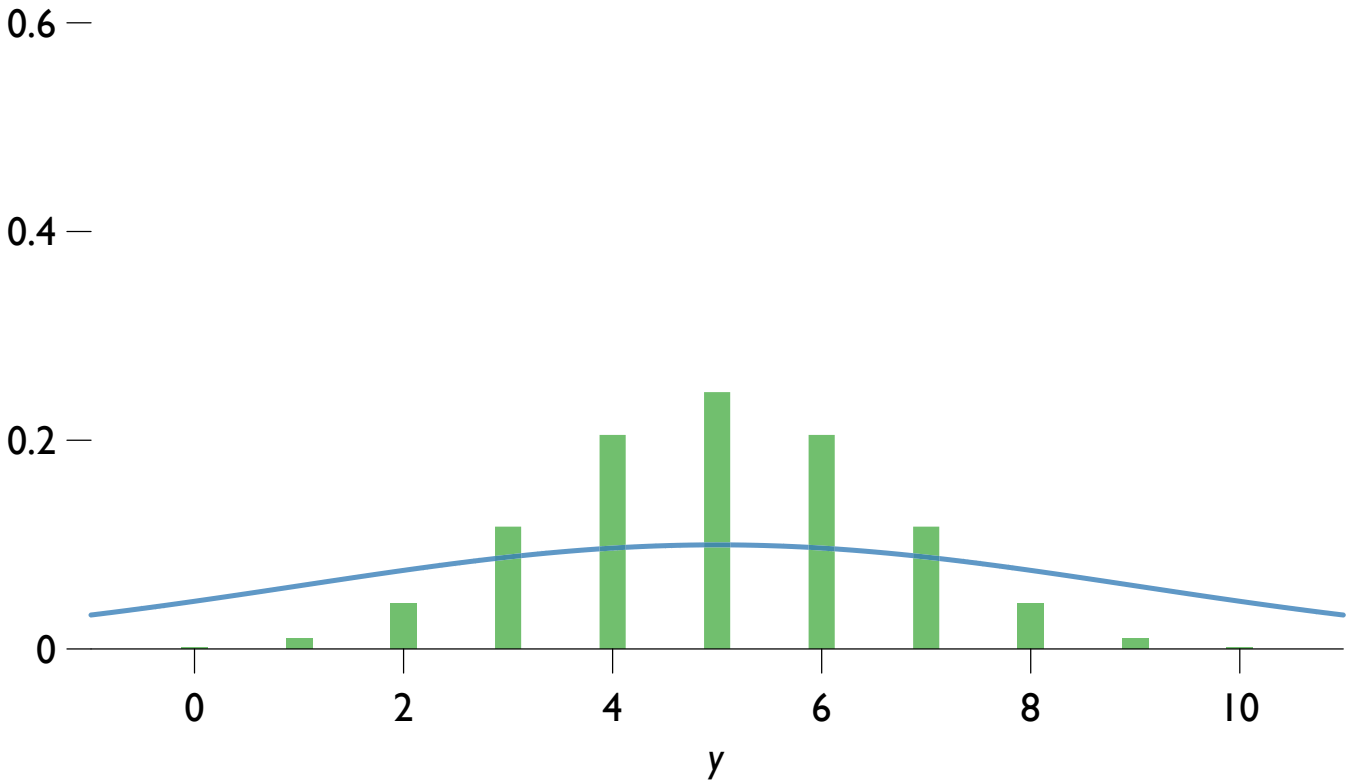
$E(y) = \mu = 5$

$\text{var}(y) = \sigma^2 = 0.5$

This means a Binomial over 10 trials with $\pi = 0.5$ with always have mean $10 \times 0.5 = 5$ and variance $10 \times 0.5 \times (1 - 0.5) = 2.5$

The Normal has two free parameters: when the mean of the Normal is 5, we can set the variance to anything at all

A Normal could have *less* dispersion than a Binomial with the same mean



Binomial($\pi = 0.5$, $M = 10$)

$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

Normal($\mu = 5$, $\sigma^2 = 16$)

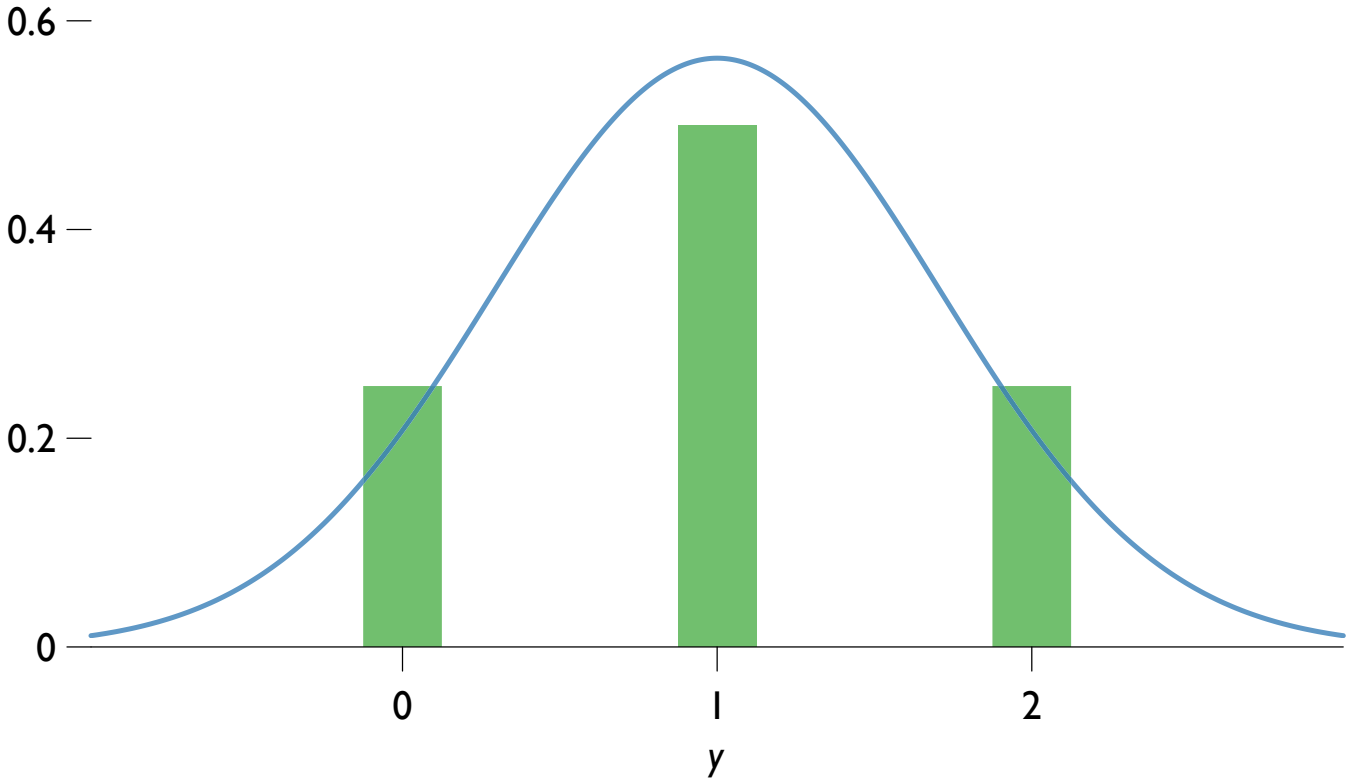
$E(y) = \mu = 5$

$\text{var}(y) = \sigma^2 = 16$

Or a Normal could have *more* dispersion than a Binomial with the same mean

The independence assumption drives the Binomial variance
to a specific level for a given mean

What if the independence assumption is wrong?



Binomial($\pi = 0.5, M = 2$)

$E(y) = M\pi = 1$

$\text{var}(y) = M\pi(1-\pi) = 0.5$

Normal($\mu = 1, \sigma^2 = 0.5$)

$E(y) = \mu = 1$

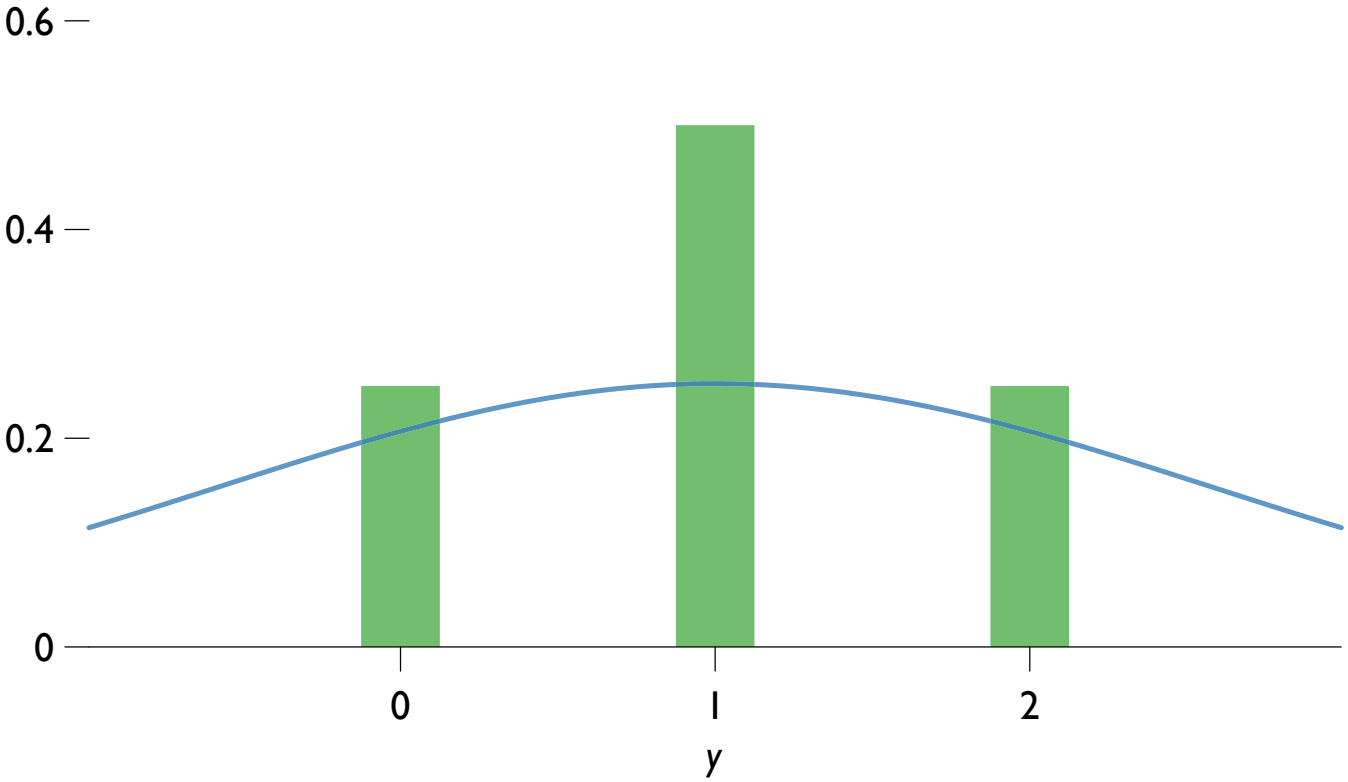
$\text{var}(y) = \sigma^2 = 0.5$

Let's explore the consequences of summing two *correlated* binary trials

Above are the Binomial and approximate Normal for this simple setup – equivalent to flipping two coins and counting the total heads

Independence implies 4 equally likely outcomes: {H,H}, {H,T}, {T,H}, and {T,T}

and produces a mean of 1 and a variance of 0.5



Binomial($\pi = 0.5, M = 2$)

$E(y) = M\pi = 1$

$\text{var}(y) = M\pi(1-\pi) = 0.5$

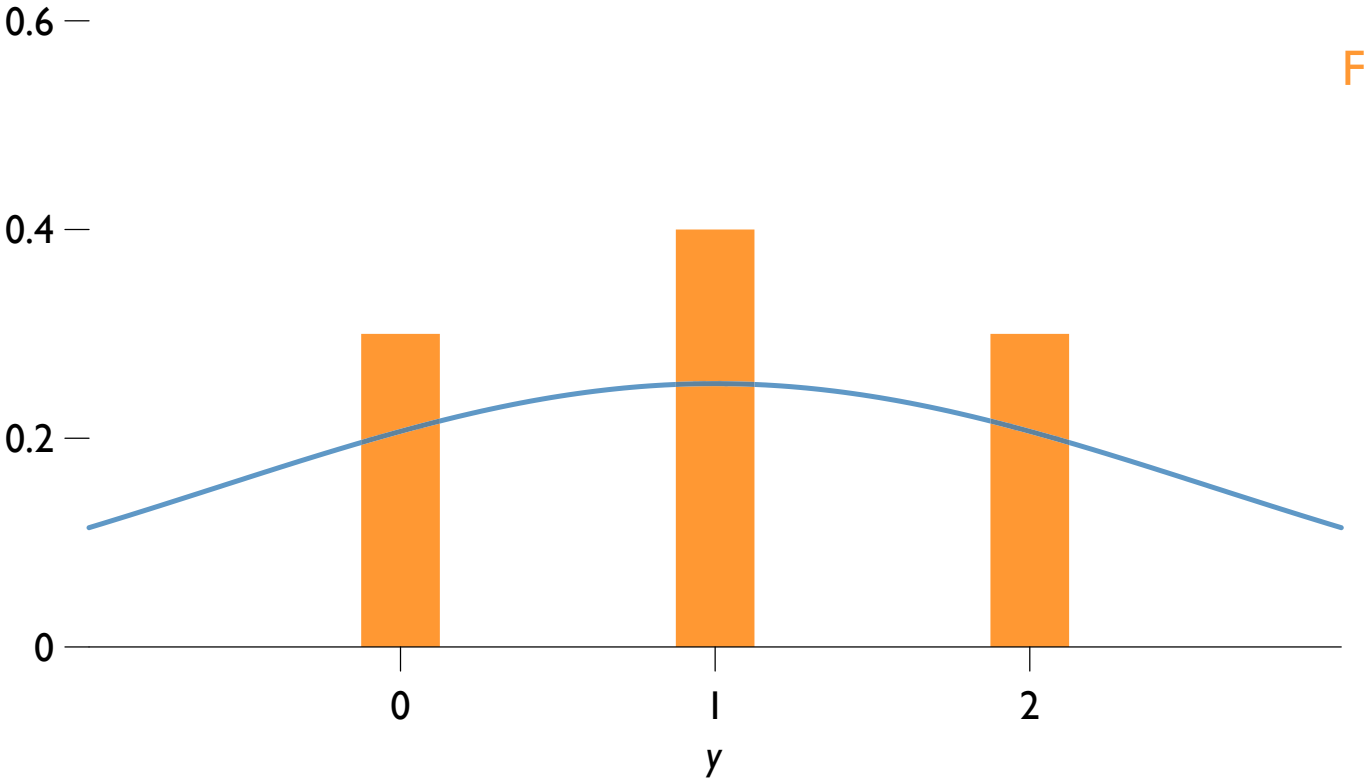
Normal($\mu = 1, \sigma^2 = 2.5$)

$E(y) = \mu = 1$

$\text{var}(y) = \sigma^2 = 2.5$

Recall that the Normal can be overdispersed relative to the Binomial

But there is no way with a strict Binomial to show greater dispersion than occurs under independent trials



Flip coin; match with $p = 0.6$

$E(y) = 0.5$

$\text{var}(y) = 0.60$

Normal($\mu = 1$, $\sigma^2 = 2.5$)

$E(y) = \mu = 1$

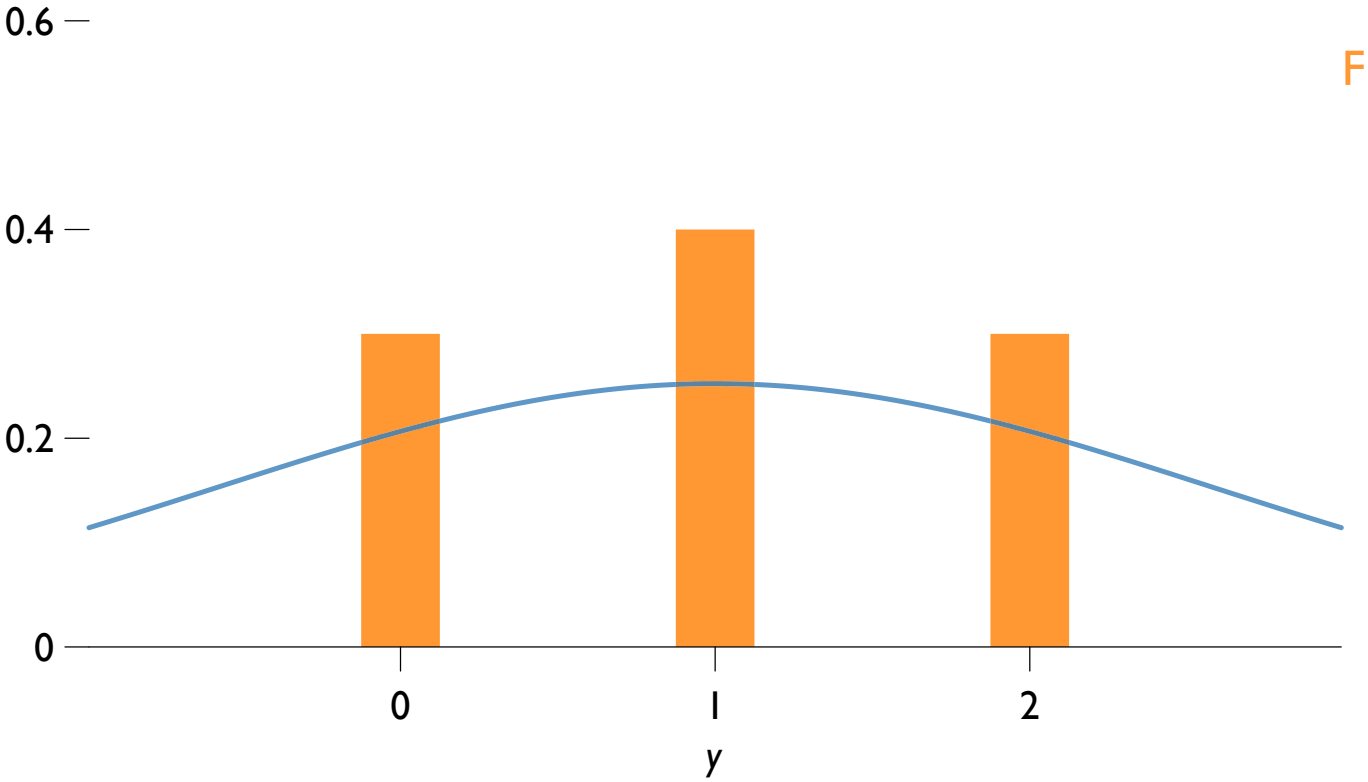
$\text{var}(y) = \sigma^2 = 2.5$

Suppose that first coin flipped is Bernoulli($\pi = 0.5$)

But the second coin is set to the "matching" side with probability 0.6

Thus we see two heads (or two tails) $0.5 \times 0.6 = 30\%$ of the time,
and mixed coins only 40% of the time

This leads to greater dispersion than for two iid Bernoulli coin flips



Flip coin; match with $p = 0.6$

$E(y) = 0.5$

$\text{var}(y) = 0.60$

$\text{Normal}(\mu = 1, \sigma^2 = 2.5)$

$E(y) = \mu = 1$

$\text{var}(y) = \sigma^2 = 2.5$

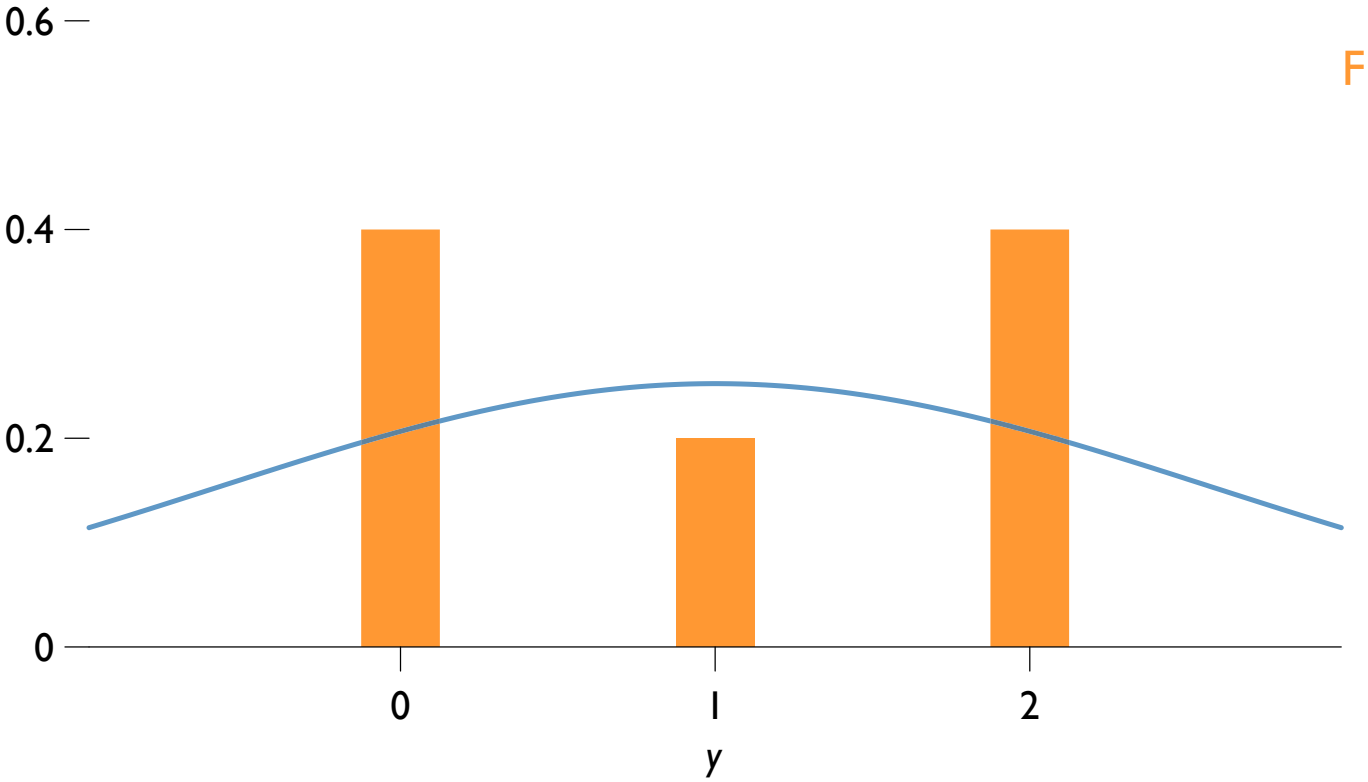
Why is the "matched coins" experiment overdispersed?

Because the average probability across the M trials varies for each observation i

When the first coin comes up heads, the probability of heads goes up

This variation in probability is *stochastic*, not deterministic

Not the result of observed covariates – we don't have any yet!



Flip coin; match with $p = 0.8$

$E(y) = 0.5$

$\text{var}(y) = 0.80$

Normal($\mu = 1$, $\sigma^2 = 2.5$)

$E(y) = \mu = 1$

$\text{var}(y) = \sigma^2 = 2.5$

Overdispersion of our event count increases as the dependence among trials rises

Upshot: if the chance of success in different trials is correlated, then we can't assume π_i is fixed conditional on covariates

π_i will also vary randomly across observations

This means we need a stochastic component for the probability of success π_i, in addition to our probability distribution over the count itself

The Beta distribution

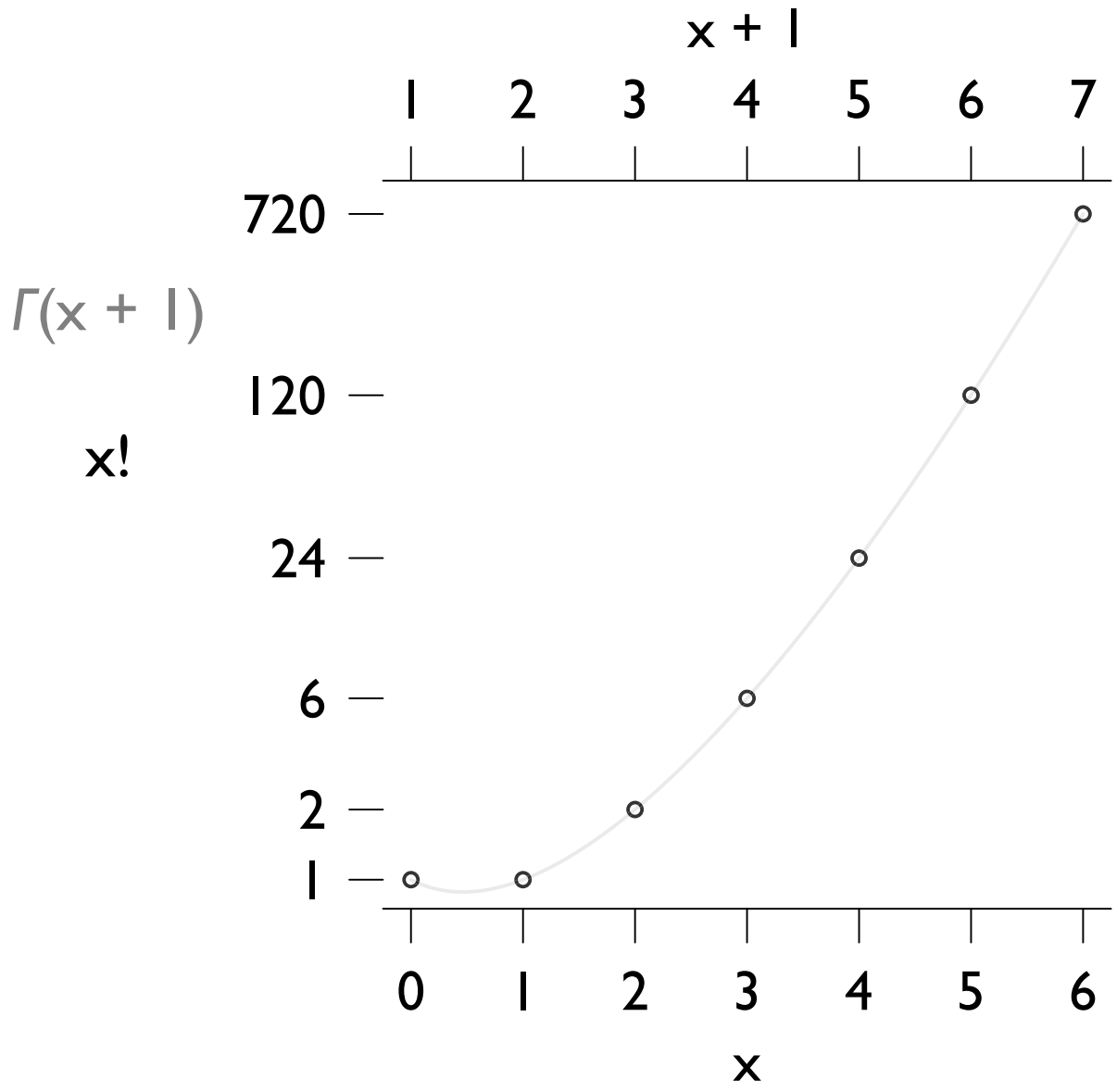
The most popular distribution for modeling outcomes that are, themselves, probabilities is the Beta distribution:

$$f_{\text{Beta}}(y_i|\alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} y_i^{\alpha_i - 1}(1 - y_i)^{\beta_i - 1}$$

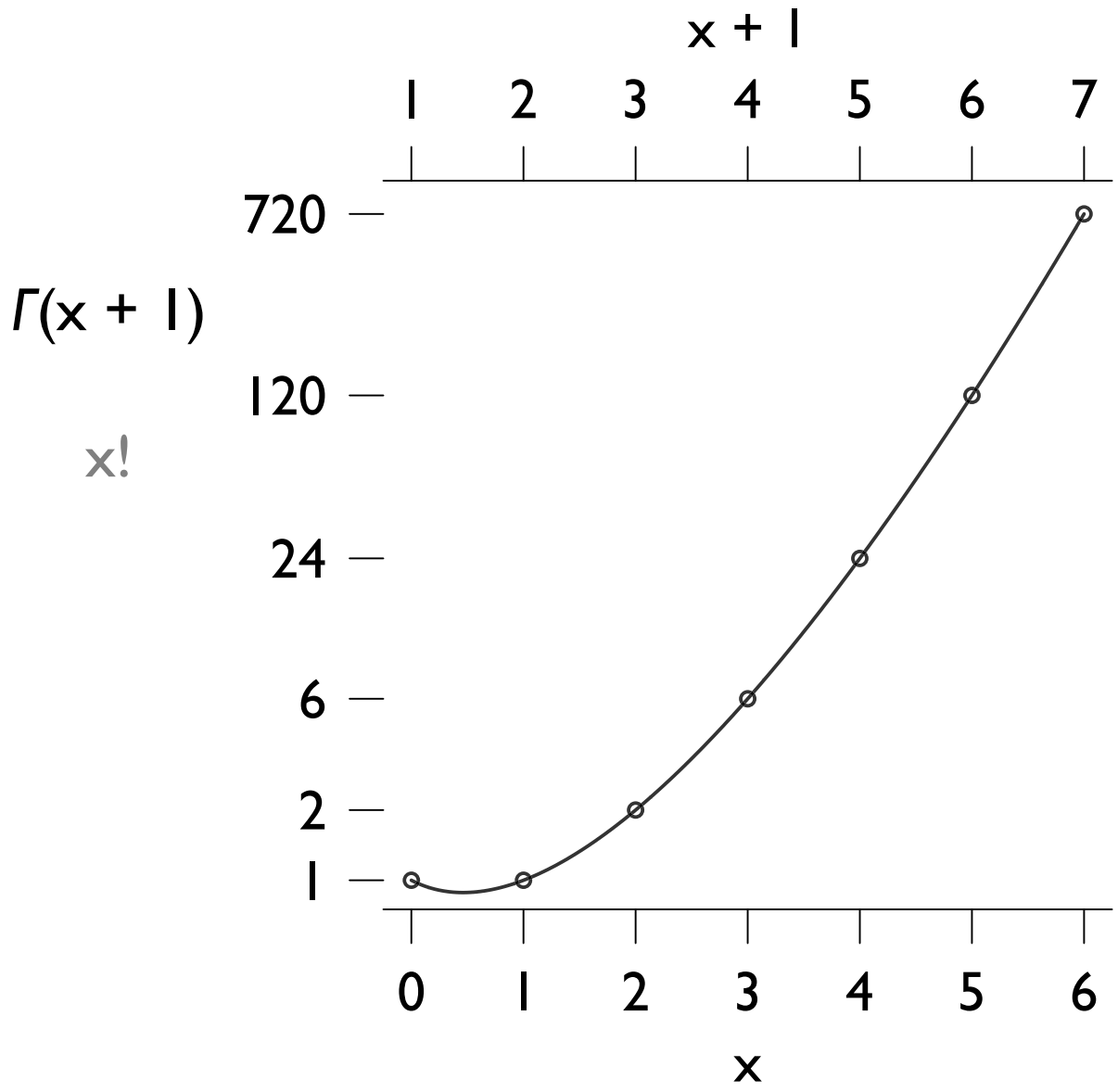
The Beta distribution has support over the interval [0,1] and is very flexible

But wait – what are those $\Gamma(\cdot)$ *functions?*

They are *Gamma functions*, or interpolated factorials. . .



Factorials like $x! = x \times (x-1) \times (x-2) \times \cdots \times 1$ are defined only for integers



The gamma function $\Gamma(x+1)$ interpolates "factorials" between the integers

It's computationally easier to work with log $\Gamma(x)$ in R: `lgamma()`

The Beta distribution

The most popular distribution for modeling outcomes that are, themselves, probabilities is the Beta distribution:

$$f_{\text{Beta}}(y_i|\alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} y_i^{\alpha_i - 1}(1 - y_i)^{\beta_i - 1}$$

The Beta distribution has support over the interval [0,1] and is very flexible

The Beta distribution

The most popular distribution for modeling outcomes that are, themselves, probabilities is the Beta distribution:

$$f_{\text{Beta}}(\pi_i|\alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \pi_i^{\alpha_i - 1}(1 - \pi_i)^{\beta_i - 1}$$

The Beta distribution has support over the interval [0,1] and is very flexible

One way to conceptualize the Beta comes from Bayesian statistics:
the Beta is the *conjugate prior distribution* for the Binomial distribution

Suppose a Bayesian wants to infer an unknown probability π
(now shown as the Beta random variable above)

Her prior beliefs about π could be represented as a historical record
of previously observed successes α and failures β

And her beliefs about the *ex ante* likely values of π
are given by the Beta distribution

The Beta distribution

The most popular distribution for modeling outcomes that are, themselves, probabilities is the Beta distribution:

$$f_{\text{Beta}}(\pi_i|\alpha_i,\beta_i) = \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\pi_i^{\alpha_i-1}(1-\pi_i)^{\beta_i-1}$$

The Beta distribution has support over the interval [0,1] and is very flexible

Moments of the Beta distribution

$$\mathbb{E}(\pi_i) = \frac{\alpha_i}{\alpha_i+\beta_i} \qquad \text{var}(\pi_i) = \frac{\alpha_i\beta_i}{(\alpha_i+\beta_i)^2(\alpha_i+\beta_i+1)}$$

The expected value of a Beta distributed variable is the success rate

The variance will become clearer as we go

The Beta distribution

$$f_{\text{Beta}}(\pi_i|\alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \pi_i^{\alpha_i - 1}(1 - \pi_i)^{\beta_i - 1}$$

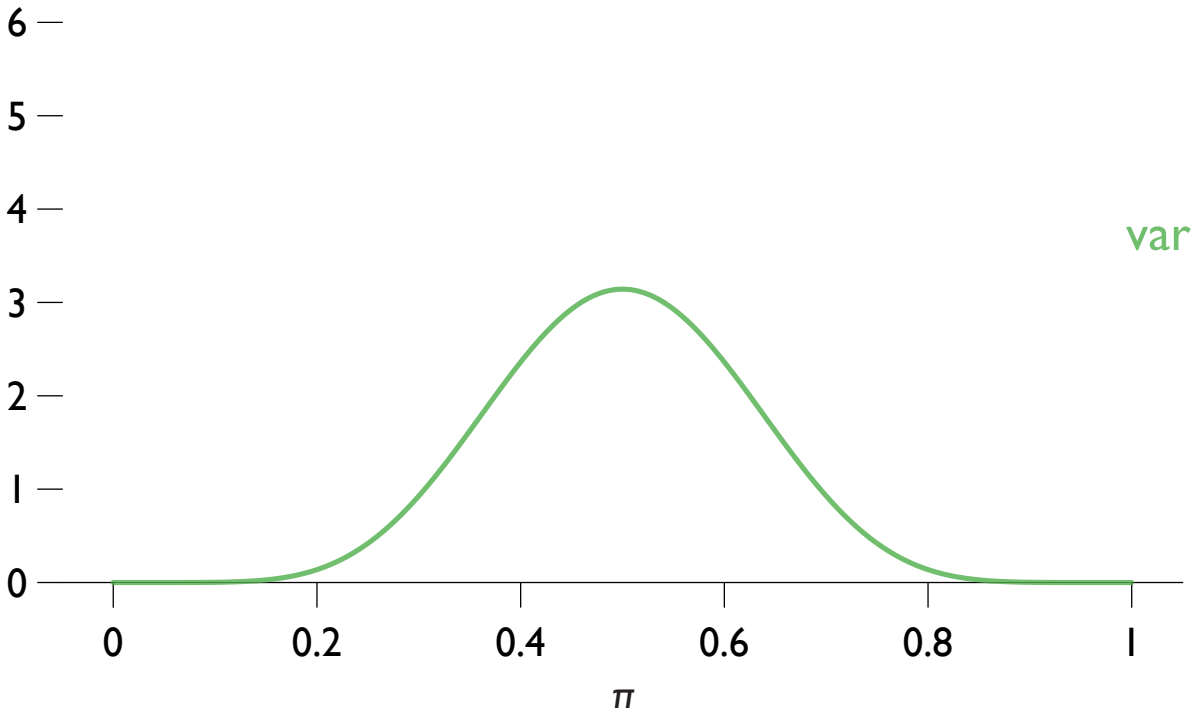
Alternative parameterization: mean and sample size

$$\mu_i = \frac{\alpha_i}{\alpha_i + \beta_i} \qquad \theta_i = \alpha_i + \beta_i$$

Moments of the Beta distribution given μ and θ

$$\mathbb{E}(\pi_i) = \mu_i \qquad \text{var}(\pi_i) = \frac{\mu_i(1 - \mu_i)}{\theta_i - 1}$$

The Beta variance resembles the Binomial variance, but shrinks as $\theta \to \infty$



$$\text{Beta}(a = 8,\ \beta = 8)$$

$$\text{E}(\pi) = a/(a+\beta) = 0.500$$

$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.015$$

This is a Beta distribution with $\alpha = 8$ and $\beta = 8$

One way to understand this distribution is to imagine
we've drawn $8 + 8 = 16$ times from a binary process with unknown probability π

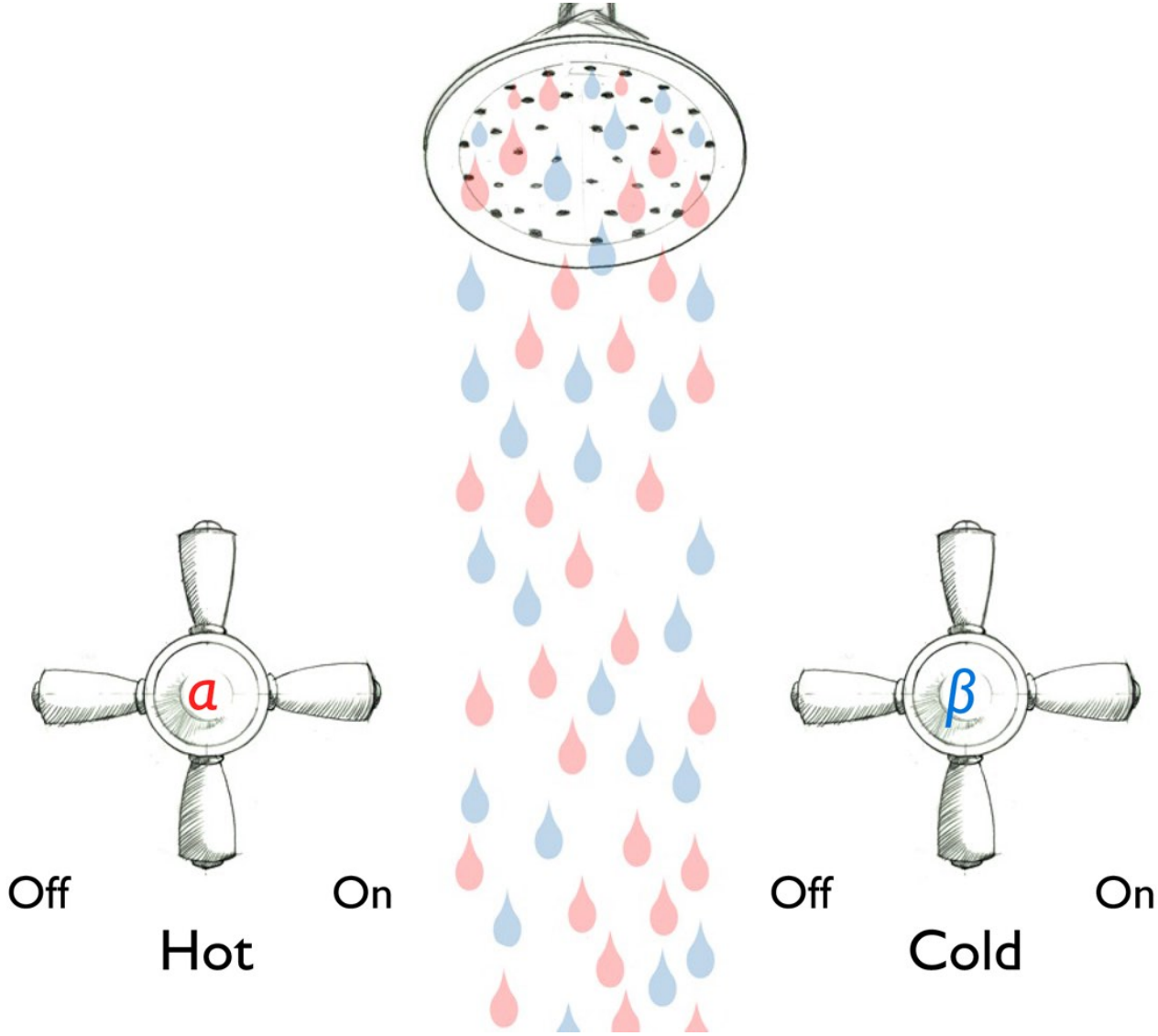
We've observed 8 "success" draws and 8 "failure" draws so far

The Beta shows the probabilities of various π's that might have produced our draws

Let's build a more tangible thought experiment. . .

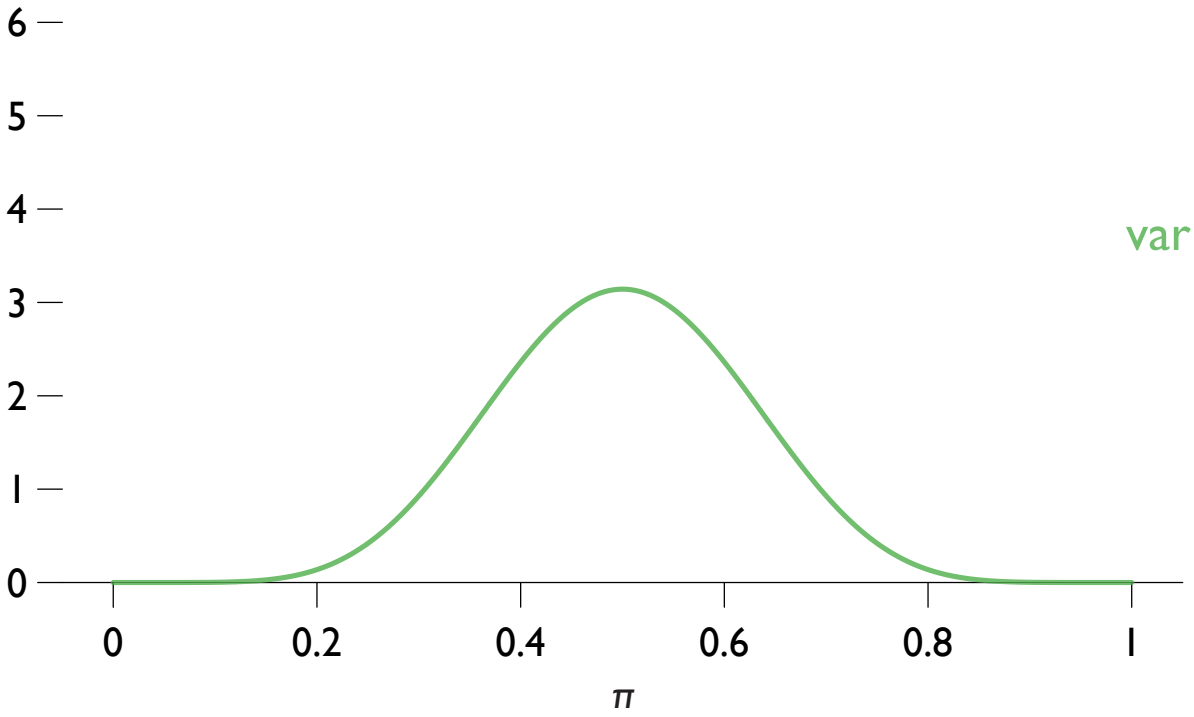
Suppose we have an unreliable shower that doesn't mix hot and cold water well

We can control the amount & average temperature of water,
but the water comes out as discrete hot and cold drops



We have two old-fashioned shower taps

The left controls the amount of hot water – analogous to α
the right controls the amount of cold water – analogous to β



$\text{Beta}(a = 8,\ \beta = 8)$

$\text{E}(\pi) = a/(a+\beta) = 0.500$

$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.015$

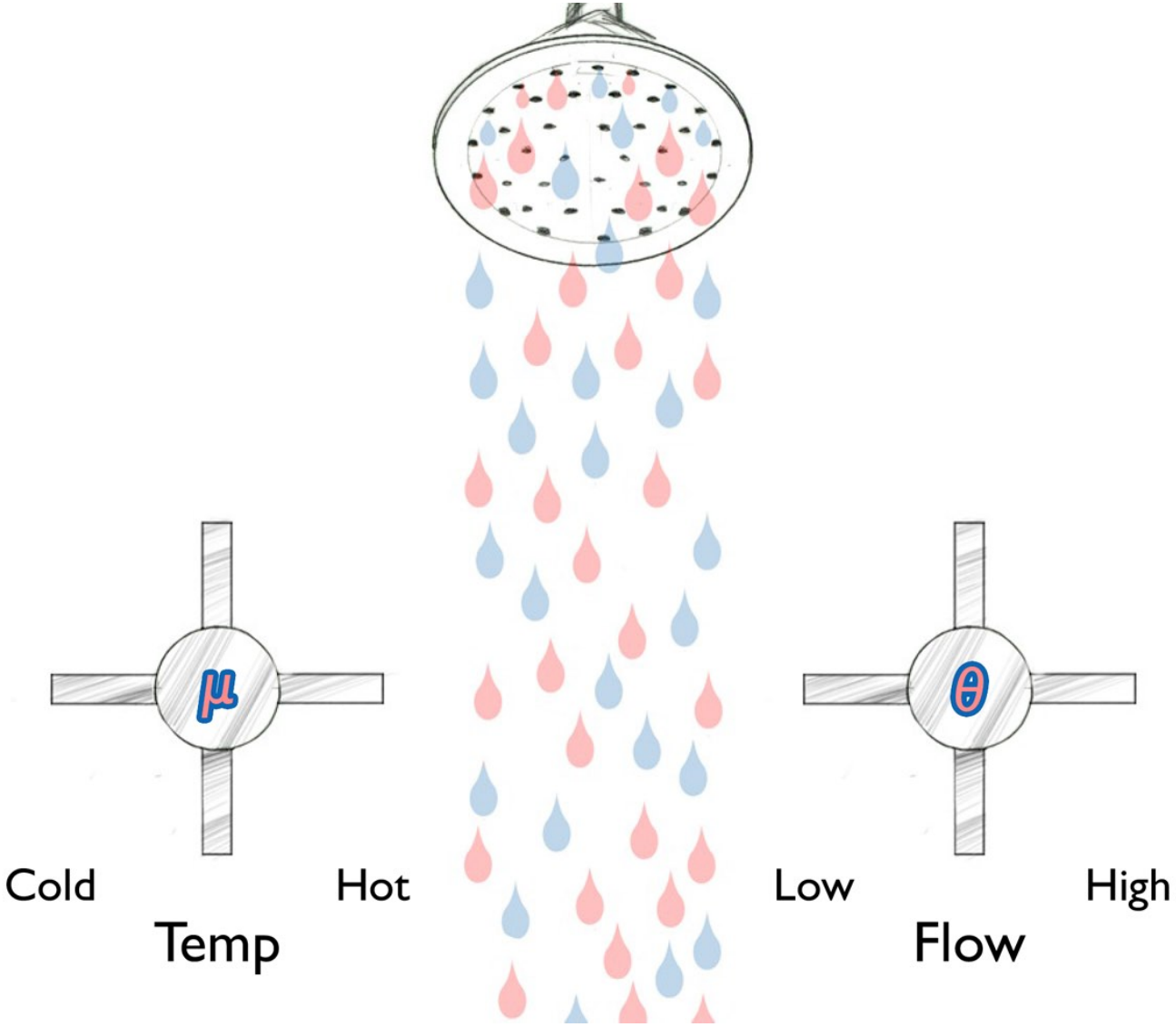
In our analogy, α represents the number of hot water drops pouring down

β is the number of cold drops

$\pi = \dfrac{\alpha}{\alpha + \beta}$ is the average water temperature

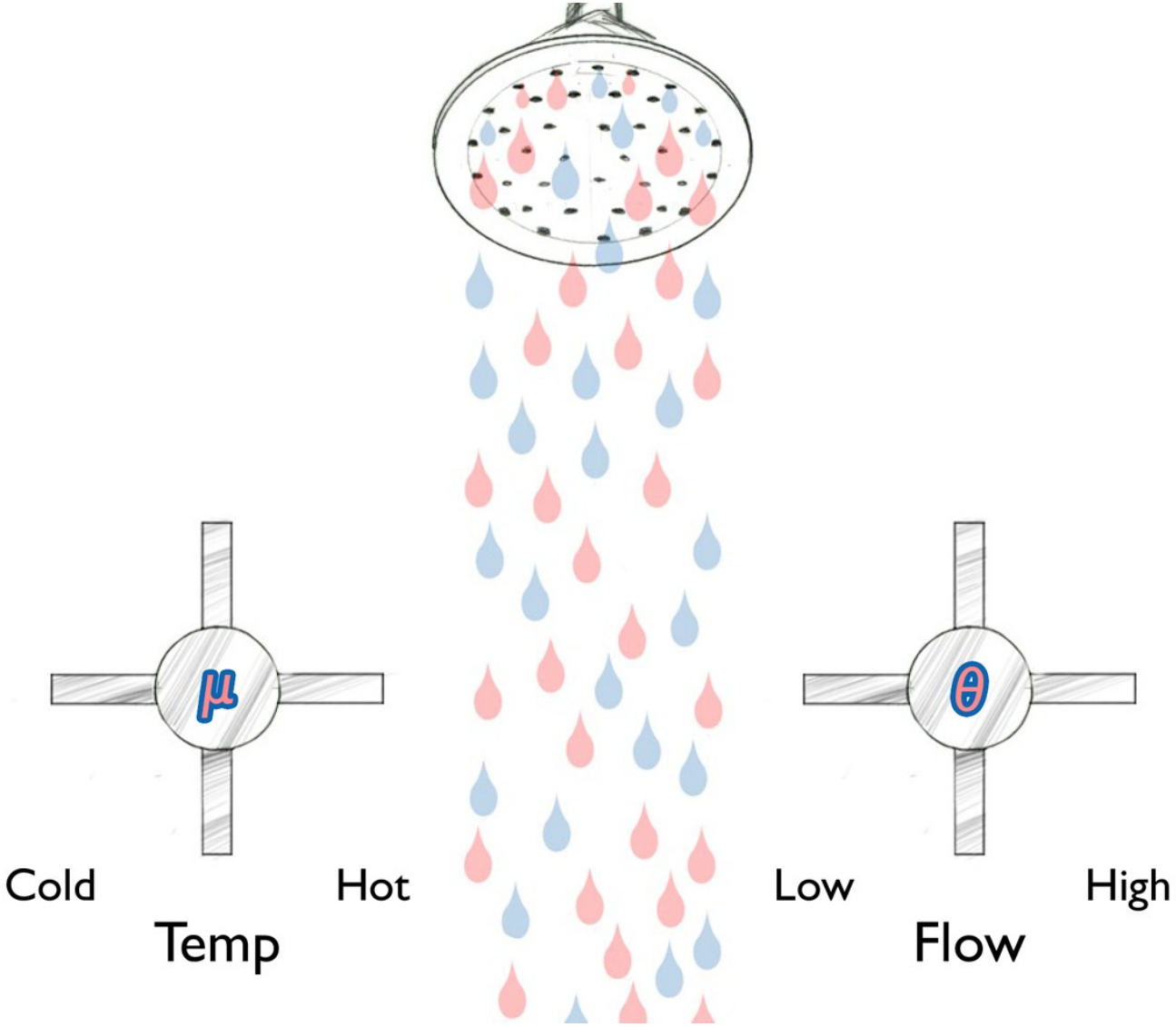
π is also the probability the next drop is hot

So far so good, but how this relates to the variance is still unclear



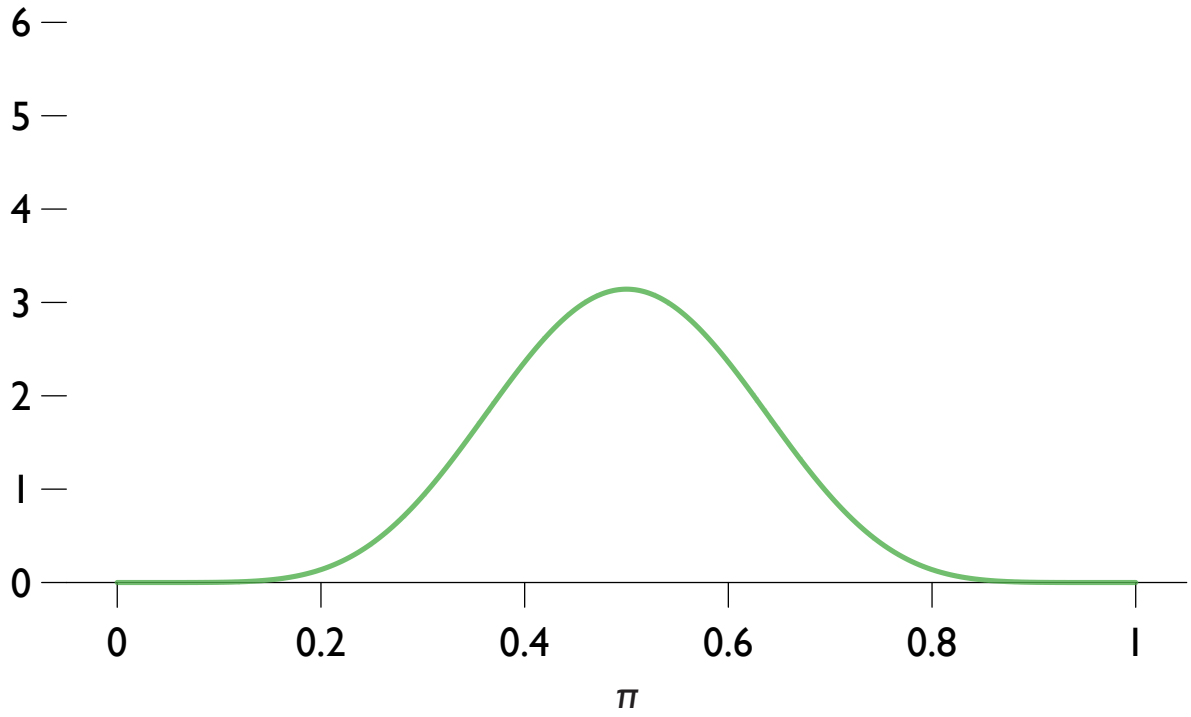
What if we replaced our old-fashioned taps with newer ones?

The left controls the average temperature
The right controls the total flow or volume of drops



The new taps can dial up any mix of cold and water the old ones could

But they do so using a different set of *parameters*



$\text{Beta}(\mu = 0.5, \theta = 16)$

$E(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.015$

Reparameterizing a distribution is like changing the way the shower taps work

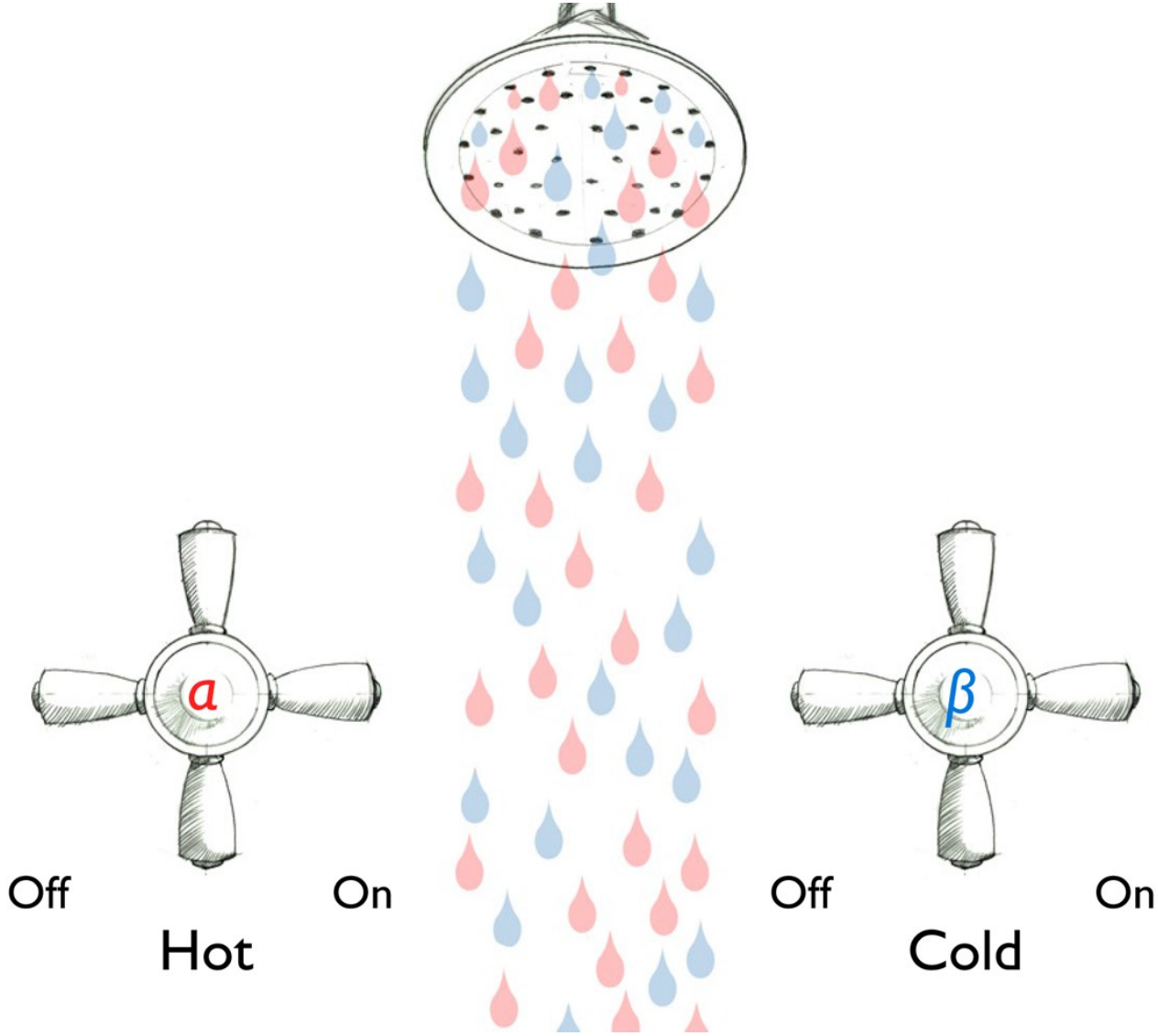
Define two new parameters

$\mu = \frac{\alpha}{\alpha + \beta}$, the average rate of successes (or the average temperature)

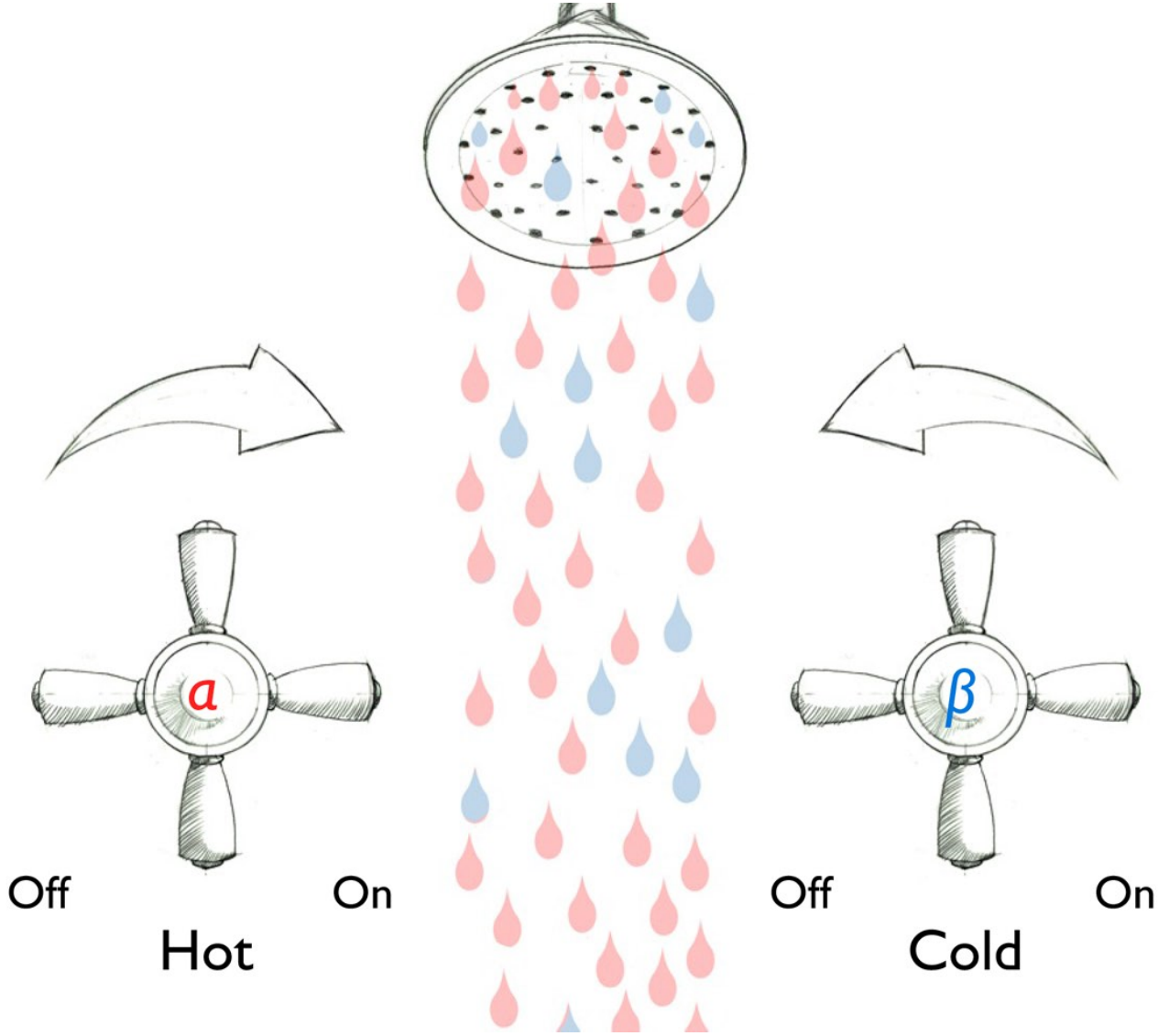
$\theta = \alpha + \beta$, the total sample size (or the volume of drops)

We can reframe the Beta($\alpha = 8$, $\beta = 8$) distribution as Beta($\mu = 0.5$, $\theta = 16$)

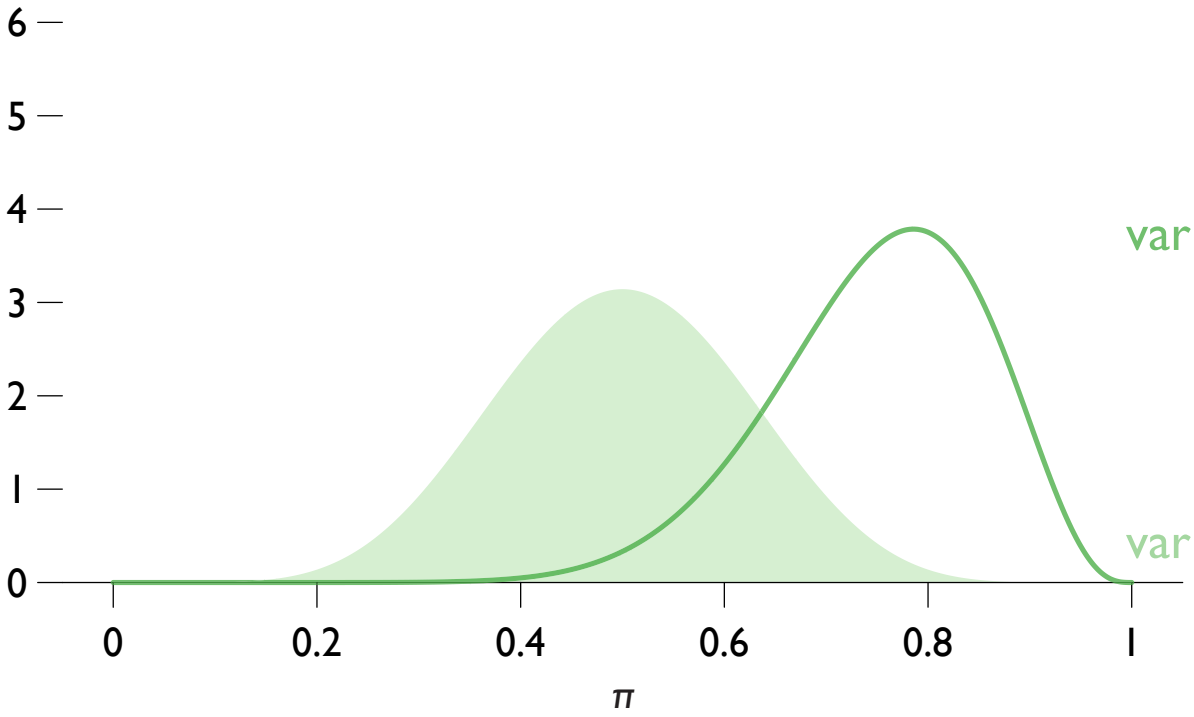
The same distribution in an easier-to-manipulate parameterization



Using the old-fashioned taps,
how would we raise the water temperature without changing the volume?



Add hot water and remove cold water!



$$\text{Beta}(a = 12,\ \beta = 4)$$

$$E(\pi) = a/(a+\beta) = 0.750$$

$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.011$$

$$\text{Beta}(a = 8,\ \beta = 8)$$

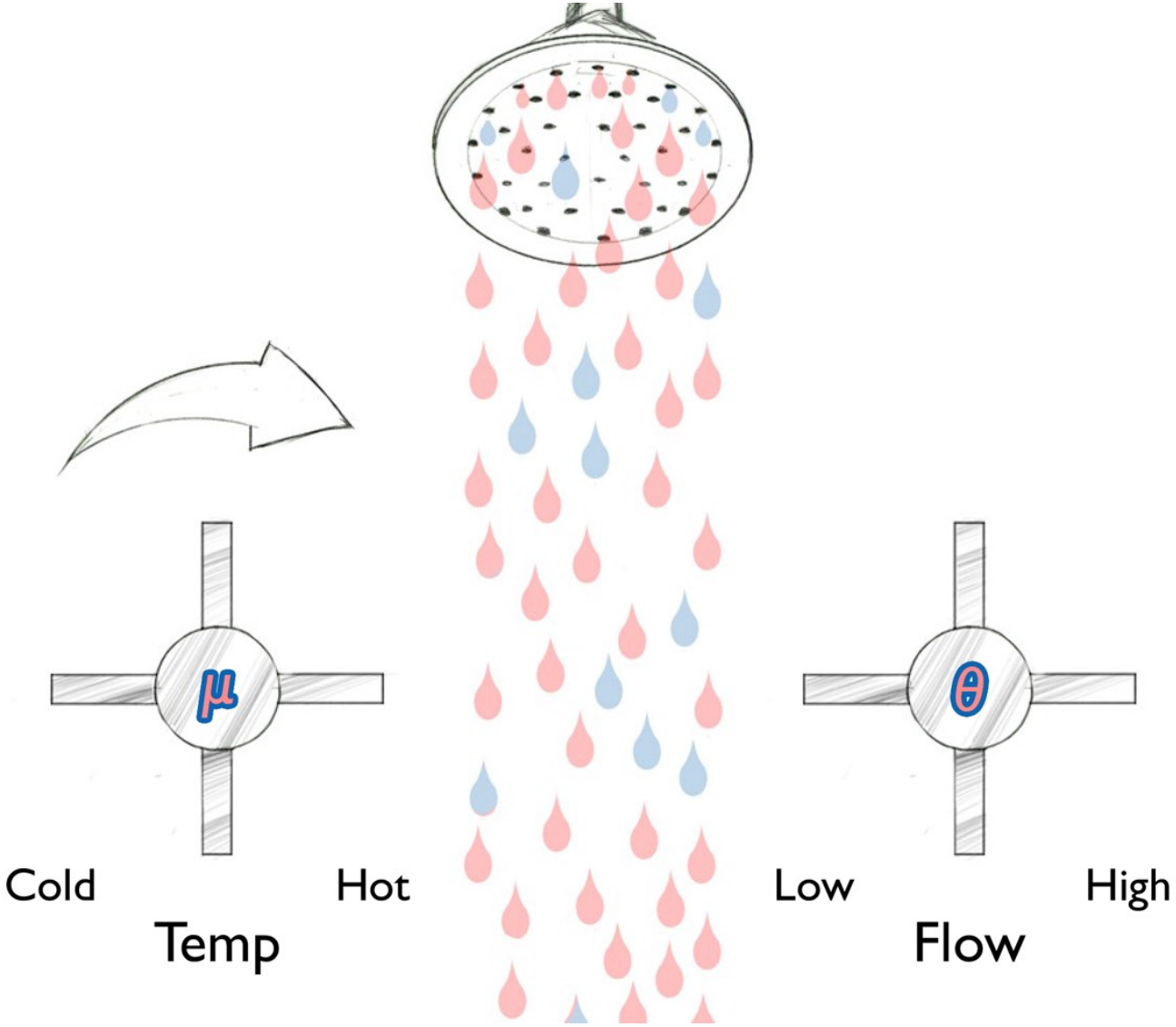
$$E(\pi) = a/(a+\beta) = 0.500$$

$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.015$$

The distribution is now centered over an expected π of 0.75

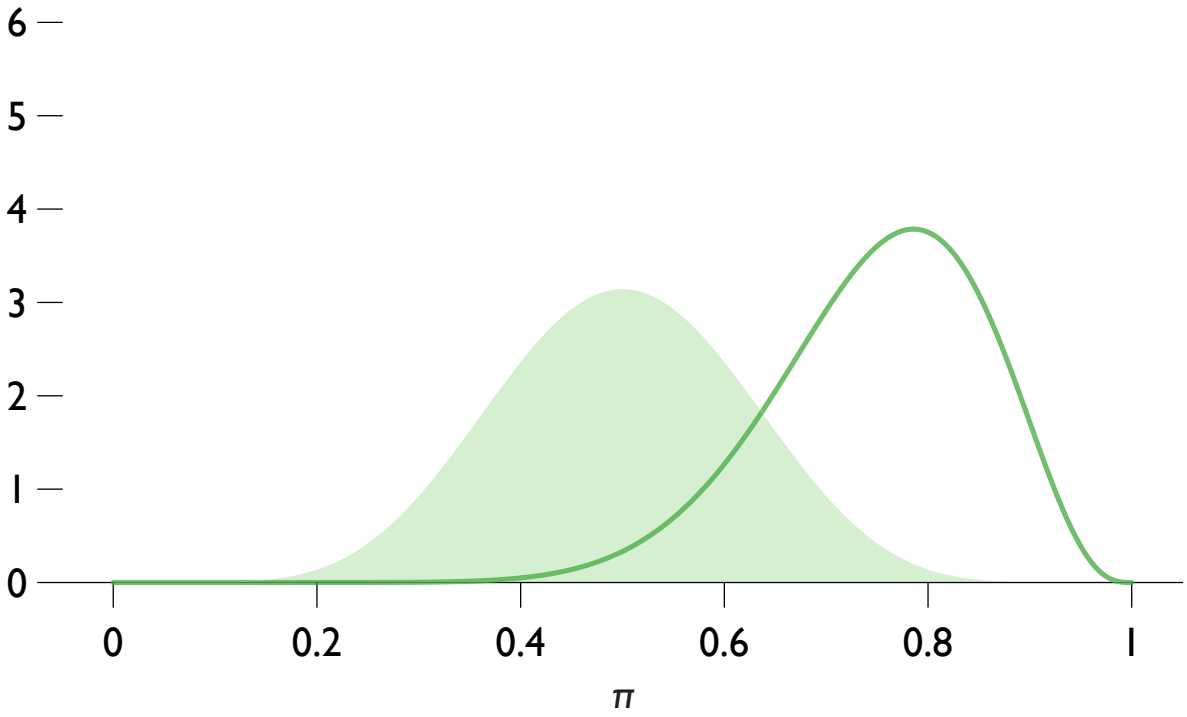
The variance is about the same – slightly lower, as it's a function of the mean, too

As in the Binomial, Beta variances are biggest when π is near 0.5



Using the new taps – which correspond to μ and θ, respectively – it's even easier to raise temperature without changing volume

Just turn the temperature tap up



$\text{Beta}(\mu = 0.75, \theta = 16)$

$\text{E}(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.011$

$\text{Beta}(\mu = 0.5, \theta = 16)$

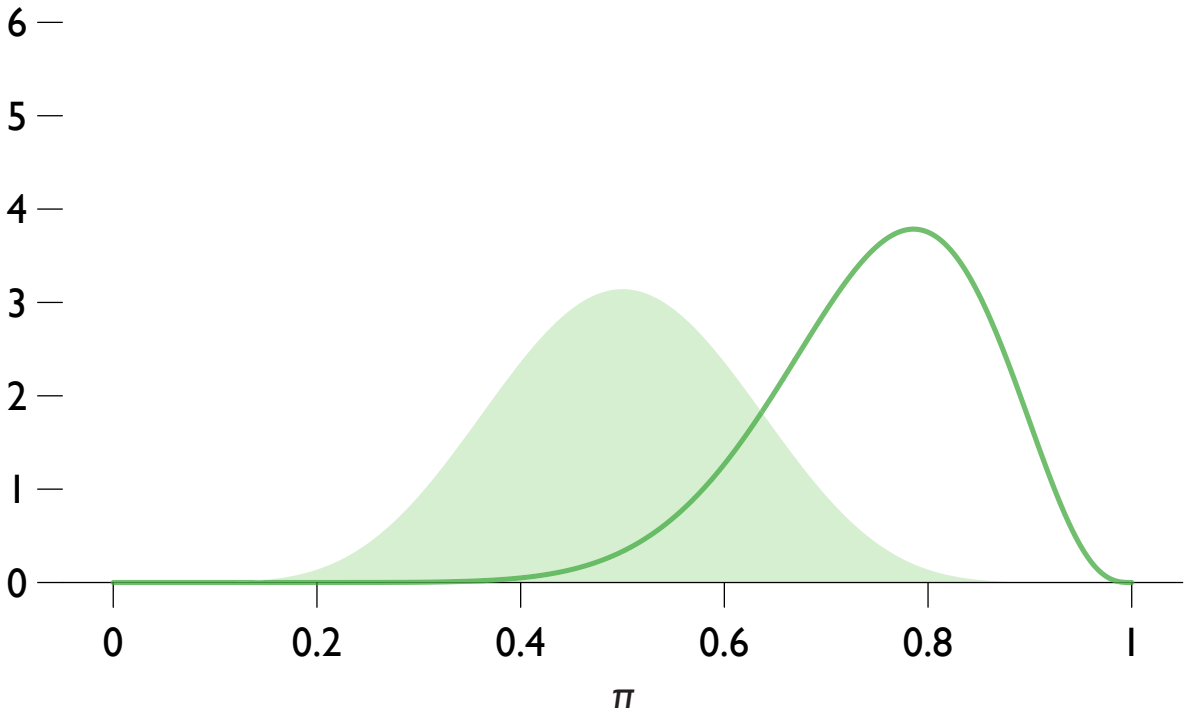
$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.015$

In terms of the Beta distribution, we have raised μ to 0.75, but kept θ at 16

This shifts the mean while keeping the variance mostly the same

What does the variance of the Beta distribution represent?



$\text{Beta}(\mu = 0.75, \theta = 16)$

$E(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.011$

$\text{Beta}(\mu = 0.5, \theta = 16)$

$E(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.015$

What does the variance of the Beta distribution represent?

We are trying to infer an unknown probability (or rate) of hot drops from a sample of hot and cold drops

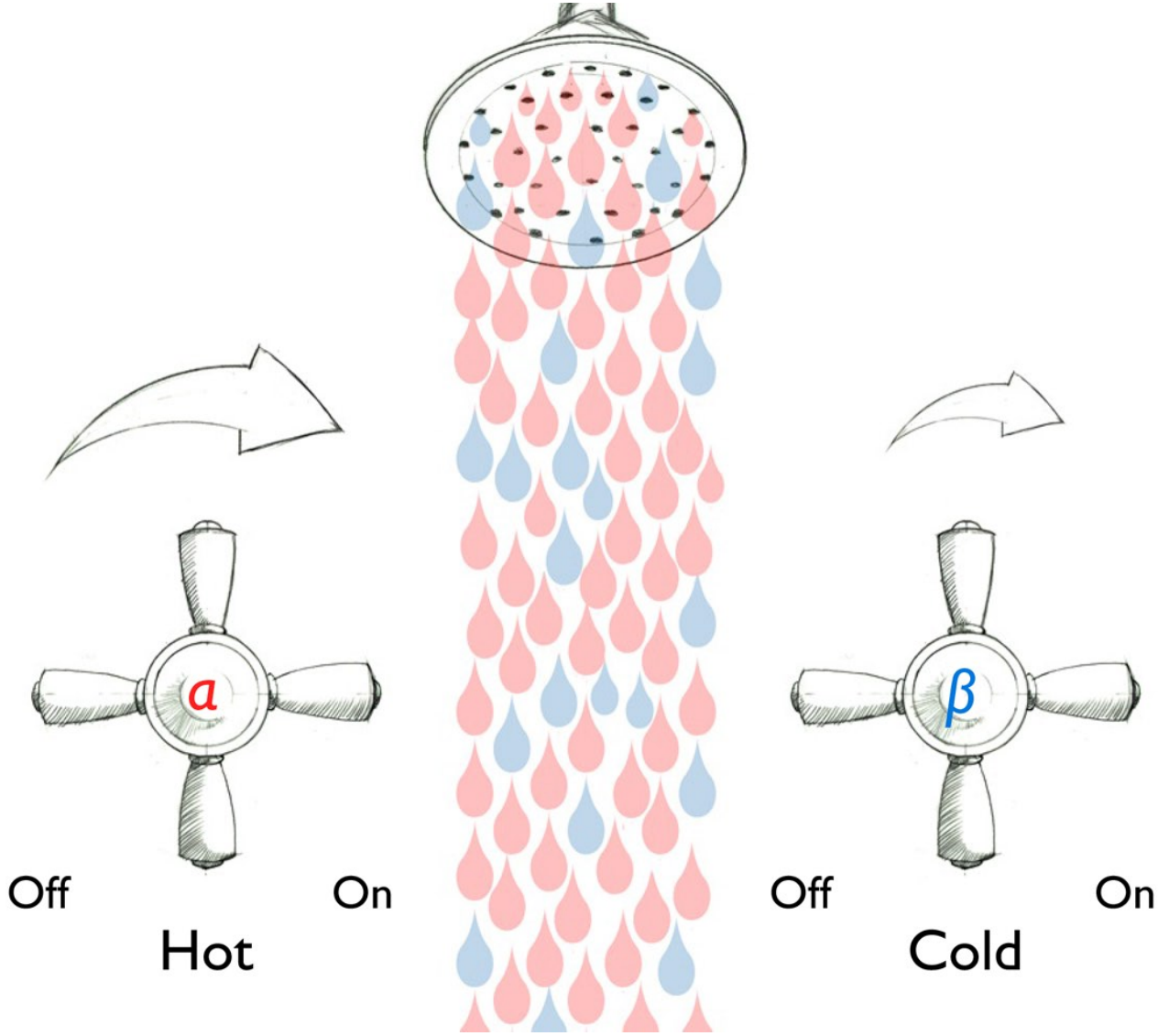
The percentage of hot drops in our sample is our best guess of that rate

But the smaller the sample, the less certain we are that μ is close to the true π

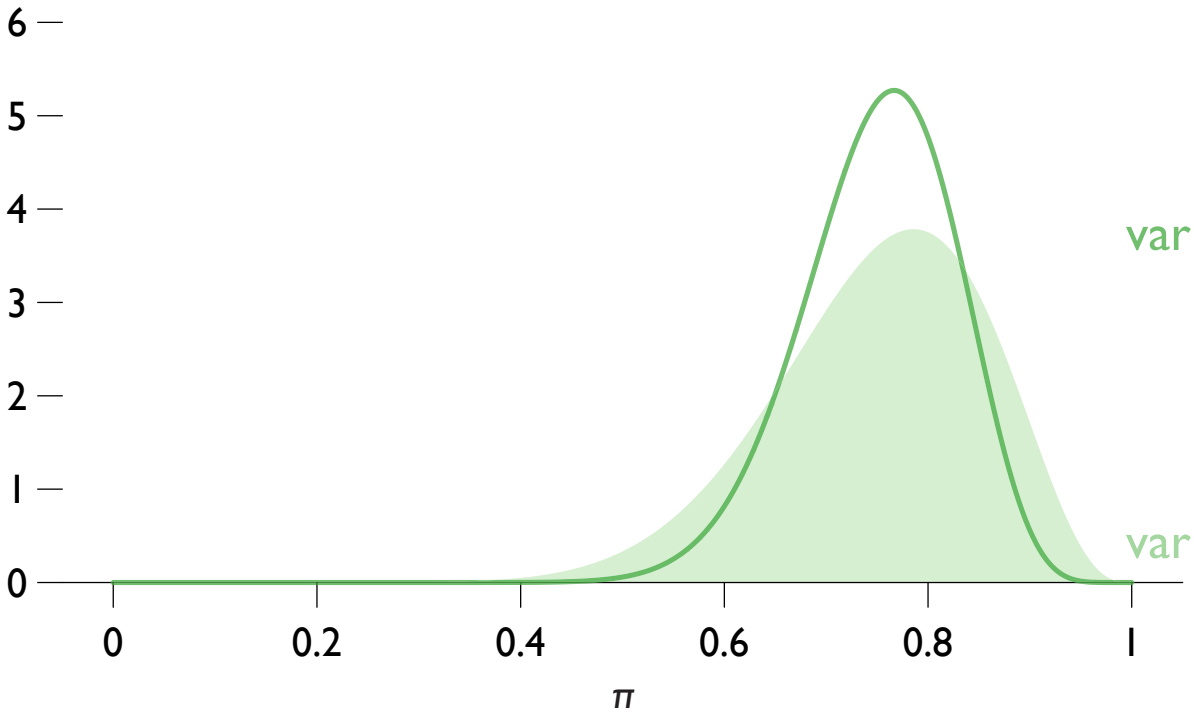
Hence the variance of our (guess) of π is larger the smaller θ is

We now have a medium flow of mostly hot water

How do we make it a torrent without changing the average temperature?



Using the old-fashioned taps,
add a lot of hot water, but also a little cold water to keep the balance

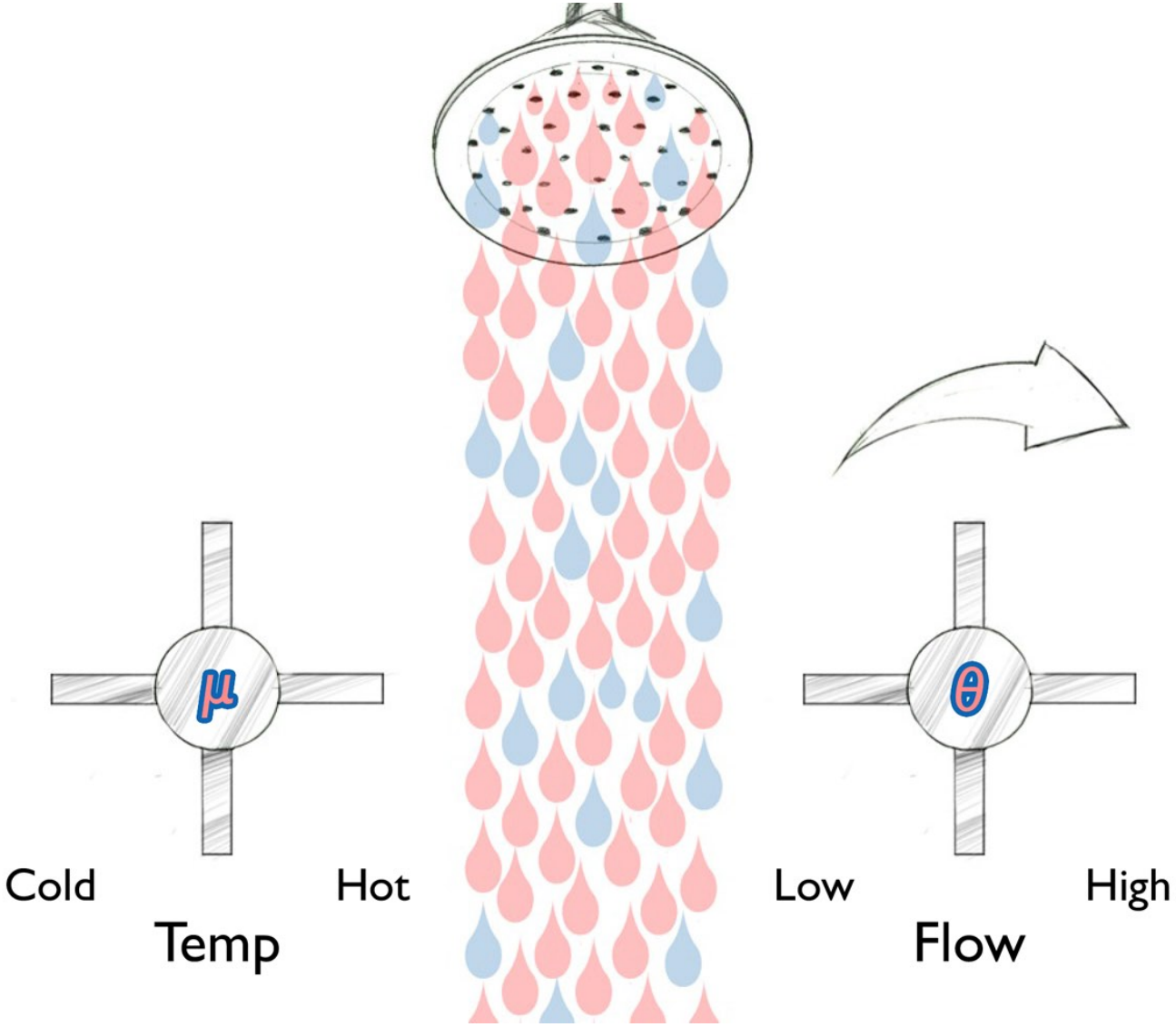


$$\text{Beta}(a = 24,\ \beta = 8)$$
$$E(\pi) = a/(a+\beta) = 0.750$$
$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.006$$

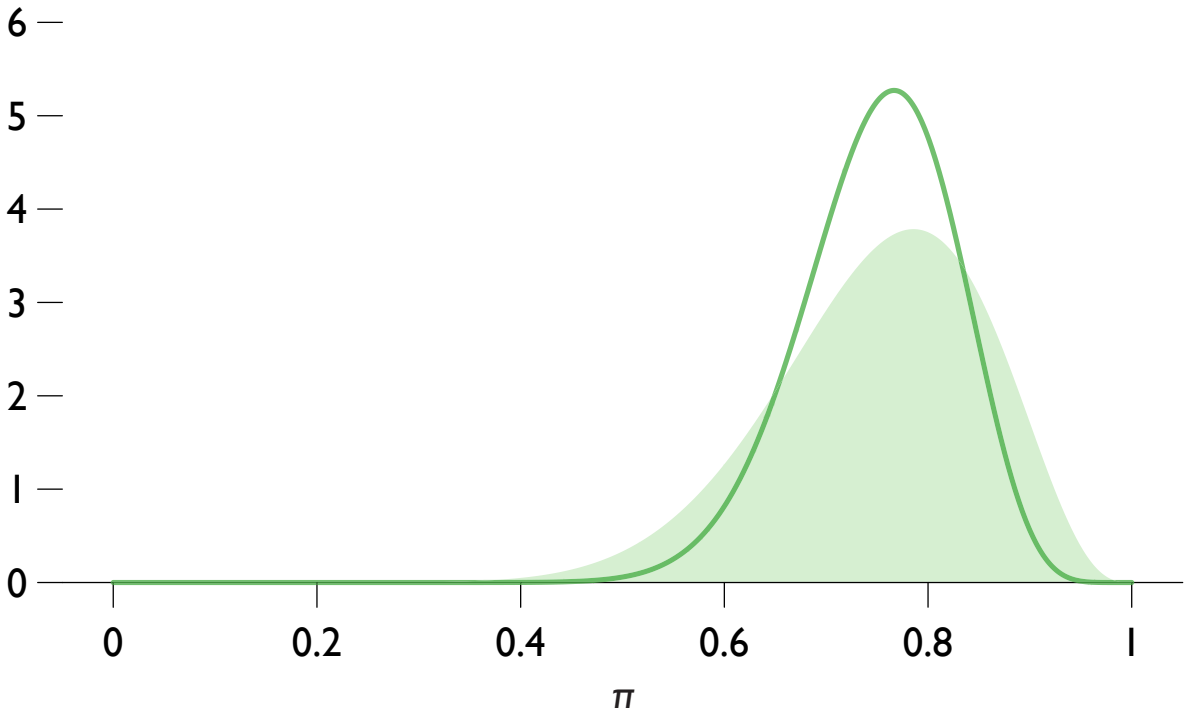
$$\text{Beta}(a = 12,\ \beta = 4)$$
$$E(\pi) = a/(a+\beta) = 0.750$$
$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.011$$

In terms of the α, β parameterization of the Beta,
we double each parameter, which keeps their ratio constant

We have many more samples now (32 versus 16),
so we can be more certain of the π we infer from the sample



To raise the flow with constant temperature using the new taps, we need adjust only the flow parameter θ



$\text{Beta}(\mu = 0.75, \theta = 32)$

$\text{E}(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.006$

$\text{Beta}(\mu = 0.75, \theta = 16)$

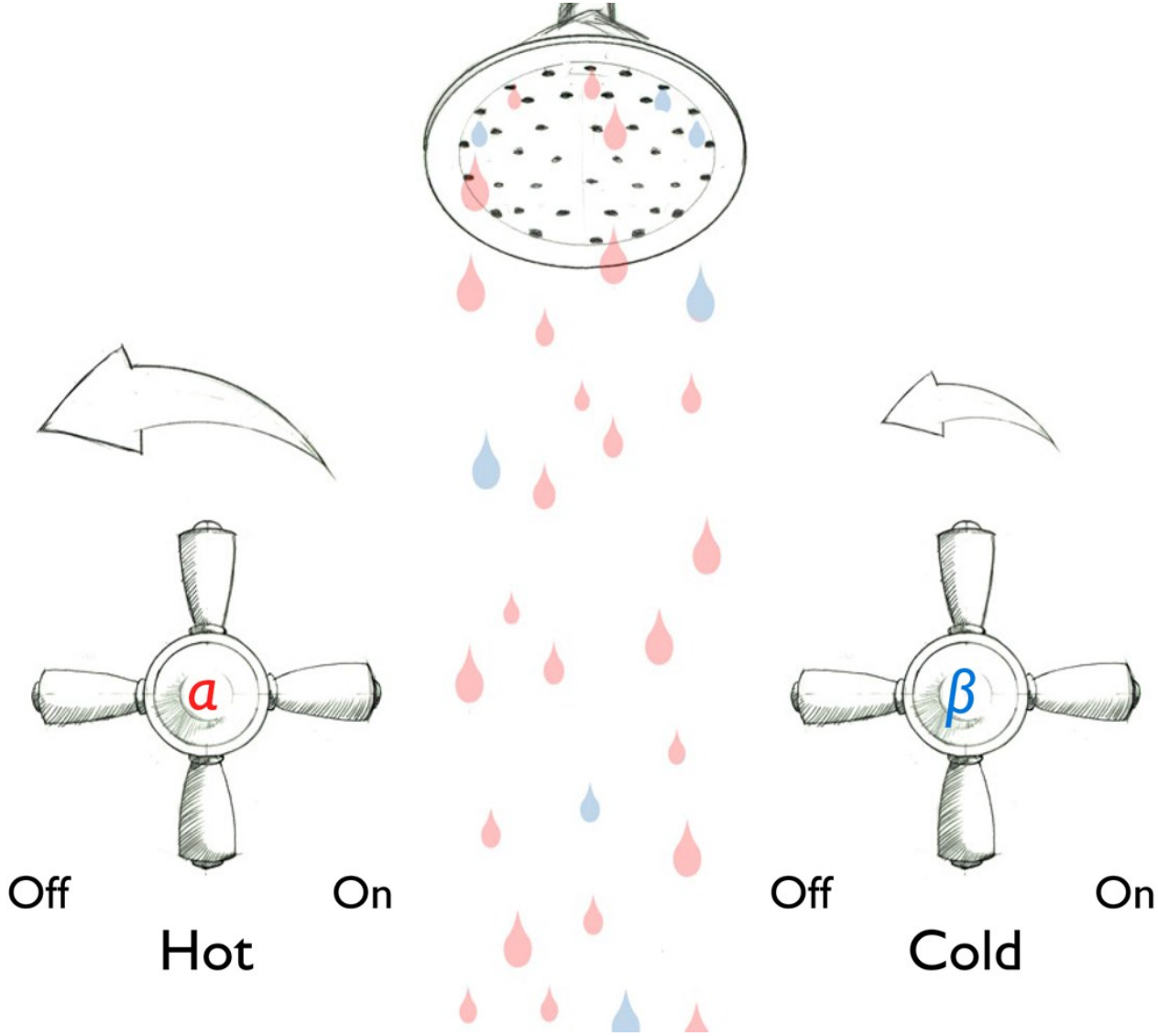
$\text{E}(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.011$

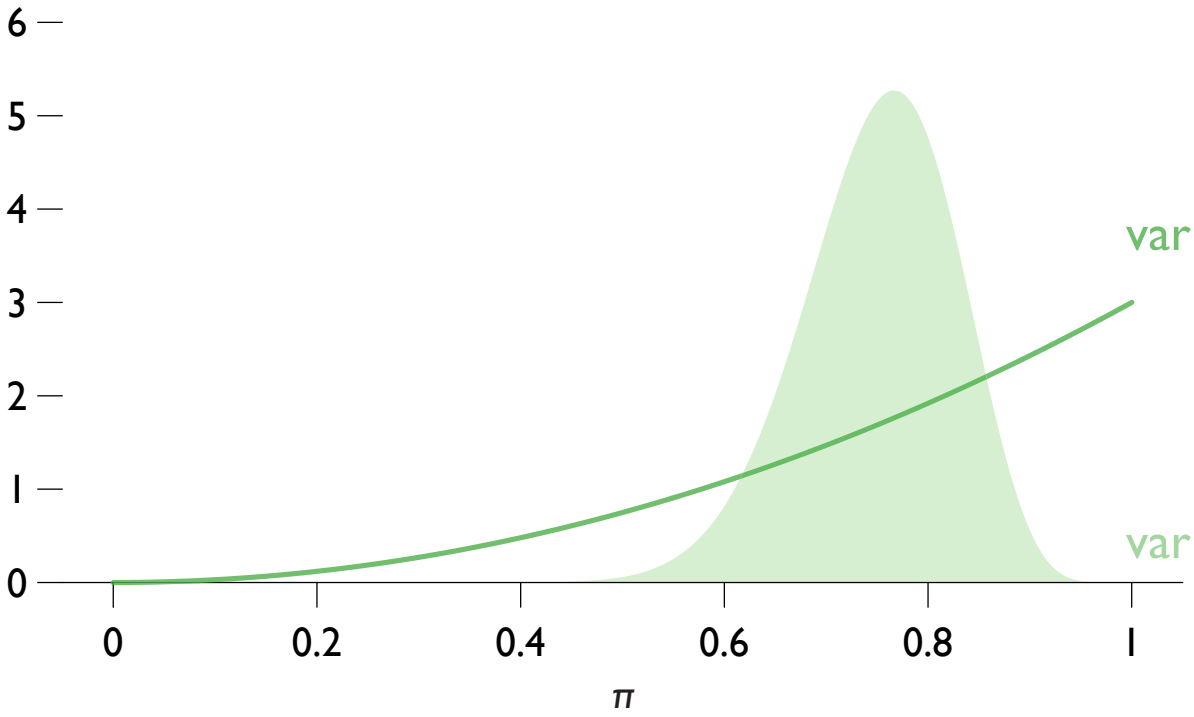
That is, we keep μ fixed at 0.75 and raise θ from 16 to 32

The Beta distribution is now tighter around the mean,
but still a bit asymmetric near the 1.0 bound

Keeping the temperature hot,
how do we mostly shut off the water?



With the old taps, cut the hot water a lot and the cold water a little



$\text{Beta}(a = 3,\ \beta = 1)$

$\text{E}(\pi) = a/(a+\beta) = 0.750$

$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.037$

$\text{Beta}(a = 24,\ \beta = 8)$

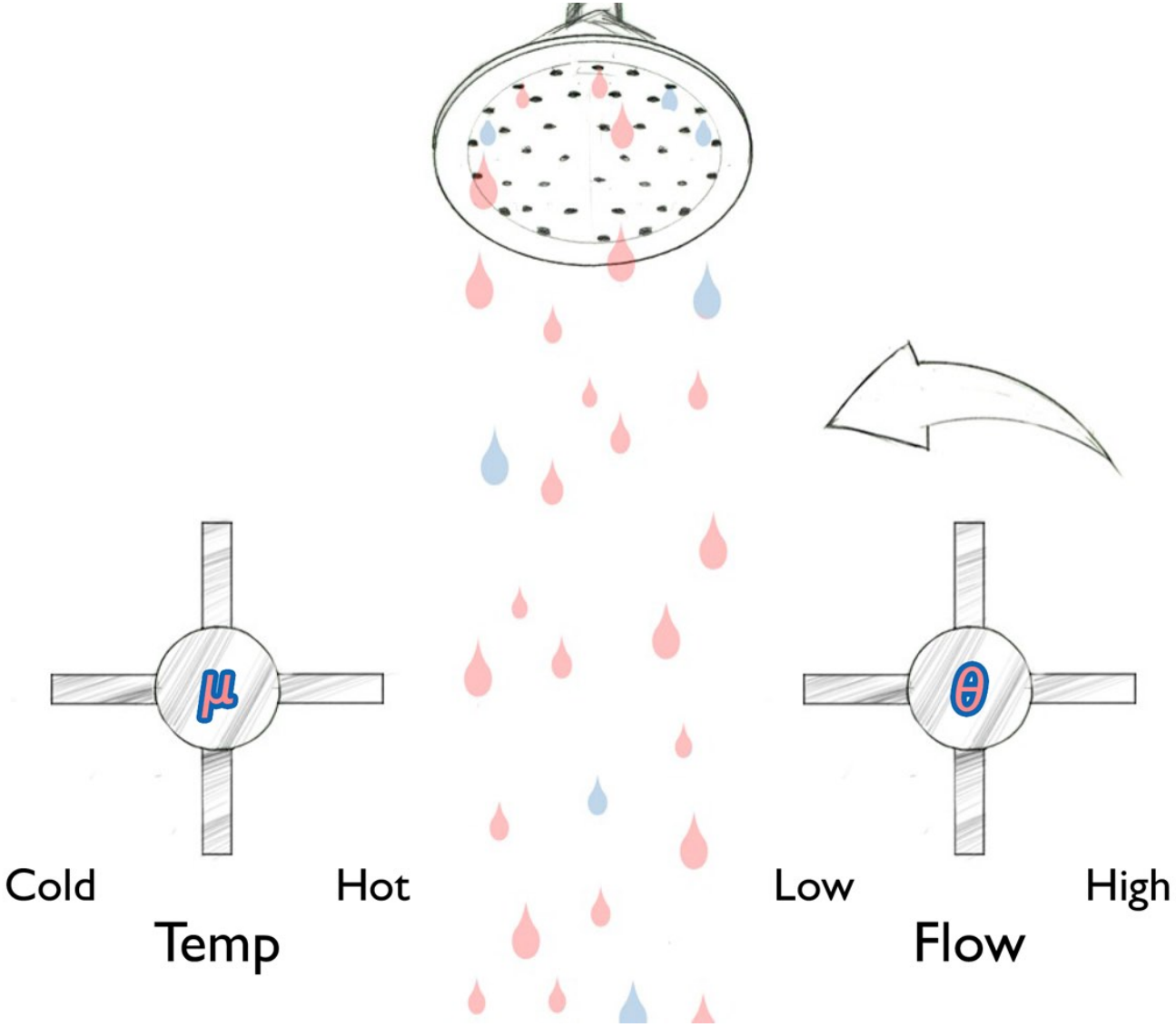
$\text{E}(\pi) = a/(a+\beta) = 0.750$

$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.006$

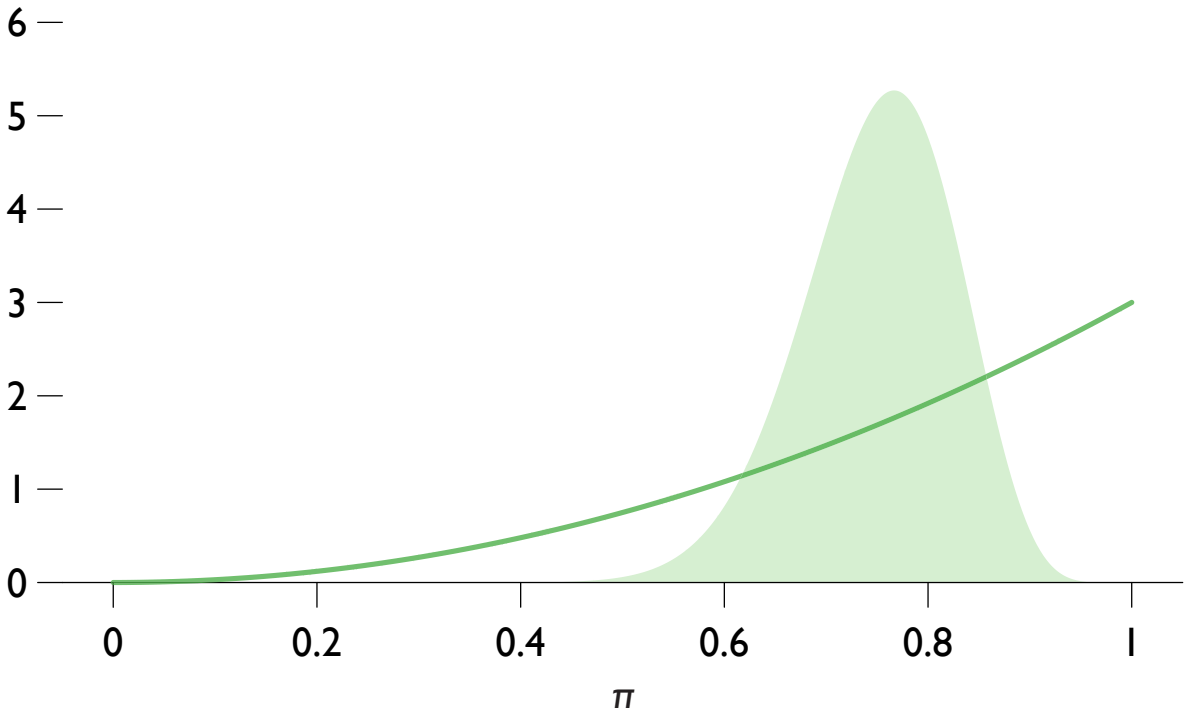
The Beta distribution remains centered on $\pi = 0.75$,
but the variance is now so large this is hard to see visually

With just a few drops to infer average temperature from,
we can't discount even very low average temperatures so easily

We might get a small run of cold drops by chance;
or we might doubt that a small run of hot drops is representative



As usual, adjusting flow downwards at a fixed temperature is more intuitive with the new style of taps, or the μ and θ parameters



Beta($\mu = 0.75$, $\theta = 4$)

$E(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.037$

Beta($\mu = 0.75$, $\theta = 32$)

$E(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.006$

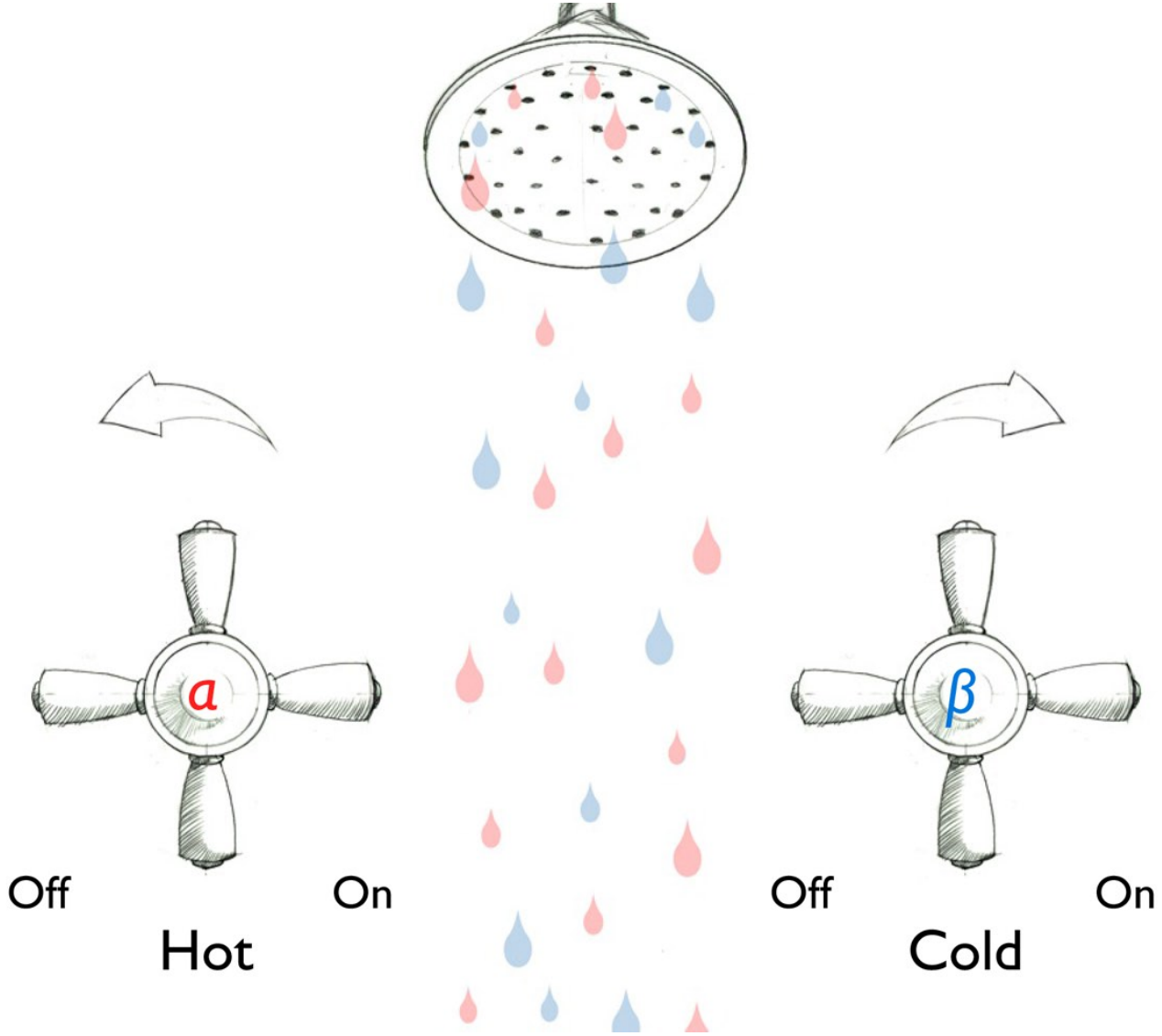
Adjusting θ downwards increases variance while keeping the mean fixed

The Beta variance is mostly determined by sample size

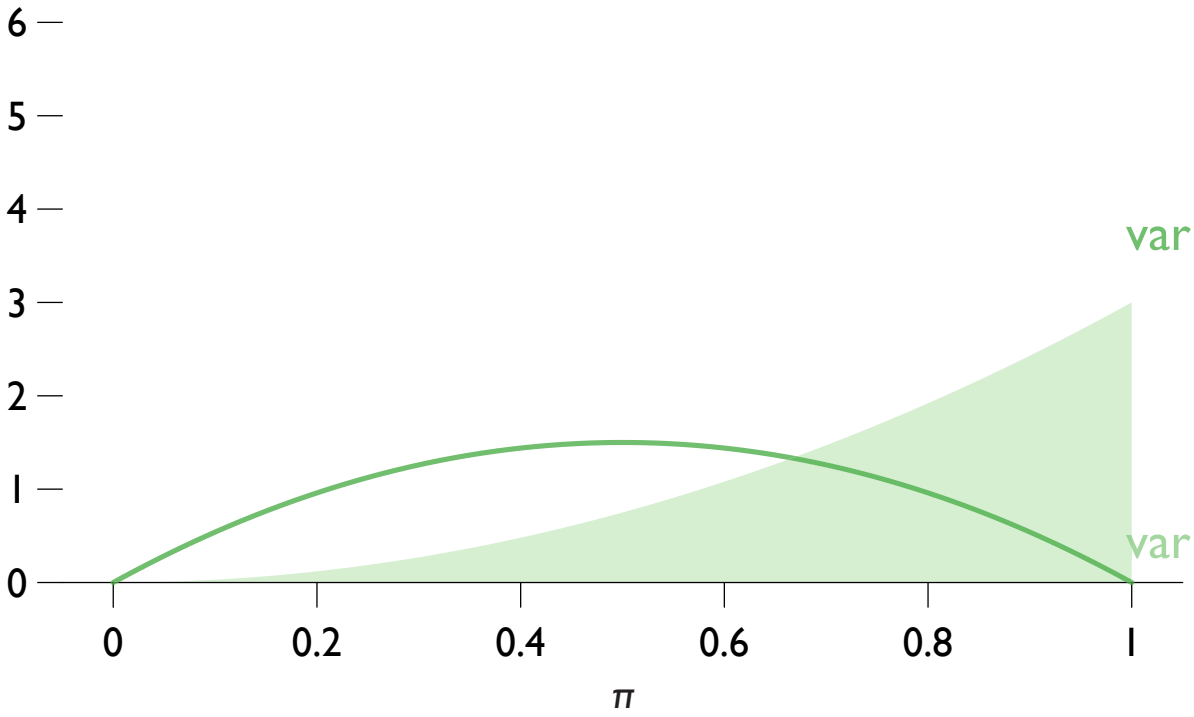
The Beta($\mu = 0.75$, $\theta = 4$) distribution peaks at $\pi = 1.0$ – *why?*

Because the Beta is limited to $\pi \in [0, 1]$, so it can't spread further right

What if we wanted to lower the mean temperature while keeping the flow low?



With the old-fashioned taps. . .



$$\text{Beta}(a = 2,\ \beta = 2)$$

$$E(\pi) = a/(a+\beta) = 0.500$$

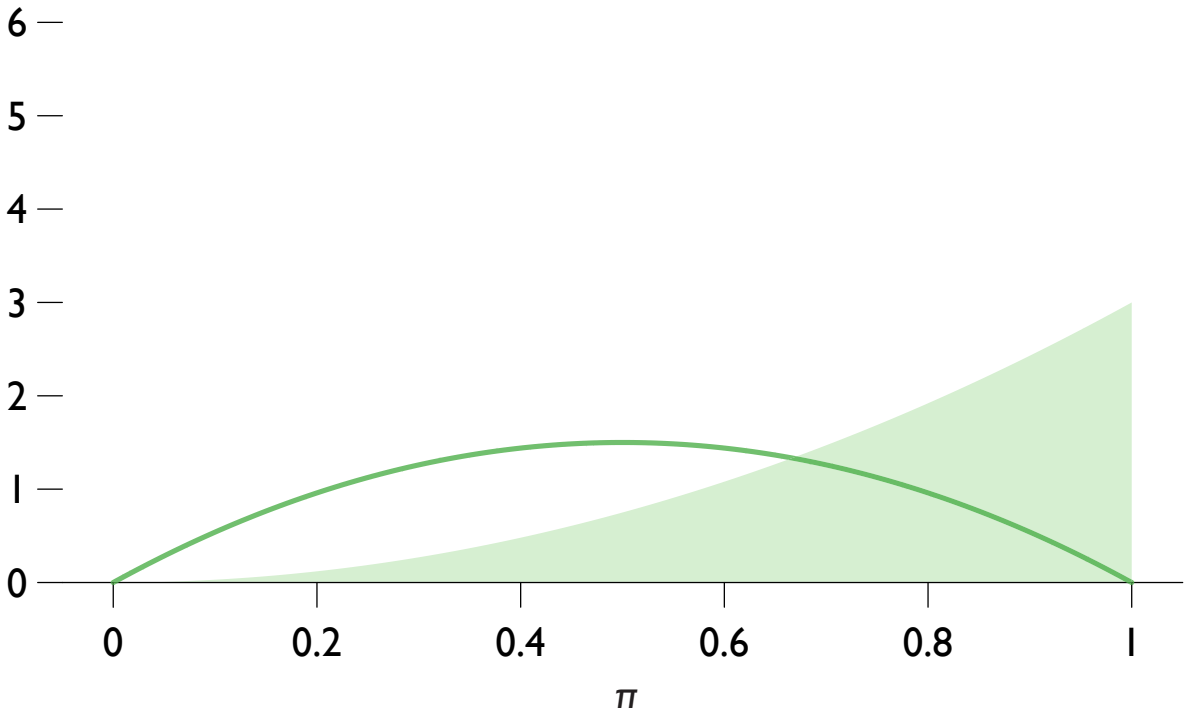
$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.050$$

$$\text{Beta}(a = 3,\ \beta = 1)$$

$$E(\pi) = a/(a+\beta) = 0.750$$

$$\text{var}(\pi) = a\beta/\left((a+\beta)^2(a+\beta+1)\right) = 0.037$$

How would this look under the alternative parameterization?



$\text{Beta}(\mu = 0.5, \theta = 4)$

$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.050$

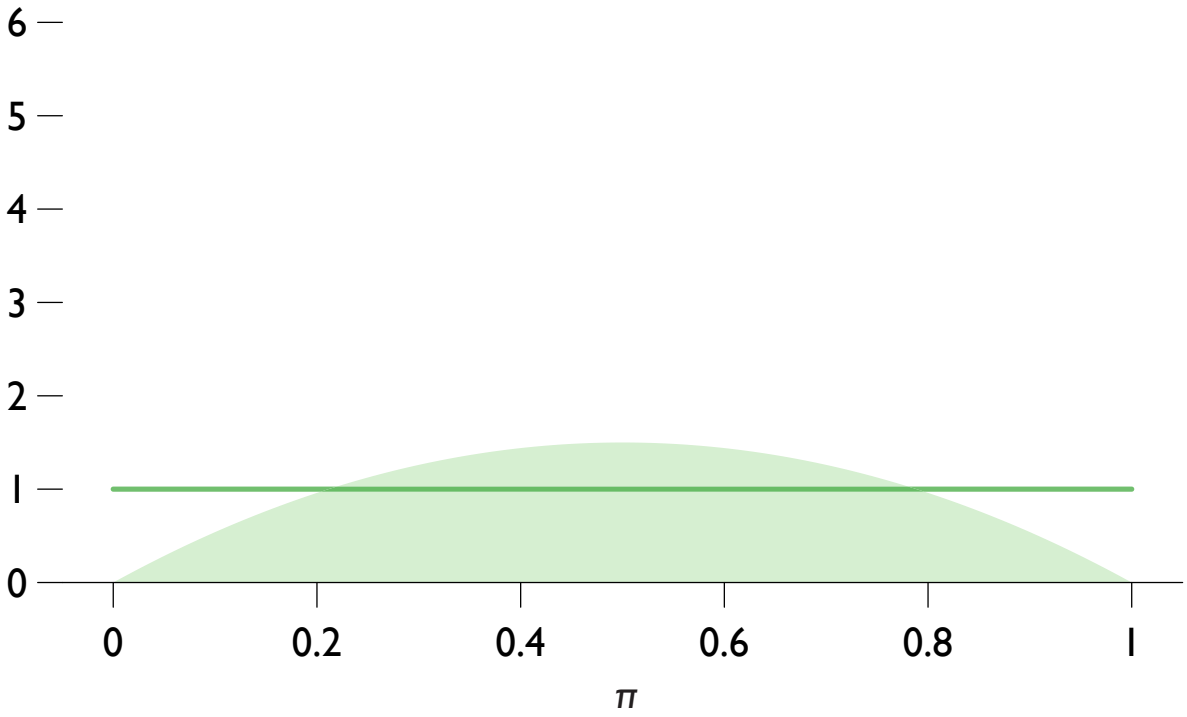
$\text{Beta}(\mu = 0.75, \theta = 4)$

$\text{E}(\pi) = \mu = 0.750$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.037$

The Beta distribution is symmetric when the mean is 0.5

The distribution is remains variable because the sample size is small



$\text{Beta}(\mu = 0.5, \theta = 2)$

$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.083$

$\text{Beta}(\mu = 0.5, \theta = 4)$

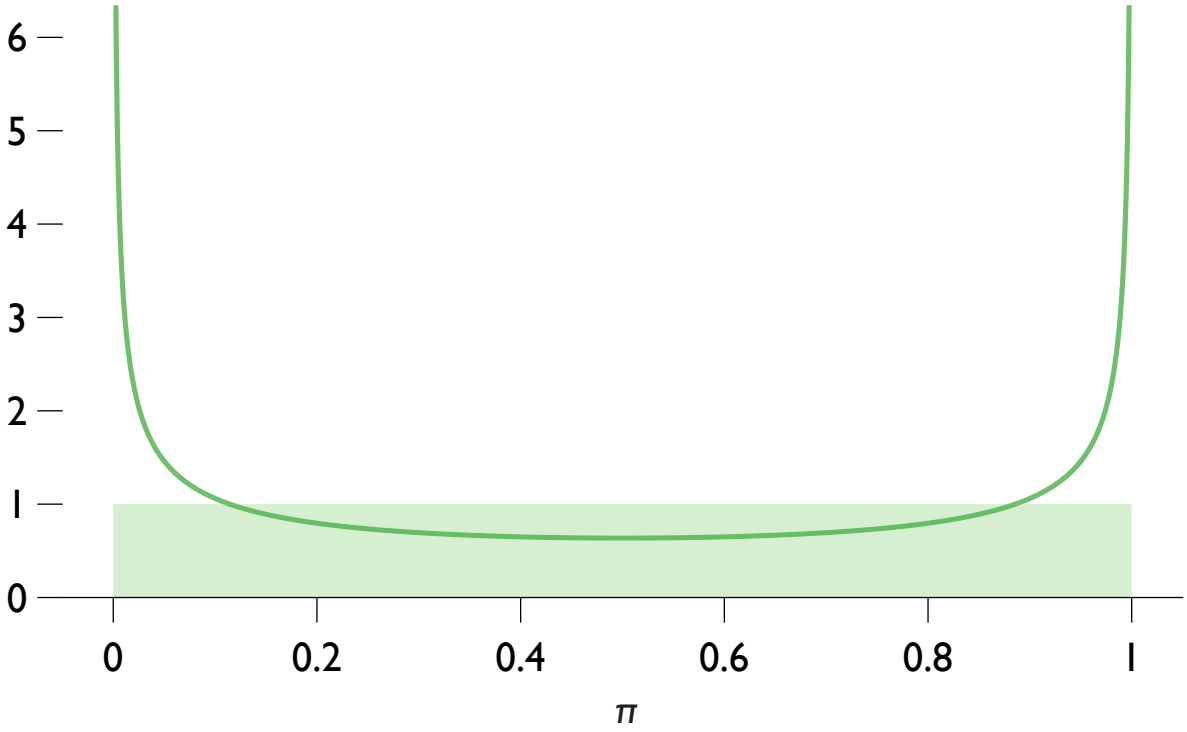
$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.050$

What if we lowered the sample size to $\theta = 2$?

If $\mu = 0.5$, this likely means we see a single success and a single failure

Now there is so little information in the sample that any $\pi \in (0, 1)$ is equally likely!



$\text{Beta}(\mu = 0.5, \theta = 1)$

$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.125$

$\text{Beta}(\mu = 0.5, \theta = 2)$

$\text{E}(\pi) = \mu = 0.500$

$\text{var}(\pi) = \mu(1-\mu)/(\theta+1) = 0.083$

What if we set $0 < \theta < 1$?

This is harder to conceptualize as a discrete set of prior samples

But it does produce a usable Beta pdf with a *bathtub* shape

Now extreme values of π are more likely than moderate ones

Can be very useful for capturing extreme dispersion in probabilities

The Beta-Binomial distribution

Let's revisit the binomial distribution,
but let π_i vary randomly across observations following a Beta distribution

$$\begin{aligned} y_i &\sim \text{Binomial}(\pi_i, M_i) \\ \pi_i &\sim \text{Beta}(\mu_i, \theta) \end{aligned}$$

This is a *compound distribution*:
We have one stochastic component (Beta) embedded inside another (Binomial)

Compound distributions are widely used, e.g., in hierarchical modeling

Note the parameters of the model are implicitly those of the Beta distribution,
as rewriting shows:

$$y_i \sim \text{Beta-Binomial}(\mu_i, \theta, M_i)$$

The Beta-Binomial distribution

$$y_i \sim \text{Beta-Binomial}(\mu_i, \theta, M_i)$$

$$\Pr(y_i) = \frac{\Gamma(M_i+1)}{\Gamma(y_i+1)\Gamma(M_i-y_i+1)} \times \frac{\Gamma(\theta)\Gamma(y_i+\mu_i\theta)\Gamma(M_i-y_i+\theta(1-\mu_i))}{\Gamma(M_i+\theta)\Gamma(\mu_i\theta)\Gamma(\theta(1-\mu_i))}$$

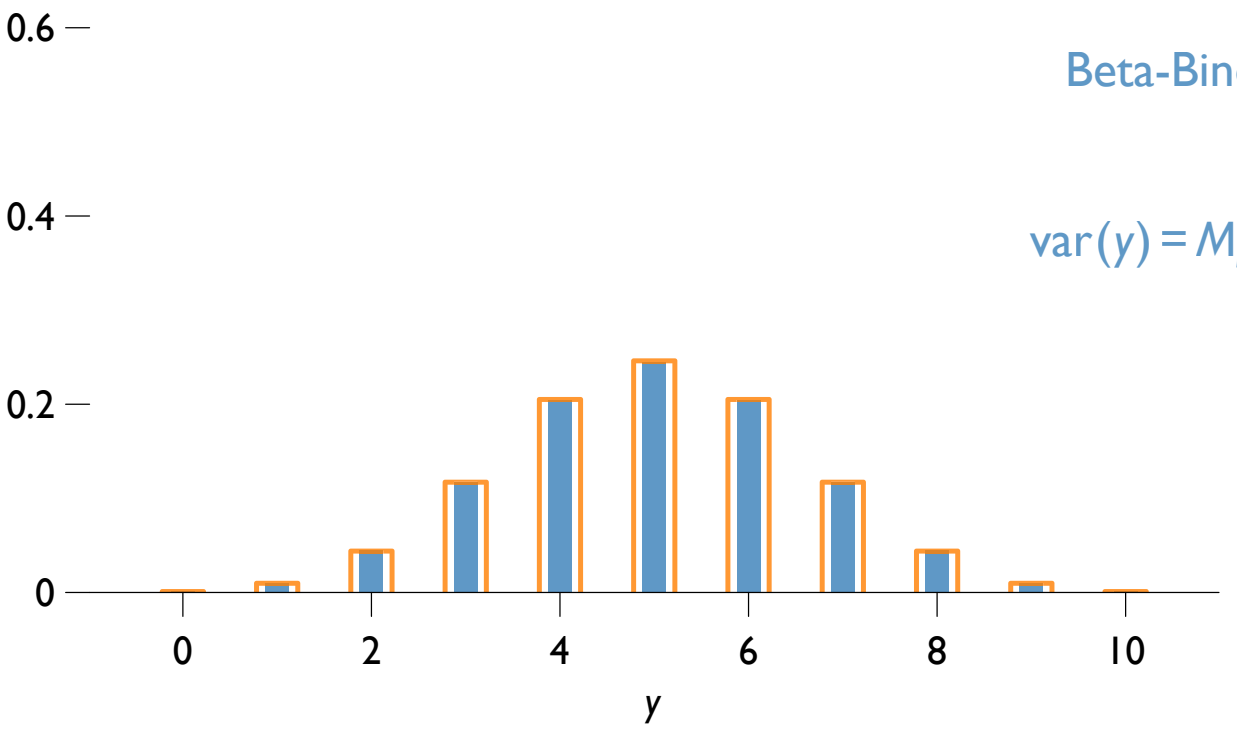
Moments of the Beta-Binomial distribution

$$\mathbb{E}(y_i) = M_i\mu_i \qquad \text{var}(y_i) = M_i\mu_i(1-\mu_i)\left(1+\frac{M_i-1}{\theta+1}\right)$$

The expected count takes the same form as the Binomial

As $\theta \to \infty$, the variance converges on the Binomial variance

For smaller θ, the variance is overdispersed



Beta-Binomial($\mu = 0.5$, $1/\theta = 0.001$, $M = 10$)

$E(y) = M\mu = 5$

$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 2.52$

Binomial($\pi = 0.5$, $M = 10$)

$E(y) = M\pi = 5$

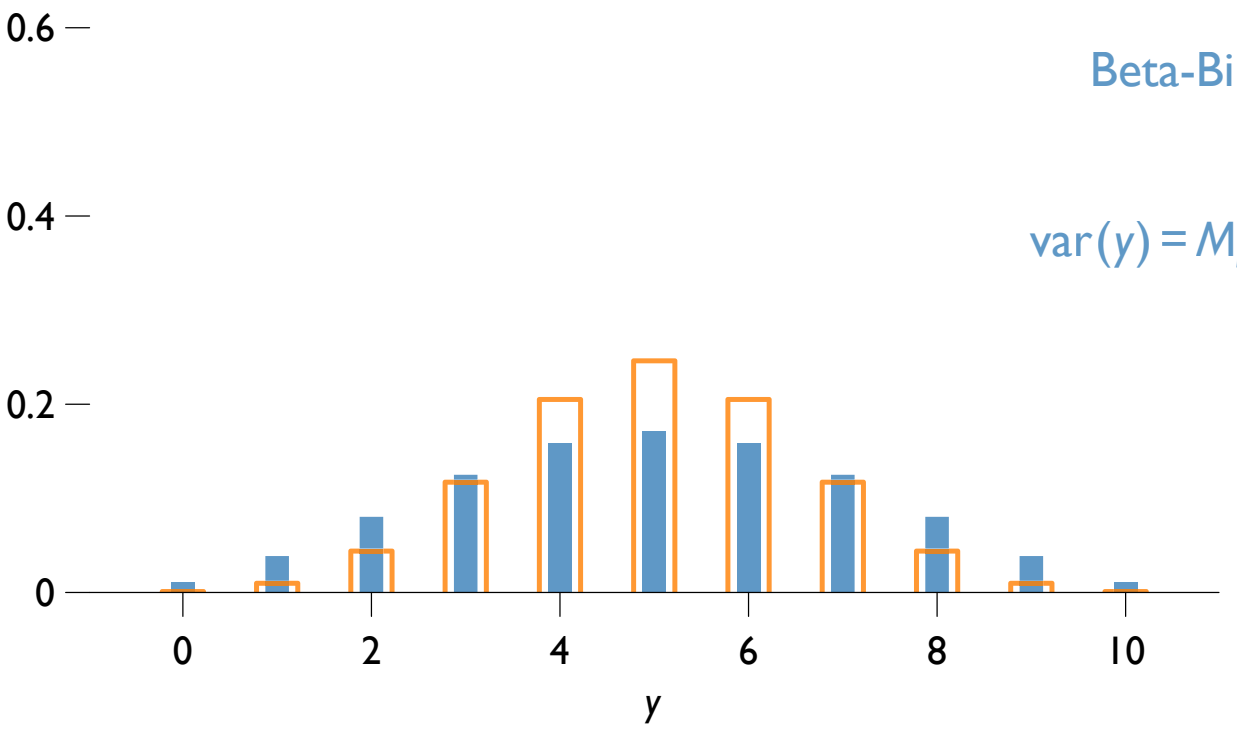
$\text{var}(y) = M\pi(1-\pi) = 2.5$

Let's compare the Beta-Binomial to the Binomial

We'll fix the average probability of success at 0.5 for both models, and the number of trials at 10

What difference does the extra θ parameter make?

As $\theta \to \infty$ or conversely, when $1/\theta \to 0$
the Beta-Binomial approaches the Binomial



Beta-Binomial($\mu = 0.5,\ 1/\theta = 0.1,\ M = 10$)

$E(y) = M\mu = 5$

$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 4.55$

Binomial($\pi = 0.5,\ M = 10$)

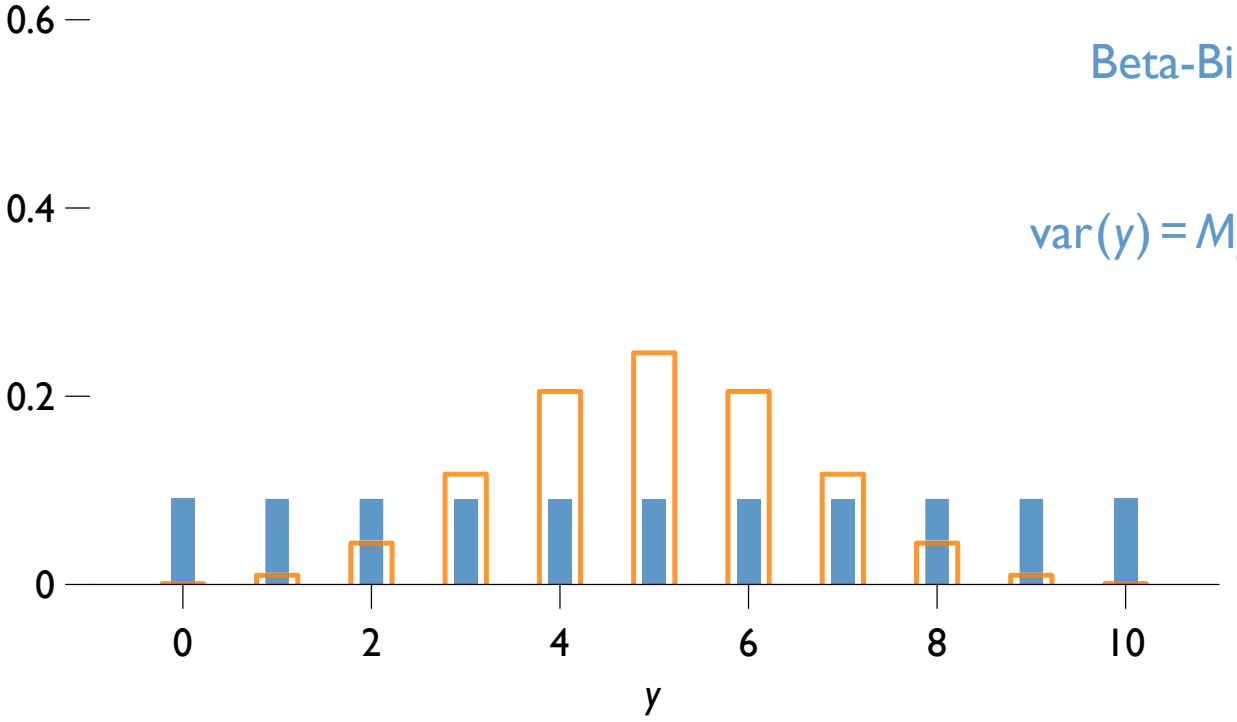
$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

But as we increase $1/\theta$ above 0,
the Beta-Binomial becomes *overdispersed* compared to the Binomial

Including θ in the model allows us to capture the possibility of correlation across trials and resulting higher variance

How to deal with overdispersed bounded counts?
Just estimate θ, a single extra parameter!



Beta-Binomial($\mu = 0.5,\ 1/\theta = 0.5,\ M = 10$)

$E(y) = M\mu = 5$

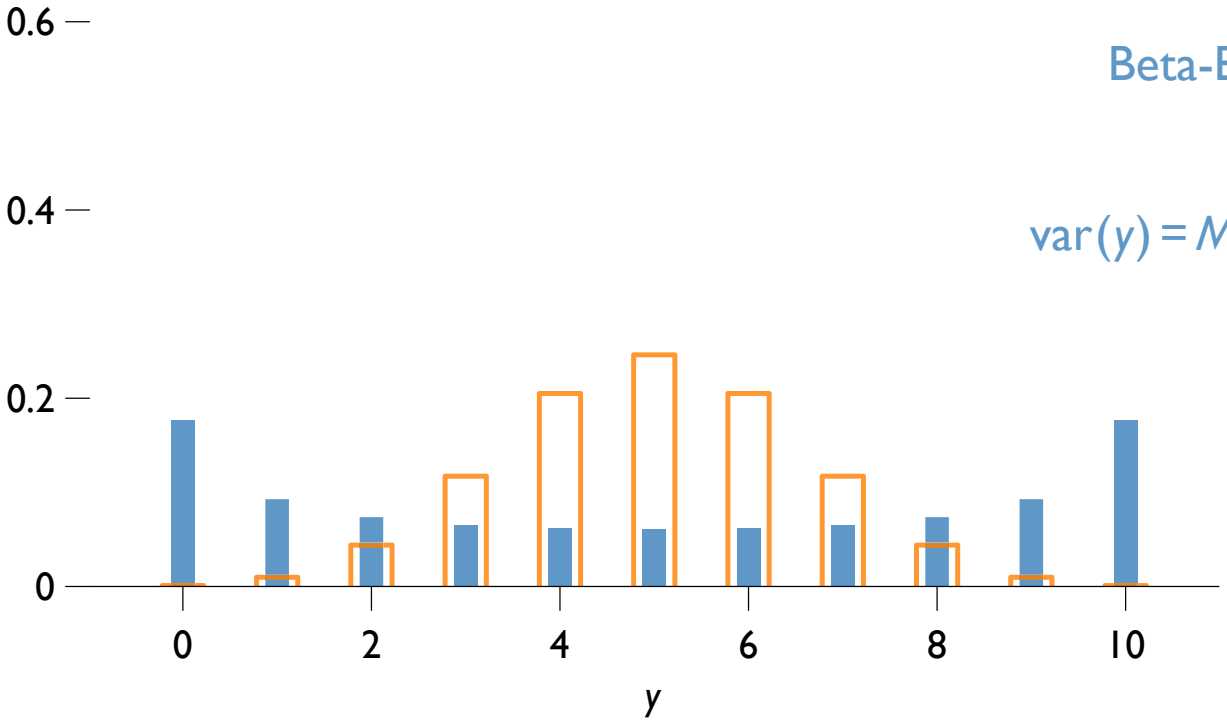
$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 10.00$

Binomial($\pi = 0.5,\ M = 10$)

$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

As in the Beta distribution, setting $\theta = 2$ (or $1/\theta = 0.5$) makes every π – and thus every count – equally likely



Beta-Binomial($\mu = 0.5$, $1/\theta = 1$, $M = 10$)

$E(y) = M\mu = 5$

$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 13.75$

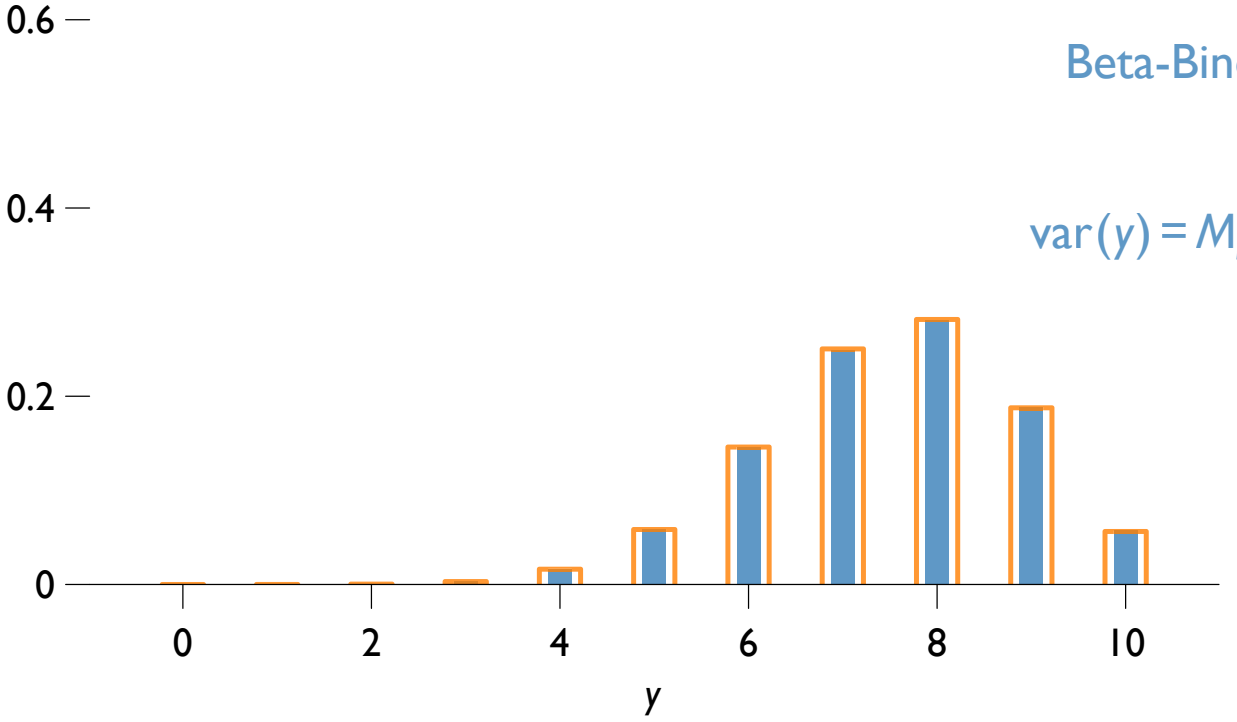
Binomial($\pi = 0.5$, $M = 10$)

$E(y) = M\pi = 5$

$\text{var}(y) = M\pi(1-\pi) = 2.5$

Setting $0 < \theta < 2$ (or $1/\theta > 0.5$) yields still greater overdispersion

All the flexibility of the Beta is present in the Beta-Binomial



Beta-Binomial($\mu = 0.75$, $1/\theta = 0.001$, $M = 10$

$E(y) = M\mu = 7.5$

$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 1.89$

Binomial($\pi = 0.75$, $M = 10$)

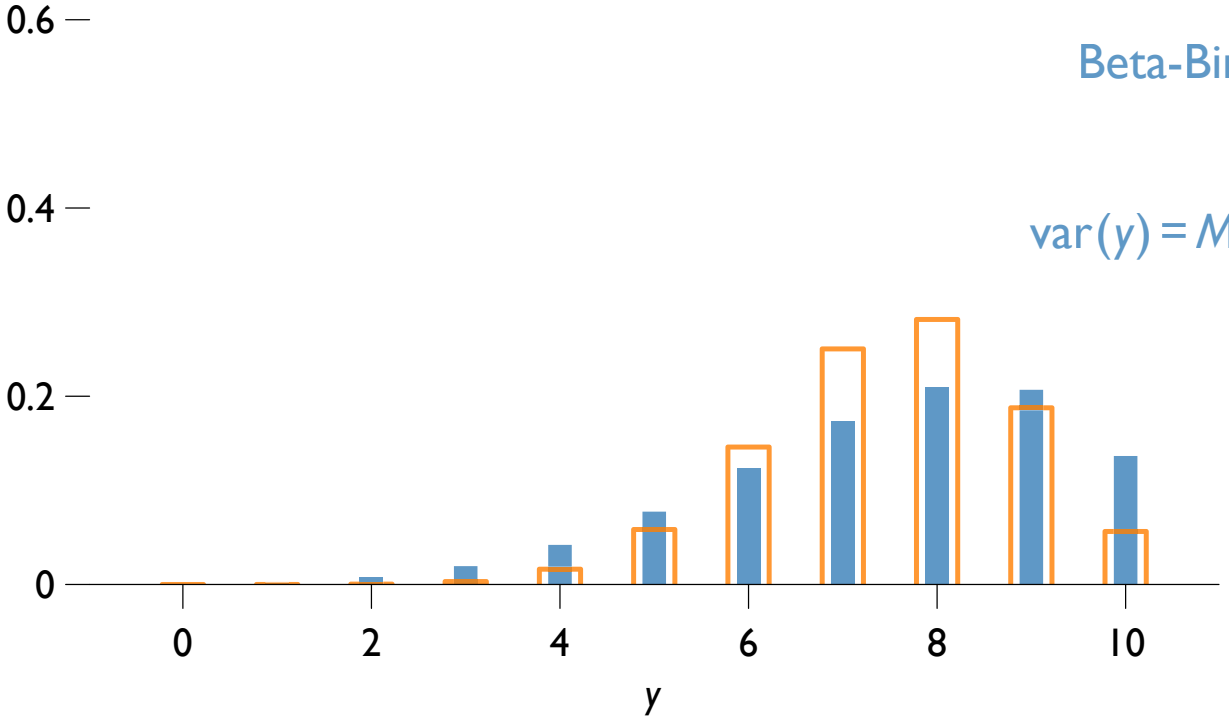
$E(y) = M\pi = 7.5$

$\text{var}(y) = M\pi(1-\pi) = 1.875$

Now let's set the average probability of a success at 0.75

This corresponds to our "hot water" Beta distributions

If $1/\theta \approx 0$, the Beta-Binomial approaches the Binomial



Beta-Binomial($\mu = 0.75,\ 1/\theta = 0.1,\ M = 10$)

$E(y) = M\mu = 7.5$

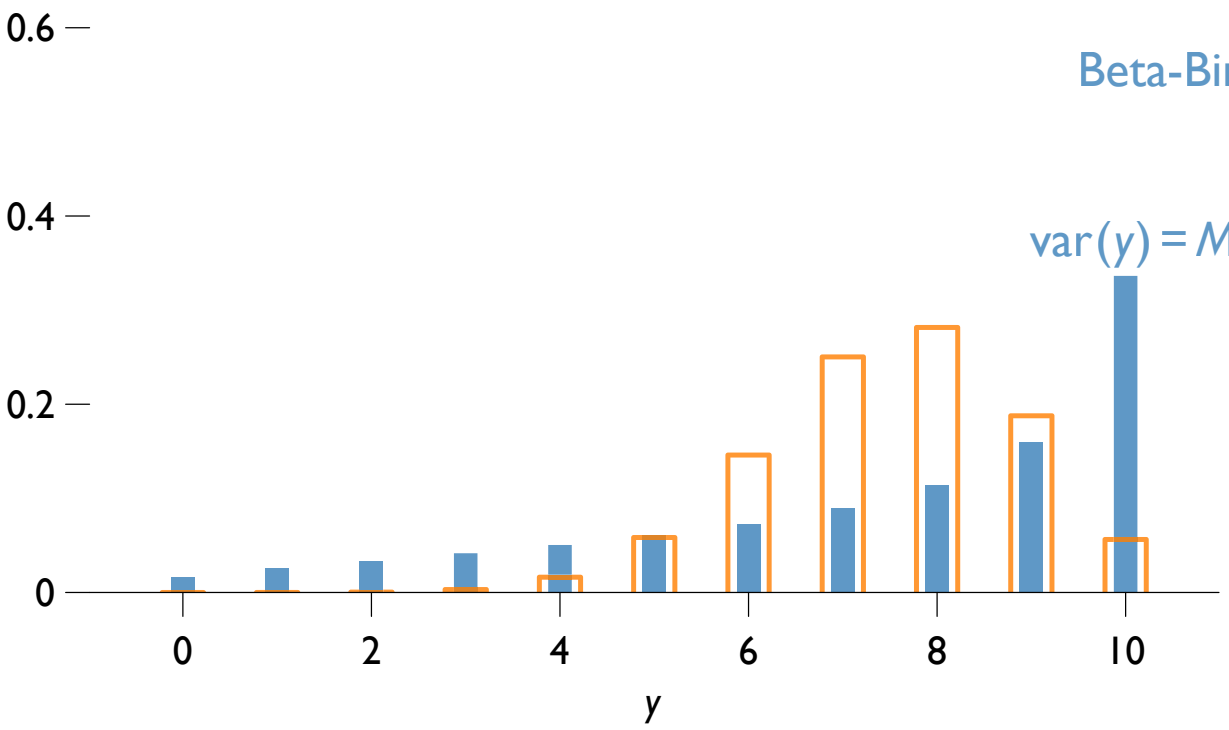
$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 3.41$

Binomial($\pi = 0.75,\ M = 10$)

$E(y) = M\pi = 7.5$

$\text{var}(y) = M\pi(1-\pi) = 1.875$

As $1/\theta$ rises, the Beta-Binomial spreads out to capture overdispersion



Beta-Binomial($\mu = 0.75,\ 1/\theta = 0.5,\ M = 10$)

$E(y) = M\mu = 7.5$

$\text{var}(y) = M\mu(1-\mu)[1 + (M-1)/(1+\theta)] = 7.50$

Binomial($\pi = 0.75,\ M = 10$)

$E(y) = M\pi = 7.5$

$\text{var}(y) = M\pi(1-\pi) = 1.875$

As $1/\theta$ gets larger, so does the dispersion

Upshot: The Beta-Binomial is a flexible model of bounded counts that can either approximate binomial dispersion or any level of overdispersion

Estimating $1/\theta$ as a dispersion parameter will:

1. Solve our binomial overconfidence problem
2. Provide better estimates of other parameters

Beta-Binomial regression

Beta-Binomial regression can be written two ways

Either as the combination of Beta and Binomial distributions

$$
\begin{aligned}
y_i &\sim \text{Binomial}(\pi_i, M_i) \\
\pi_i &\sim \text{Beta}(\mu_i, \theta) \\
\mu_i &= \text{logit}^{-1}(\text{x}_i\boldsymbol{\beta})
\end{aligned}
$$

Or reparameterized into a single stochastic component

$$
\begin{aligned}
y_i &\sim \text{Beta-Binomial}(\mu_i, \theta, M_i) \\
\mu_i &= \text{logit}^{-1}(\text{x}_i\boldsymbol{\beta})
\end{aligned}
$$

Beta-Binomial regression

$$
\begin{aligned}
y_i &\sim \text{Beta-Binomial}(\mu_i, \theta, M_i) \\
\mu_i &= \text{logit}^{-1}(\text{x}_i\boldsymbol{\beta})
\end{aligned}
$$

$\boldsymbol{\beta}$ coefficients can be interpreted as in the Binomial or logit models

Estimation is by maximum likelihood as usual;
use `vglm(model, data, family=betabinomial)` in the `VGAM` package

Beta-Binomial regression

$$
\begin{aligned}
y_i &\sim \text{Beta-Binomial}(\mu_i, \rho, M_i) \\
\mu_i &= \text{logit}^{-1}(\text{x}_i\boldsymbol{\beta})
\end{aligned}
$$

$\boldsymbol{\beta}$ coefficients can be interpreted as in the Binomial or logit models

Estimation is by maximum likelihood as usual;
use `vglm(model, data, family=betabinomial)` in the `VGAM` package

Note that `vglm()` uses yet another parameterization: μ and ρ

ρ is the correlation of trials within an observation: $1/(1+\theta)$

So the variance of y_i is now $M_i\mu_i(1-\mu_i)\left[1+\rho(M_i-1)\right]$

	Least Squares	Binomial	Beta-Binomial
log Income	-0.05	-0.10	-0.31
	(0.05)	(0.05)	(0.32)
College	0.27	0.82	1.73
	(0.12)	(0.09)	(0.78)
Intercept	1.28	2.17	4.35
	(0.55)	(0.49)	(3.30)
log ρ			-4.89
			(0.27)
N	39	39	39
log $\mathcal{L}$	—	-7437	-326
AIC	—	14881	660
In-sample MAE (null=3.25%)	2.84%	2.96%	2.82%
5-fold CV MAE (null=3.34%)	3.13%	3.28%	3.07%

The Beta-Binomial is complicated to understand, but easy to estimate in `R`

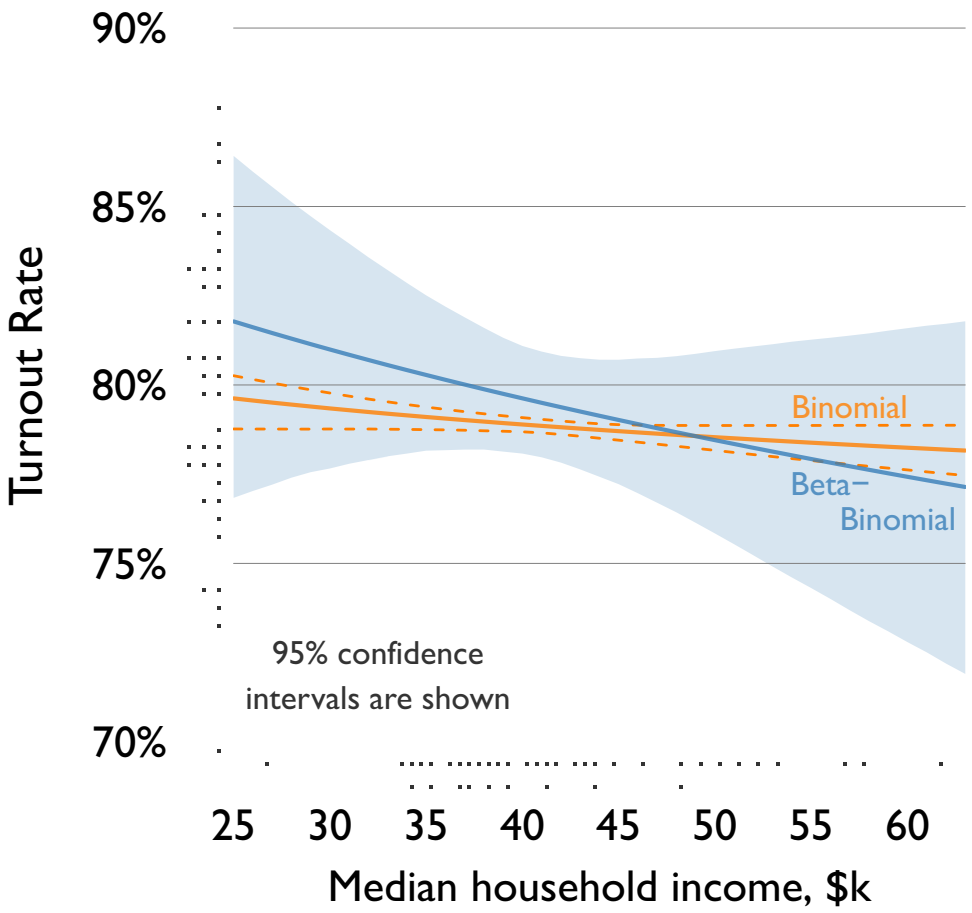
Caveat: `vglm(..., family=betabinomial)` reports log(correlation), rather than a dispersion or sample size parameters

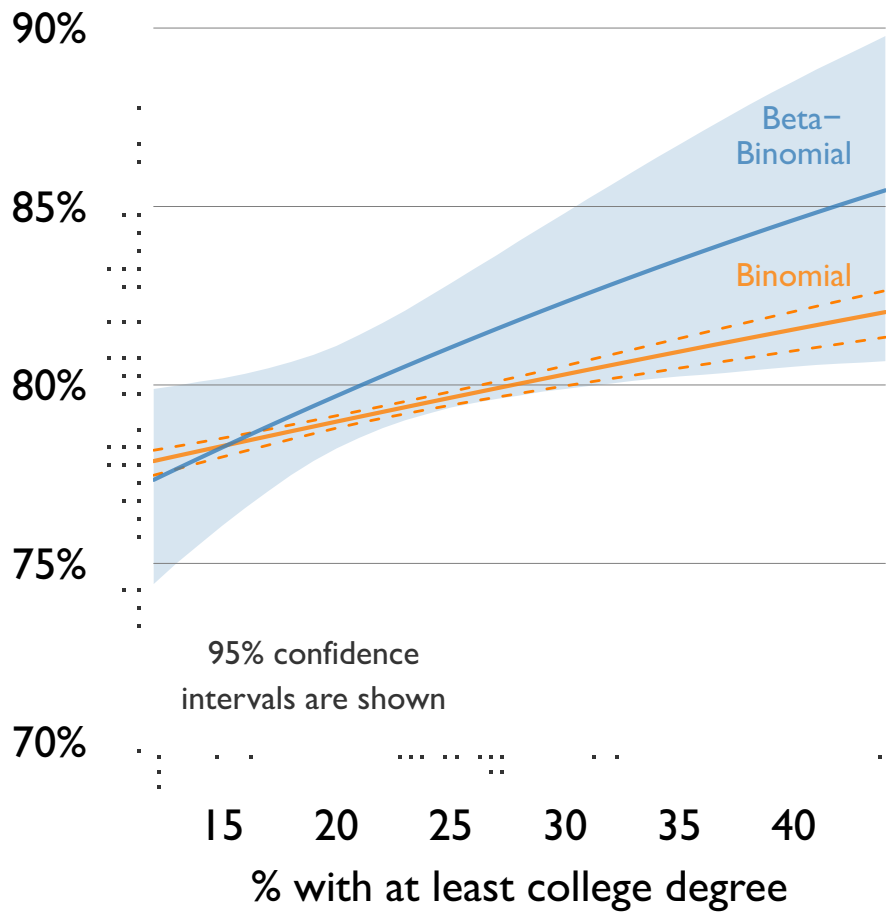
	Least Squares	Binomial	Beta-Binomial
log Income	-0.05	-0.10	-0.31
	(0.05)	(0.05)	(0.32)
College	0.27	0.82	1.73
	(0.12)	(0.09)	(0.78)
Intercept	1.28	2.17	4.35
	(0.55)	(0.49)	(3.30)
Correlation (ρ)			0.0078
			(0.0021)
N	39	39	39
log $\mathcal{L}$	—	-7437	-326
AIC	—	14881	660
In-sample MAE (null=3.25%)	2.84%	2.96%	2.82%
5-fold CV MAE (null=3.34%)	3.13%	3.28%	3.07%

A bit of simulation turns this into the correlation ρ and its standard error

You could also convert it to θ or $1/\theta$

What can we learn from comparing the Beta-Binomial results?

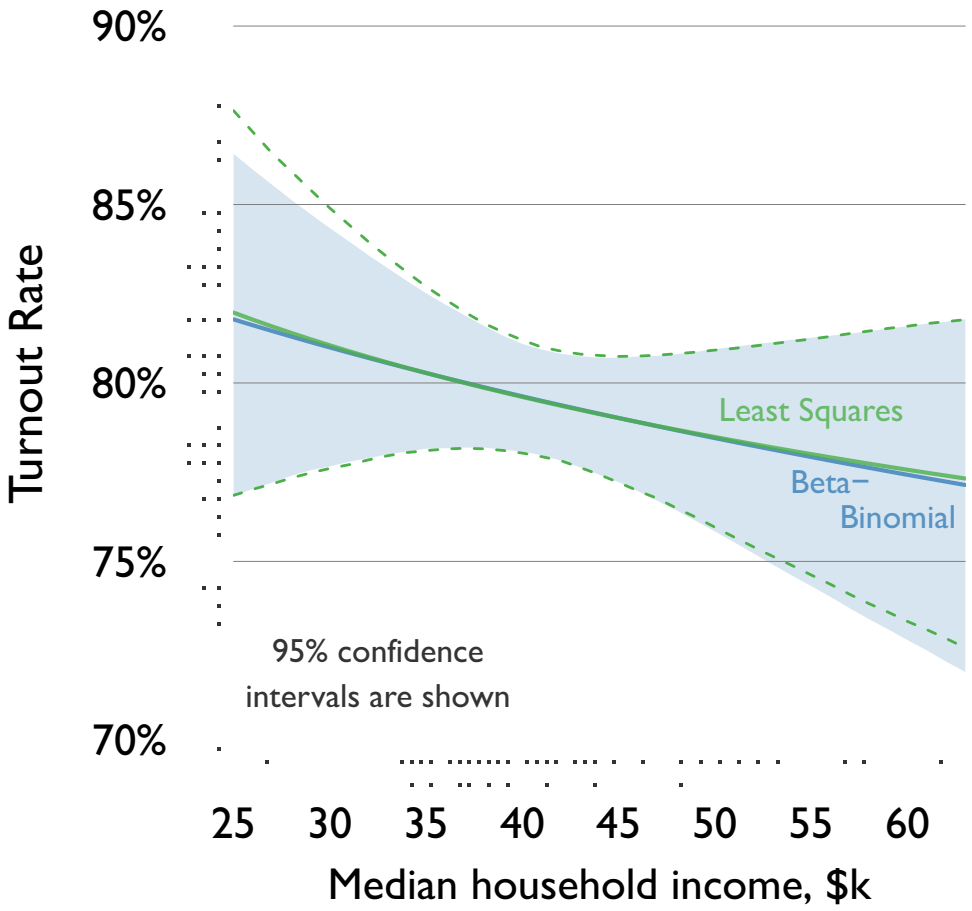


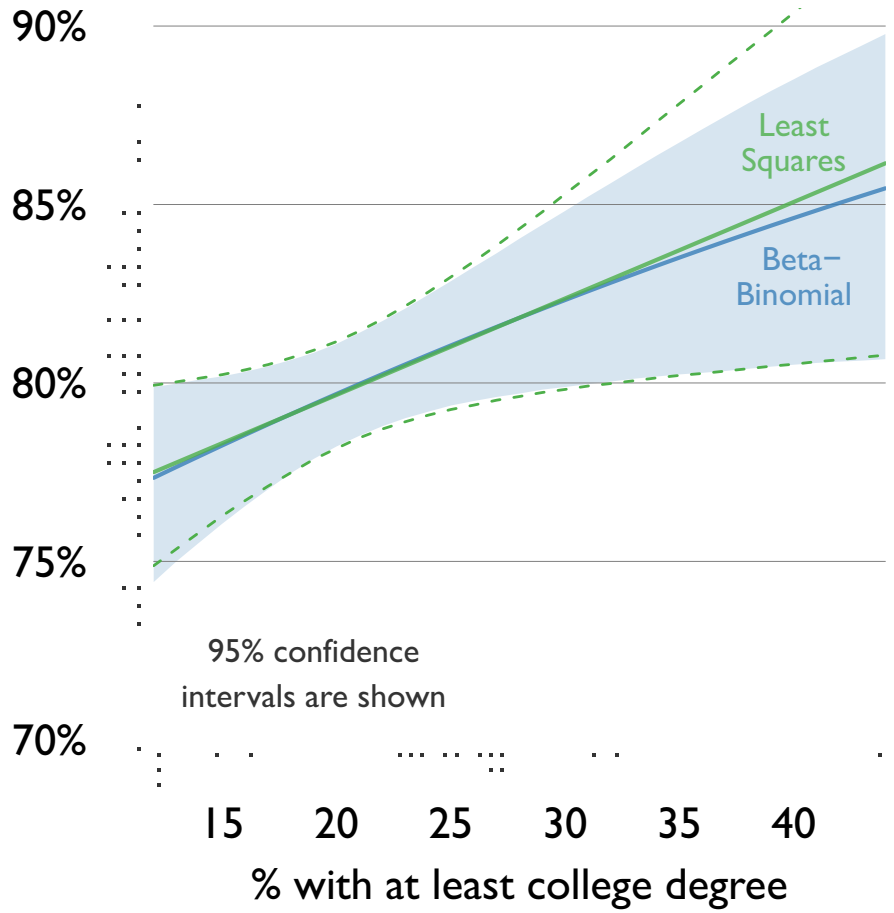


The Beta-Binomial finds somewhat different relationships than the Binomial

And *much* larger and more realistic CIs

Correlation within vote counts implies less independent information about turnout





Compared to least squares, Beta-Binomial regression yields similar expected values

CIs are even fairly similar – least squares accounts for overdispersion using σ^2

Different models won't always be so similar:
(1) if turnout rates had wider range; (2) if turnout rates clumped at 0 or 1;
(3) if M_i had a wider range; (4) did imputation make everything more "Normal"?

Quasilikelihood and the quasibinomial

Binomial regression is clearly untrustworthy when grouped counts are correlated

As contagion across trials is typical in social science settings,
we expect binomial results to be overconfident in general

We've seen one solution:
Derive a probability distribution with overdispersion from correlated counts

Quasilikelihood and the quasibinomial

Binomial regression is clearly untrustworthy when grouped counts are correlated

As contagion across trials is typical in social science settings,
we expect binomial results to be overconfident in general

We've seen one solution:
Derive a probability distribution with overdispersion from correlated counts

Another solution:
Take the binomial, and multiply its variance to make it more dispersed

This approach produces the *quasibinomial*, so-called because it isn't quite binomial –
or even a proper probability distribution at all!

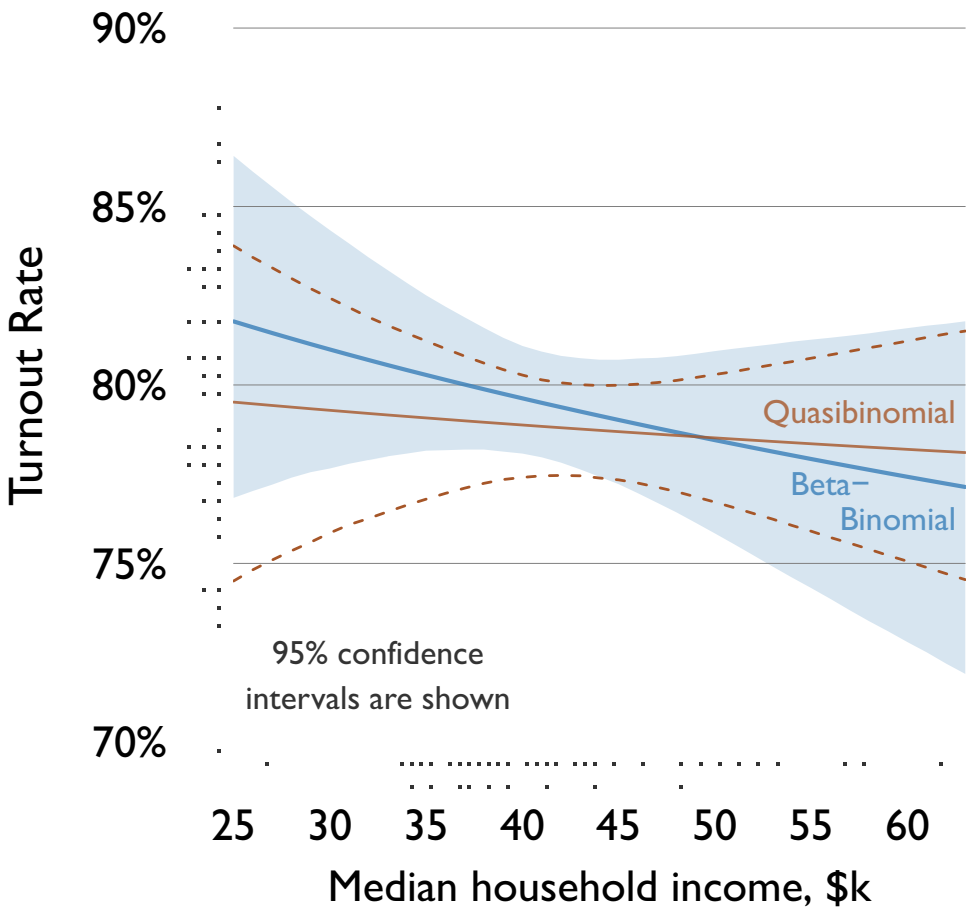
Using the rescaled "quasibinomial" instead of the binomial in a GLM
model is a quick and dirty fix for overdispersion

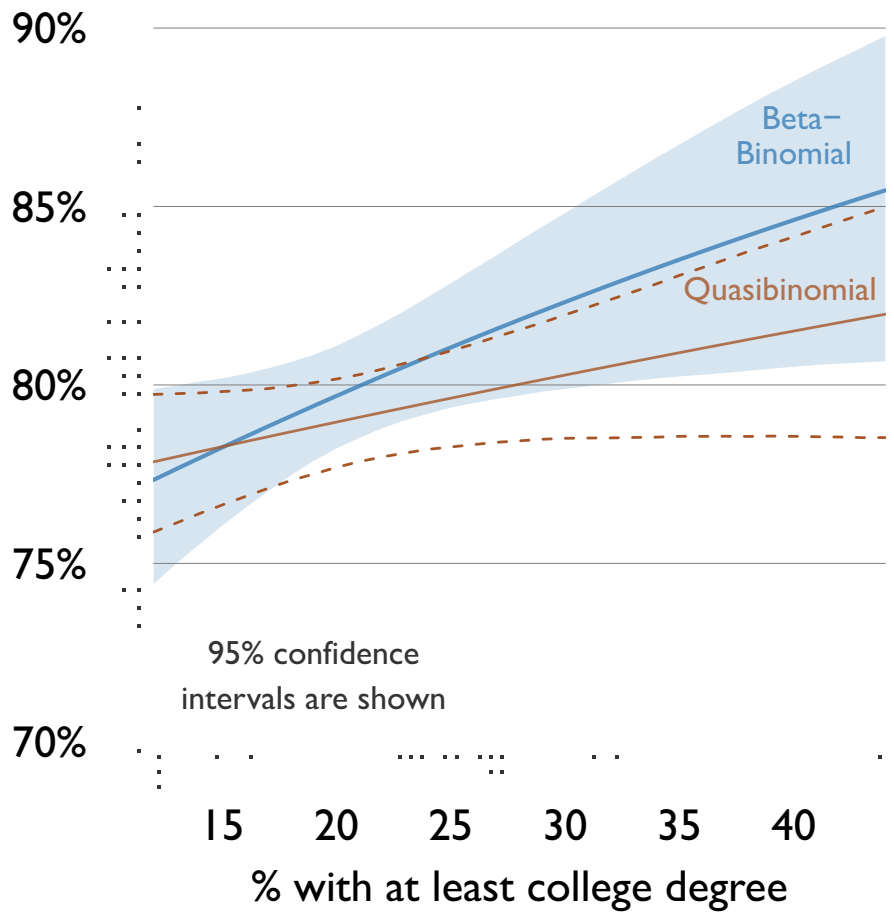
Not technically an MLE – doesn't produce a proper likelihood –
so the method is called *quasilikelihood*

	Least Squares	Binomial	Quasi-Binomial	Beta-Binomial
log Income	-0.05	-0.10	-0.10	-0.31
	(0.05)	(0.05)	(0.26)	(0.32)
College	0.27	0.82	0.82	1.73
	(0.12)	(0.09)	(0.46)	(0.78)
Intercept	1.28	2.17	2.17	4.35
	(0.55)	(0.49)	(2.67)	(3.30)
Correlation (ρ)				0.0078
				(0.0021)
N	39	39	39	39
log $\mathcal{L}$	—	-7437	—	-326
AIC	—	14881	—	660
In-sample MAE (null=3.25%)	2.84%	2.96%	2.96%	2.82%
5-fold CV MAE (null=3.34%)	3.13%	3.28%	3.29%	3.07%

The Quasibinomial parameters and fit are just like the Binomial

But the standard errors of parameter estimates are more like the Beta-Binomial

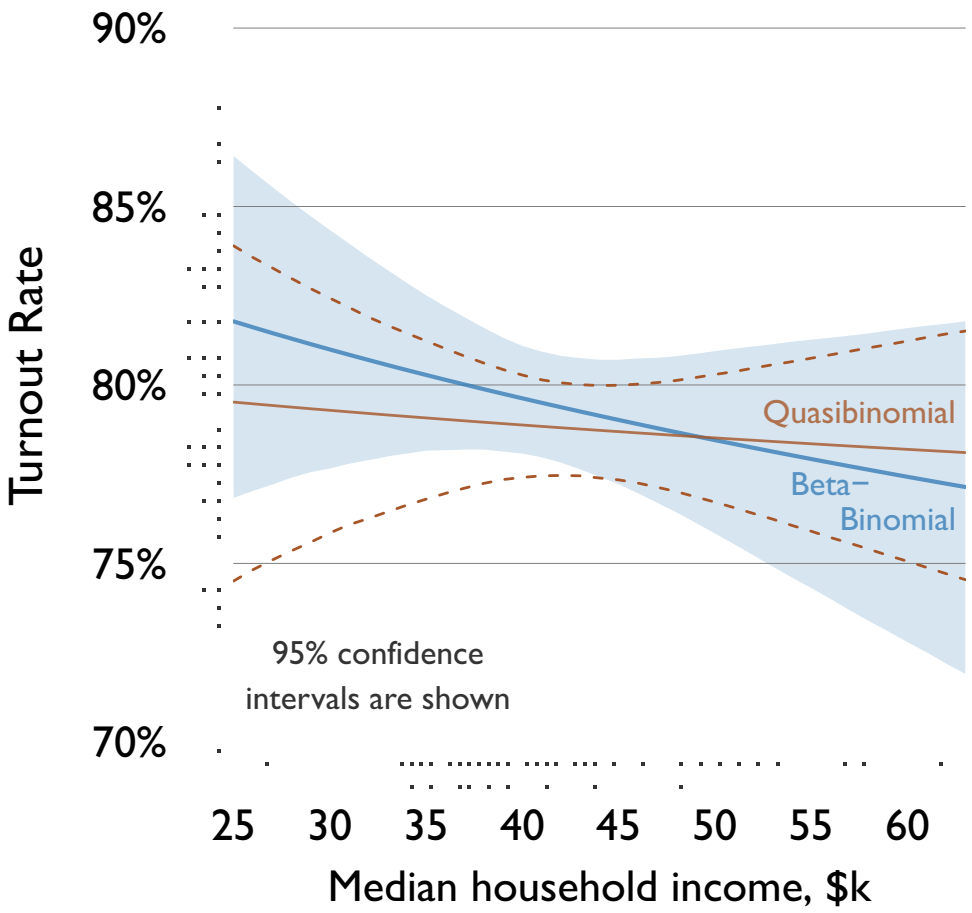


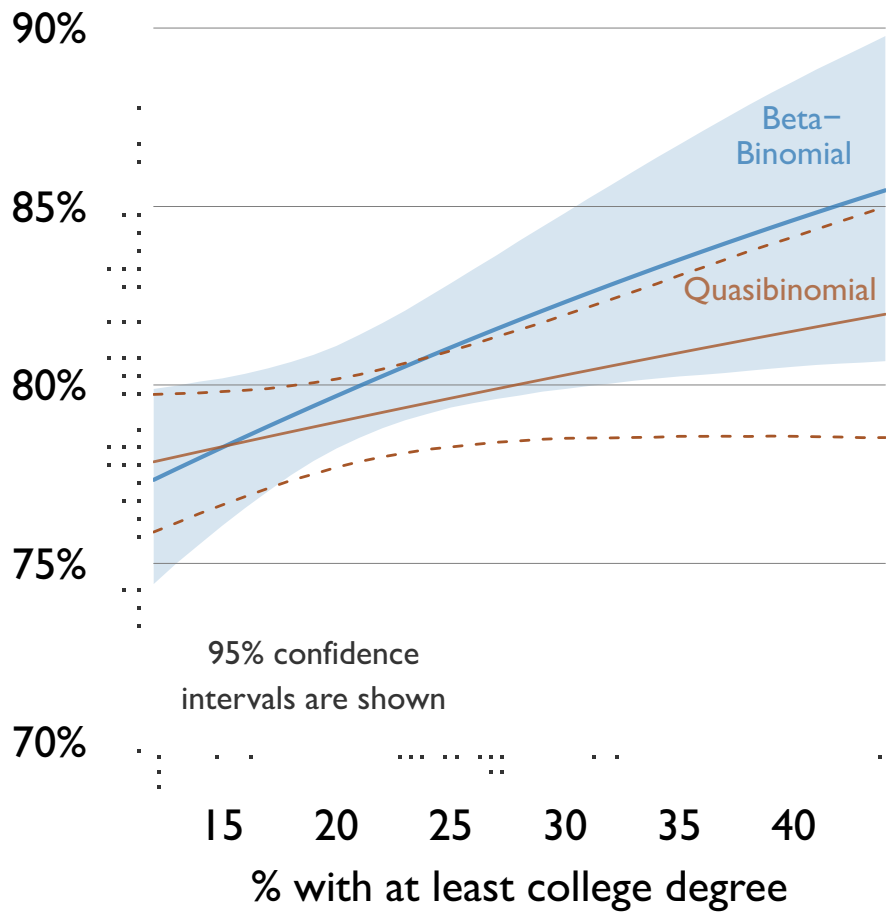


The quasibinomial starts with the binomial estimates, then scales up the variance

So expected values are exactly the same as for the binomial

But confidence intervals are *much* wider, similar to the Beta-Binomial





Which model should you use?

Dispersion was very important here; the binomial nature of the data less so

Can't know that *ex ante*, so try all reasonable options

When models offer substantively different results based on mostly arbitrary differences in assumptions, report the differences & the goodness of fit

When to use (beta-)binomial regression

When grouped counts are uncorrelated,
the binomial produces much more efficient results than other models, including least squares

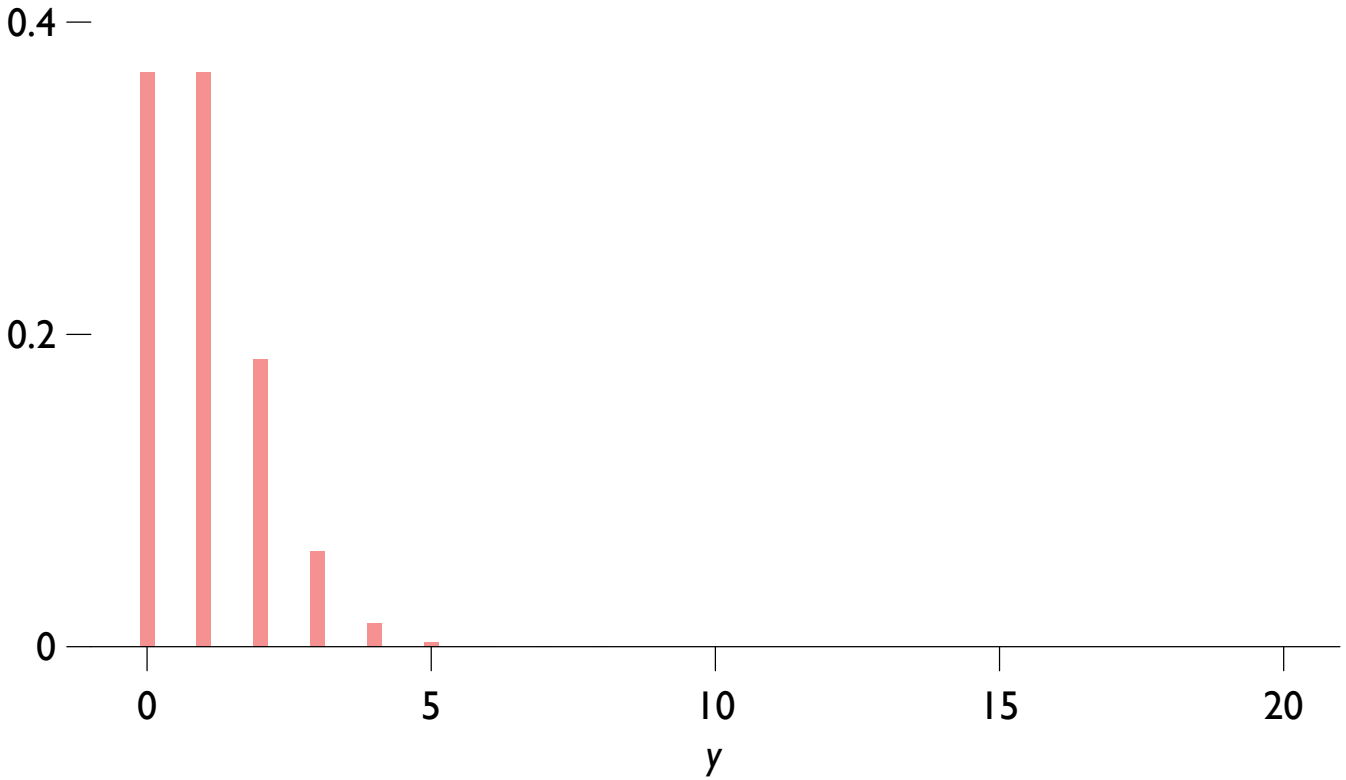
But when counts are correlated, the binomial is drastically overconfident

Beta-binomial or quasibinomial regression are often the best choice for counts with clear upper bounds, such as aggregate voting behavior

But they're not always appropriate

- Beta-binomial and quasibinomial regression are unsuitable when M_i is unknown
- Both models are really unsuitable when M_i is undefined or infinite
- May be biased when π_i is small (see rare-events logit)

Fortunately, there are a better options tailor-made for these situations . . .



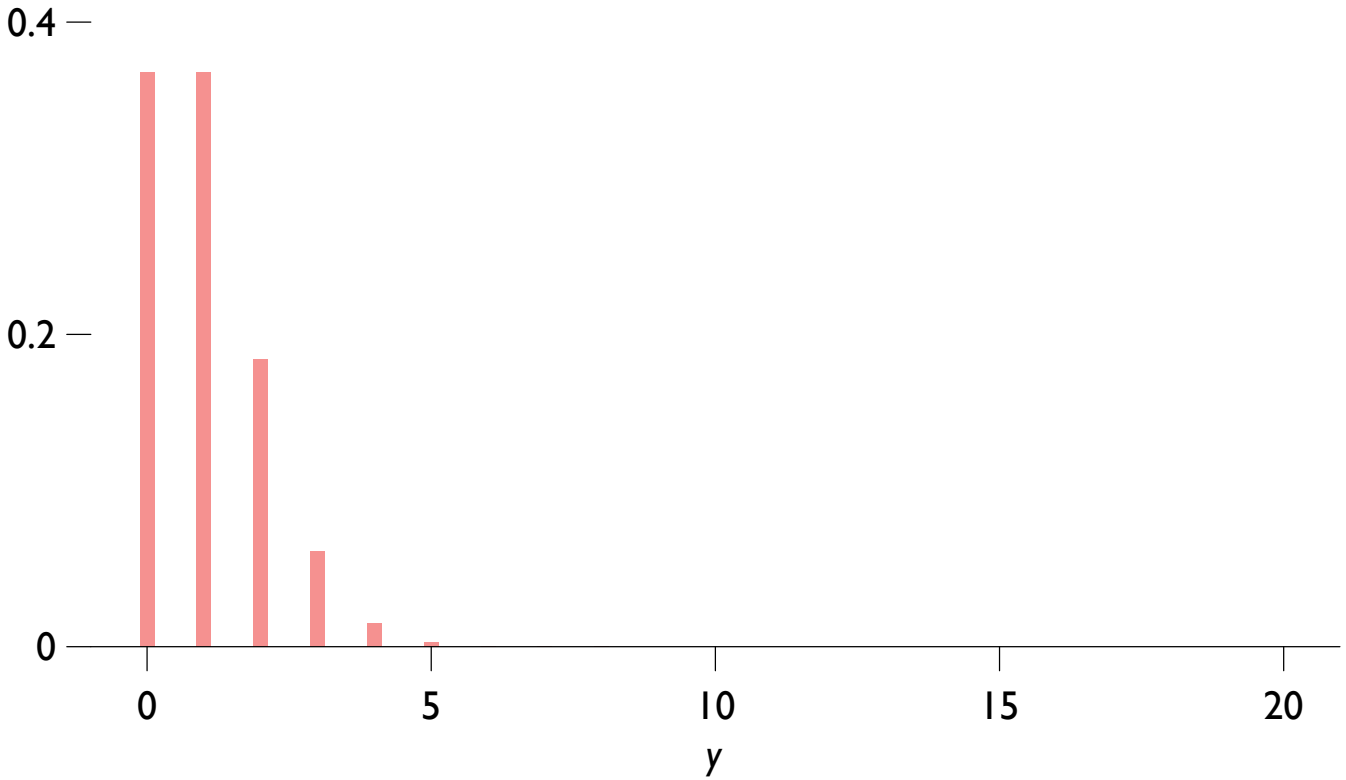
$\text{Poisson}(\lambda = 1)$

$E(y) = \lambda = 1$

$\text{var}(y) = \lambda = 1$

The Poisson distribution describes
the *unbounded* number of events occurring in a period of continuous time

These periods – more generally, the population at risk – can vary in size



Poisson($\lambda = 1$)

$E(y) = \lambda = 1$

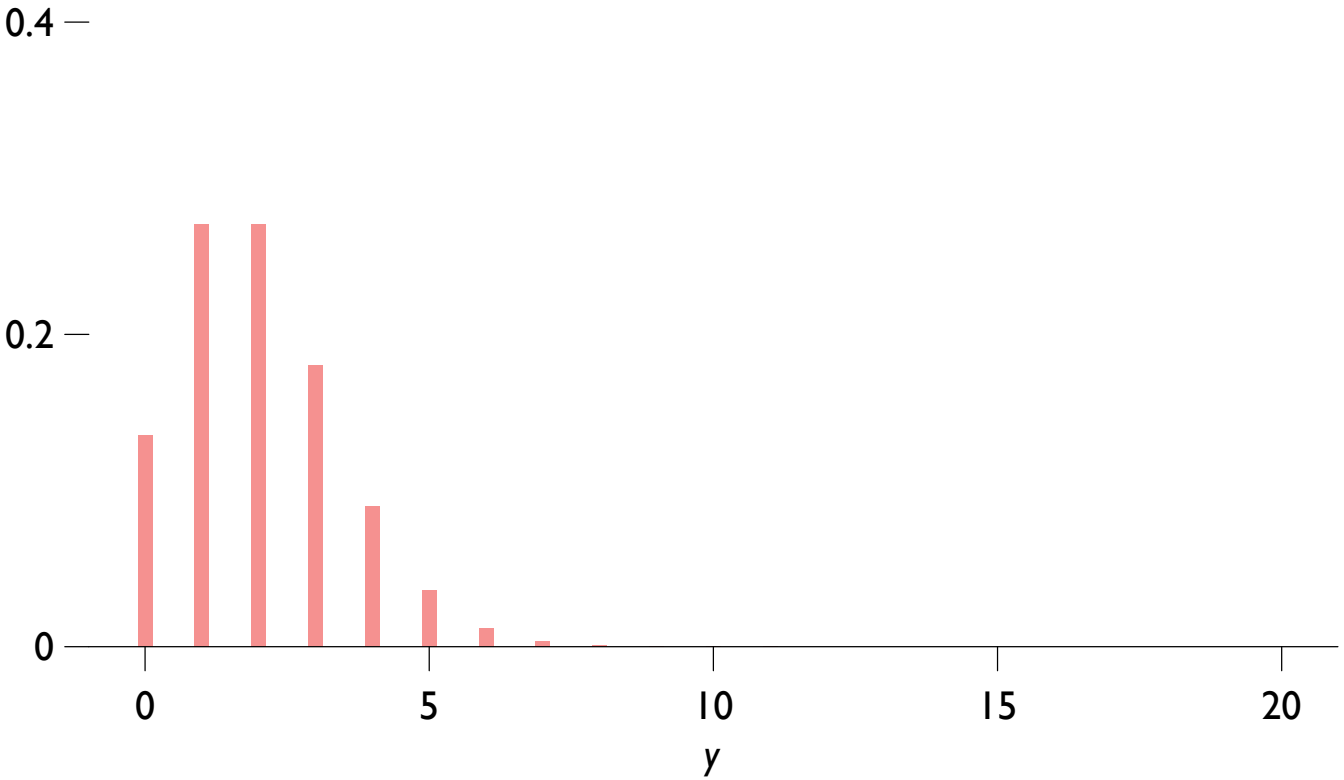
$\text{var}(y) = \lambda = 1$

Poisson distribution assumptions

1. The starting count is zero

2. Only one event can occur at a given time

3. Pr(event at time t) is constant –
that is, independent of the occurence of previous events



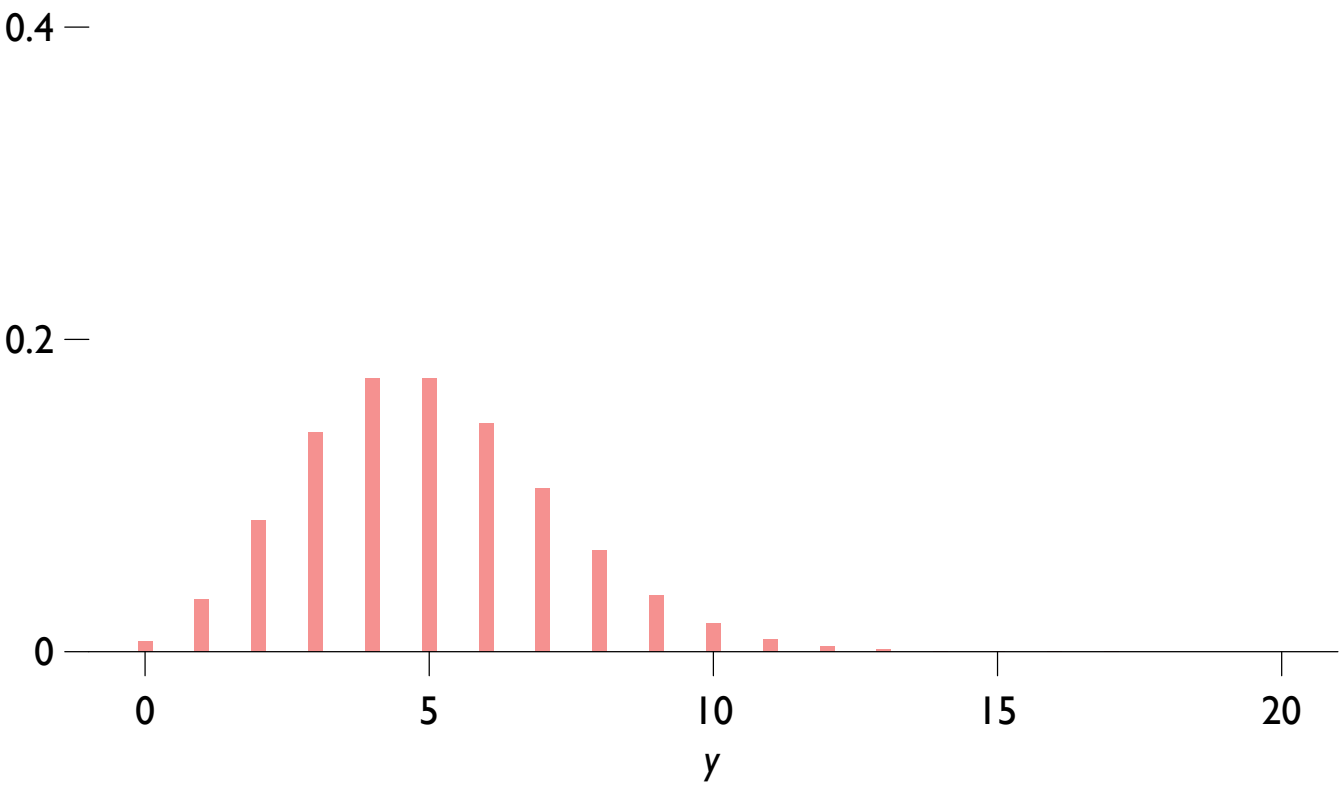
Poisson($\lambda = 2$)

$E(y) = \lambda = 2$

$\text{var}(y) = \lambda = 2$

The Poisson pdf, graphed above for example λ's

$$\Pr(y_i|\lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$



Poisson($\lambda = 5$)

$E(y) = \lambda = 5$

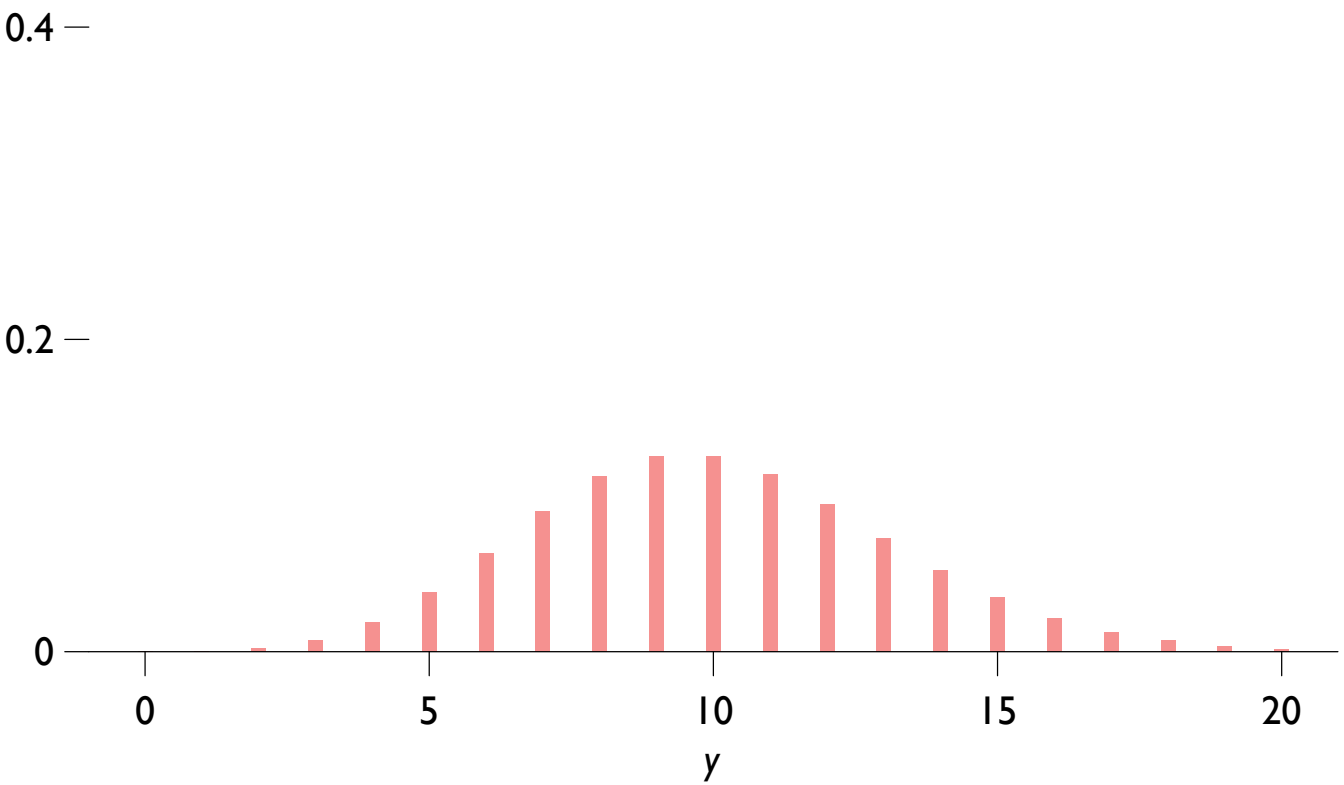
$\text{var}(y) = \lambda = 5$

The Poisson has a single parameter, λ, that gives both the mean and variance

$$\mathbb{E}(y) = \text{var}(y) = \lambda$$

This means that overdispersion is impossible in the Poisson distribution

A Poisson with a given expected count *always* has the same variance



Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$\text{var}(y) = \lambda = 10$

As λ increases, the Poisson becomes more symmetric

But not perfectly so:

Poisson is always truncated on the left at 0

and always has a long right tail (going to $+\infty$)

Deriving the Poisson from the Binomial distribution

$$\frac{m!}{y!(m-y)!}\pi^y(1-\pi)^{m-y}$$

Using the definition of factorials & multiplying π by m/m

$$\frac{m(m-1)\cdots(m-y+1)}{y!}\left(\frac{m\pi}{m}\right)^y\left(1-\frac{m\pi}{m}\right)^{m-y}$$

Deriving the Poisson from the Binomial distribution

$$\frac{m!}{y!(m-y)!}\pi^y(1-\pi)^{m-y}$$

Using the definition of factorials & multiplying π by m/m

$$\frac{m(m-1)\cdots(m-y+1)}{y!}\left(\frac{m\pi}{m}\right)^y\left(1-\frac{m\pi}{m}\right)^{m-y}$$

Separating terms . . .

$$m(m-1)\cdots(m-y+1)\frac{1}{y!}(m\pi)^y\left(\frac{1}{m}\right)^y\left(1-\frac{m\pi}{m}\right)^m\left(1-\frac{m\pi}{m}\right)^{-y}$$

Deriving the Poisson from the Binomial distribution

$$\frac{m!}{y!(m-y)!}\pi^y(1-\pi)^{m-y}$$

Using the definition of factorials & multiplying π by m/m

$$\frac{m(m-1)\cdots(m-y+1)}{y!}\left(\frac{m\pi}{m}\right)^y\left(1-\frac{m\pi}{m}\right)^{m-y}$$

Separating terms . . .

$$m(m-1)\cdots(m-y+1)\frac{1}{y!}(m\pi)^y\left(\frac{1}{m}\right)^y\left(1-\frac{m\pi}{m}\right)^m\left(1-\frac{m\pi}{m}\right)^{-y}$$

. . . and recollecting them gives us something we take limits of

$$\frac{m(m-1)\cdots(m-y+1)}{m^y}\times\frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m(1-\pi)^{-y}$$

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Term 1 $\to 1$ as $m \to \infty$

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Term 1 $\to 1$ as $m \to \infty$

Term 2 $\to \frac{\lambda^y}{y!}$ as $m\pi \to \lambda$ (trivially)

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Term 1 $\to 1$ as $m \to \infty$

Term 2 $\to \frac{\lambda^y}{y!}$ as $m\pi \to \lambda$ (trivially)

Term 3 $\to \exp(-\lambda)$ as $m \to \infty$ and $m\pi \to \lambda$

Why? The definition of $\exp(a)$ is in fact this very limit:

$$\exp(a) = \lim_{m\to\infty}\left(1+\frac{a}{m}\right)^m$$

Substituting $-m\pi$ for a, then taking the limit as $m\pi \to \lambda$, yields Term 3 above

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Term 1 $\to 1$ as $m \to \infty$

Term 2 $\to \frac{\lambda^y}{y!}$ as $m\pi \to \lambda$ (trivially)

Term 3 $\to \exp(-\lambda)$ as $m \to \infty$ and $m\pi \to \lambda$

Term 4 $\to 1$ as $\pi \to 0$

Deriving the Poisson from the Binomial distribution

Now take the limit as $m \to \infty$, $\pi \to 0$, and $m\pi \to \lambda$

$$\frac{m(m-1)\cdots(m-y+1)}{m^y} \times \frac{(m\pi)^y}{y!}\left(1-\frac{m\pi}{m}\right)^m (1-\pi)^{-y}$$

Term 1 $\to 1$ as $m \to \infty$

Term 2 $\to \frac{\lambda^y}{y!}$ as $m\pi \to \lambda$ (trivially)

Term 3 $\to \exp(-\lambda)$ as $m \to \infty$ and $m\pi \to \lambda$

Term 4 $\to 1$ as $\pi \to 0$

So, we get $\dfrac{\exp(-\lambda)\lambda^y}{y!}$, the Poisson distribution

The Poisson is the limiting distribution of the Binomial
as the number of trials gets very large
and the probability of success gets very small

Poisson Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{N} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Poisson Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{N} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Take logs

$$\log \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{N} y_i \log \lambda_i - \lambda_i - \log y_i!$$

Poisson Likelihood

Form the likelihood from the probability

$$\mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{N} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Take logs

$$\log \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{N} y_i \log \lambda_i - \lambda_i - \log y_i!$$

Reduce to sufficient statistics

$$\log \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{N} y_i \log \lambda_i - \lambda_i$$

(simplest likelihood we've seen yet . . .)

So much for the stochastic component. How do we model λ?

Poisson Likelihood

The Poisson parameter λ must be positive valued

Intuition suggests it should also be nonlinear. . .

Suppose $\mathbb{E}(y) = \lambda$ is "small," such as $\lambda = 1$

At this level, suppose the expected effect of $\Delta x = 1$ is a $\beta \uparrow$ in y

Poisson Likelihood

The Poisson parameter λ must be positive valued

Intuition suggests it should also be nonlinear. . .

Suppose $\mathbb{E}(y) = \lambda$ is "small," such as $\lambda = 1$

At this level, suppose the expected effect of $\Delta x = 1$ is a $\beta \uparrow$ in y

If $\mathbb{E}(y)$ is much larger, we might expect the same $\Delta x = 1$ to have a (proportionately) larger effect

Getting from 10 to 100 votes is harder than getting from 10,010 to 10,100.

Poisson Likelihood

The Poisson parameter λ must be positive valued

Intuition suggests it should also be nonlinear. . .

Suppose $\mathbb{E}(y) = \lambda$ is "small," such as $\lambda = 1$

At this level, suppose the expected effect of $\Delta x = 1$ is a $\beta \uparrow$ in y

If $\mathbb{E}(y)$ is much larger, we might expect the same $\Delta x = 1$ to have a (proportionately) larger effect

Getting from 10 to 100 votes is harder than getting from 10,010 to 10,100.

One intuitive choice is $\dfrac{\partial y}{\partial x} = \beta\lambda$

This implies an exponential systematic component, $\lambda = \exp(\mathbf{x}\boldsymbol{\beta})$

In GLM terms, we call this a log link, or a log-linear model: $\log \lambda = \mathbf{x}\boldsymbol{\beta}$

Poisson Likelihood

$$\log \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{N} y_i \log \lambda_i - \lambda_i$$

Substitute the systematic component, $\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$

Poisson Likelihood

$$\log \mathcal{L}(\boldsymbol{\lambda}|\mathbf{y}) = \sum_{i=1}^{N} y_i \log \lambda_i - \lambda_i$$

Substitute the systematic component, $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$

$$\log \mathcal{L}(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} y_i \mathbf{x}_i\boldsymbol{\beta} - \exp(\mathbf{x}_i\boldsymbol{\beta})$$

We could code this into `R` and estimate the Poisson model using `optim()`.

Or we could use `glm(..., family=poisson)`

Aside: What exactly *are* Generalized Linear Models (GLM)?

Generalized Linear Models

Many common probability distributions are special cases of the **exponential family**

If we can rewrite a probability distribution as

$$f(y_i|\mu_i) = h(y_i) \exp\left[A(\mu_i)B(\mu_i) - C(\mu_i)\right]$$

for some functions A, B, and C, then it is a member of the exponential family

Generalized Linear Models

Many common probability distributions are special cases of the **exponential family**

If we can rewrite a probability distribution as

$$f(y_i|\mu_i) = h(y_i) \exp\left[A(\mu_i)B(\mu_i) - C(\mu_i)\right]$$

for some functions A, B, and C, then it is a member of the exponential family

Exponential family distributions include: the Normal, the Poisson, the Binomial, the Multinomial, the Exponential, and the Inverse-Normal, and the Gamma distribution

Generalized Linear Models

Many common probability distributions are special cases of the **exponential family**

If we can rewrite a probability distribution as

$$f(y_i|\mu_i) = h(y_i)\exp\left[A(\mu_i)B(\mu_i) - C(\mu_i)\right]$$

for some functions A, B, and C, then it is a member of the exponential family

Exponential family distributions include: the Normal, the Poisson, the Binomial, the Multinomial, the Exponential, and the Inverse-Normal, and the Gamma distribution

Nelder and Wedderburn showed that you can estimate the MLE for any exponential family member using *iteratively re-weighted least squares*

- quick convergence
- provides "residuals" of a sort

They call this approach to modeling GLM *GLM is a* subset *of MLE*

GLM notation and terminology

Thus far, we have set up our models as:

Stochastic component: $y \sim f(\mu, \alpha)$

Systematic component: $\mu = g(\mathbf{x}\boldsymbol{\beta})$

GLM notation and terminology

Thus far, we have set up our models as:

Stochastic component: $y \sim f(\mu, \alpha)$

Systematic component: $\mu = g(\mathbf{x}\boldsymbol{\beta})$

GLMs use an equivalent but different nomenclature:

Distribution family: $y \sim f(\mu, \alpha)$

Link function: $g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$

MLE systematic components and GLM link functions are inverses of each other

Canonical link: the link(s) which make $\mu = A(\mu)$ above

GLM notation and terminology

Distribution family: $y \sim f(\mu, \alpha)$

GLM notation and terminology

Distribution family: $y \sim f(\mu, \alpha)$

Link function: $g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$

Distribution	Systematic component		Canonical link
Normal	$\mathbb{E}(y) = \mathbf{x}\boldsymbol{\beta}$	Identity	$g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$

GLM notation and terminology

Distribution family: $y \sim f(\mu, \alpha)$

Link function: $g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$

Distribution	Systematic component		Canonical link
Normal	$\mathbb{E}(y) = \mathbf{x}\boldsymbol{\beta}$	Identity	$g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$
Poisson	$\mathbb{E}(y) = \exp(\mathbf{x}\boldsymbol{\beta})$	Log	$g^{-1}(\mu) = \log(\mathbf{x}\boldsymbol{\beta})$

GLM notation and terminology

Distribution family: $y \sim f(\mu, \alpha)$

Link function: $g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$

Distribution	Systematic component		Canonical link
Normal	$\mathbb{E}(y) = \mathbf{x}\boldsymbol{\beta}$	Identity	$g^{-1}(\mu) = \mathbf{x}\boldsymbol{\beta}$
Poisson	$\mathbb{E}(y) = \exp(\mathbf{x}\boldsymbol{\beta})$	Log	$g^{-1}(\mu) = \log(\mathbf{x}\boldsymbol{\beta})$
Binomial	$\mathbb{E}(y) = \frac{1}{1+\exp(-\mathbf{x}\boldsymbol{\beta})}$	Logit	$g^{-1}(\mu) = \log\left(\frac{\mathbf{x}\boldsymbol{\beta}}{1-\mathbf{x}\boldsymbol{\beta}}\right)$

How this nomenclature works

Because the Poisson model is linear in the log of $\mathbf{x}\boldsymbol{\beta}$,
it's an example of a "log-linear" model

Poisson with Unequal Exposure Periods

What if different observations have different period lengths or at-risk populations?

Option (1): use the MLE from HW2

Option (2): use a fixed *offset* in either the Poisson MLE or GLM

For periods of variable length t_i,
using an offset entails adding t_i as an extra covariate with a fixed coefficient of 1

Just use: `glm(..., family=poisson, offset=log(t))`

Then be sure to multiply any fitted values by t_i,
and any counterfactual expected values & CIs by t_{hyp}

Why does this work? *It's equivalent to the variable period Poisson MLE*

$$\begin{aligned} \lambda_i &= t_i \exp(\mathbf{x}_i \boldsymbol{\beta}) \\ \log(\lambda_i) &= \log(t_i) + \mathbf{x}_i \boldsymbol{\beta} \\ \lambda_i &= \exp(\mathbf{x}_i \boldsymbol{\beta} + \log(t_i)) \end{aligned}$$

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\mathbb{E}(y|\mathbf{x}) \quad = \quad \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 (x_2 + \delta)\right)$$

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + \delta)\right) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2 \delta)\end{aligned}$$

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + \delta)\right) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2 \delta)\end{aligned}$$

Now let's consider the factor change in $\mathbb{E}(y)$
("factor change": how many *times* y increases for an increment in x_2)

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + \delta)\right) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2 \delta)\end{aligned}$$

Now let's consider the factor change in $\mathbb{E}(y)$
("factor change": how many *times* y increases for an increment in x_2)

$$\frac{\mathbb{E}(y|x_1, x_2 + \delta)}{\mathbb{E}(y|x_1, x_2)} = \frac{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2 \delta)}{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)}$$

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + \delta)\right) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2\delta)\end{aligned}$$

Now let's consider the factor change in $\mathbb{E}(y)$
("factor change": how many *times* y increases for an increment in x_2)

$$\begin{aligned}\frac{\mathbb{E}(y|x_1, x_2 + \delta)}{\mathbb{E}(y|x_1, x_2)} &= \frac{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2\delta)}{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)} \\ &= \exp(\beta_2\delta)\end{aligned}$$

Interpreting Poisson Coefficients

Consider a Poisson regression of y on x_1 and x_2

Suppose we increase x_2 by δ. What is the change in $\mathbb{E}(y)$?

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2(x_2 + \delta)\right) \\ &= \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2\delta)\end{aligned}$$

Now let's consider the factor change in $\mathbb{E}(y)$
("factor change": how many *times* y increases for an increment in x_2)

$$\begin{aligned}\frac{\mathbb{E}(y|x_1, x_2 + \delta)}{\mathbb{E}(y|x_1, x_2)} &= \frac{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)\exp(\beta_2\delta)}{\exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 x_2)} \\ &= \exp(\beta_2\delta)\end{aligned}$$

For a δ increase in x_2, the expected count increases by $\exp(\beta_2\delta)$ times

For a unit increase in x_2, the expected count increases by $\exp(\beta_2)$ times

A warning about Poisson standard errors

Just like the Binomial regression model from which it's derived,
the Poisson regression model assumes *event independence*

This assumption is violated by correlation across events

In theory, events could be positively or negatively correlated

Positive correlation – "contagion" – leads to overdispersion: $\text{var}(y|\lambda) > \mathbb{E}(y|\lambda)$

Just as with the binomial, when data are overdispersed,
the Poisson model provides (very) overconfident standard errors

A warning about Poisson standard errors

Just like the Binomial regression model from which it's derived,
the Poisson regression model assumes *event independence*

This assumption is violated by correlation across events

In theory, events could be positively or negatively correlated

Positive correlation – "contagion" – leads to overdispersion: $\text{var}(y|\lambda) > \mathbb{E}(y|\lambda)$

Just as with the binomial, when data are overdispersed,
the Poisson model provides (very) overconfident standard errors

If $\text{var}(y|\lambda) < \mathbb{E}(y|\lambda)$, or underdispersion,
we see the opposite—underconfident standard errors

We could call this process "inhibition," the opposite of contagion (very rare)

Contagion, Inhibition, and Independence require different models
Poisson is for independence only

Example: Homeowner Associations

Foreclosure filings of homeowners' associations in Harris County (Houston), TX, during 1995–2001. *Source:* `www.HOAdata.org`

Houston-area HOAs file for foreclosure against members

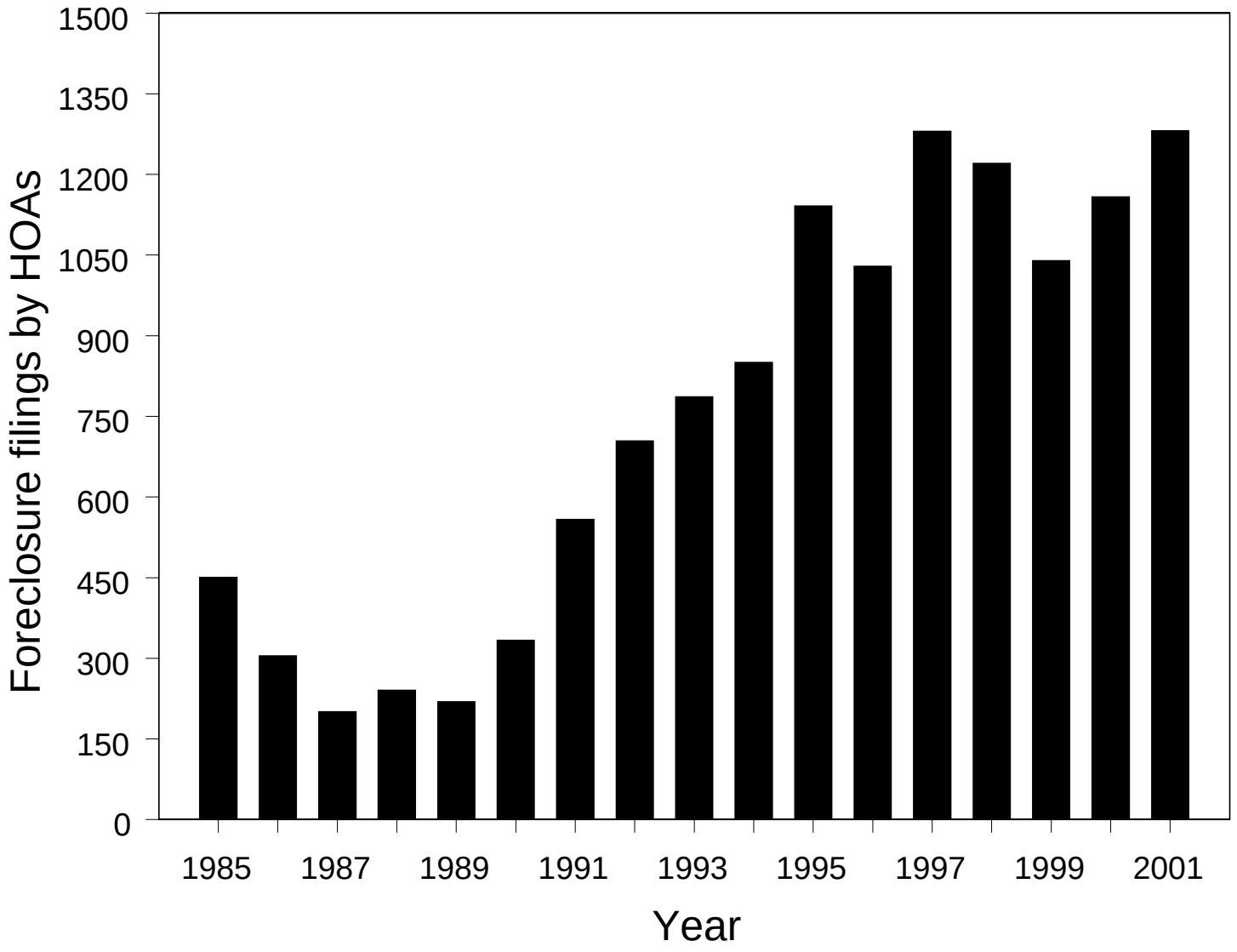
- who are behind on dues
- who have refused to pay fines

Only a small percentage of filings result in actual foreclosure,
but most result in $k's in attorneys' fees

The data consist of the following:

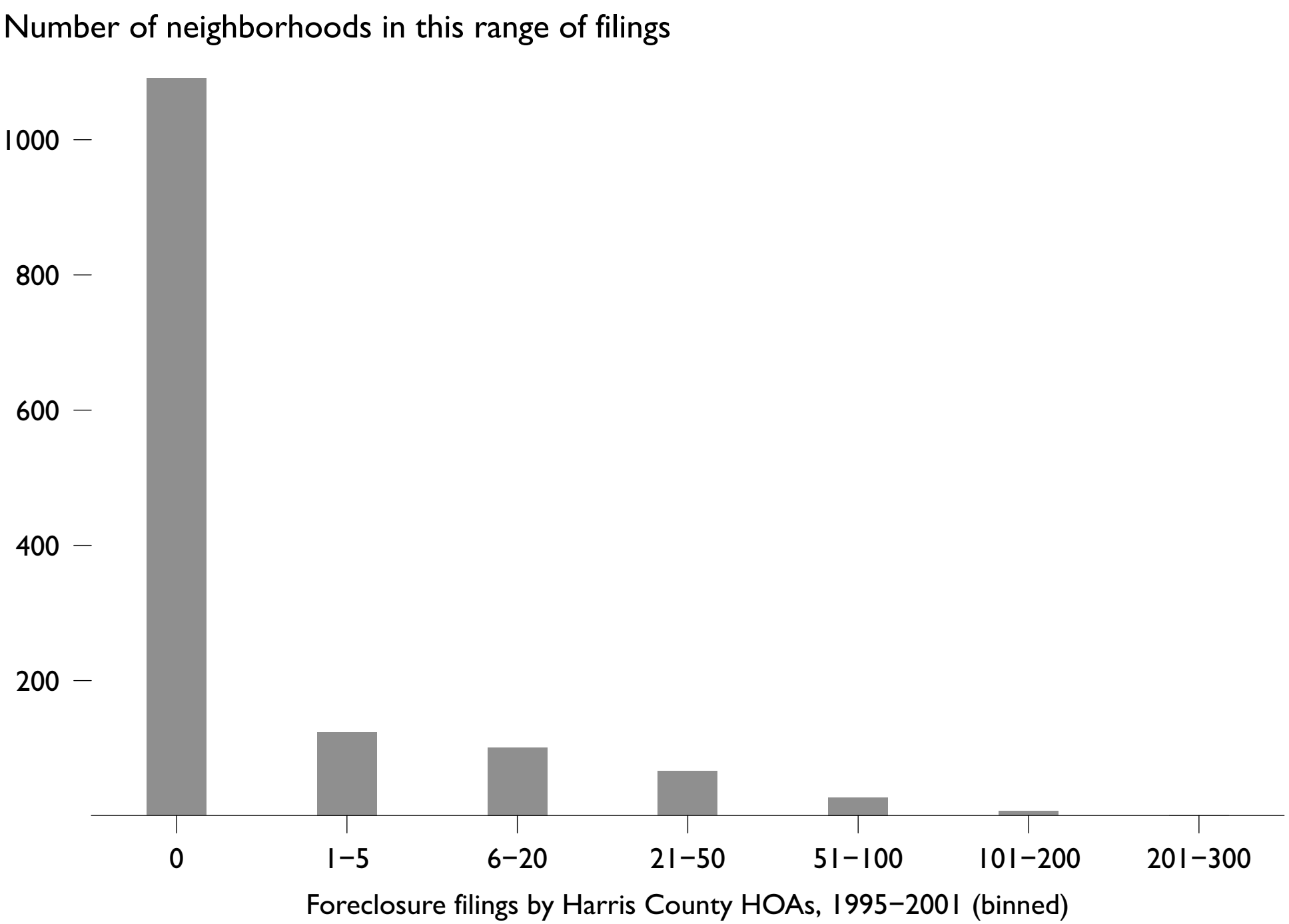
Foreclosure filings	Number of filings in a neighborhood during 1995–2001
Median valuation	Median log home price in the neighborhood
Post-1975 neighborhood	Was the median home built after 1975?

The unit of observation is the neighborhood,
which may or may not have an organized HOA (unobserved)

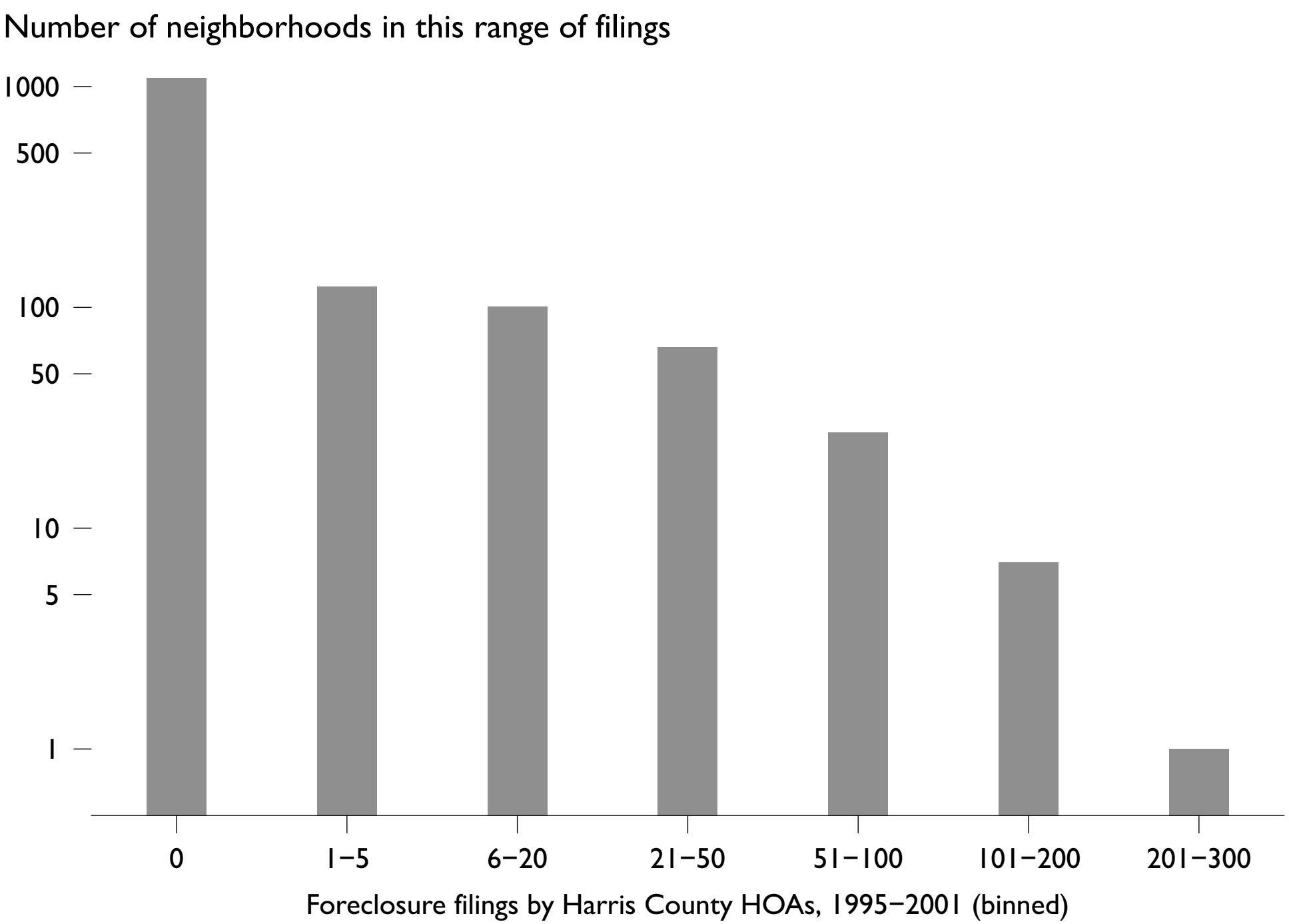


There are surprisingly many filings . . .

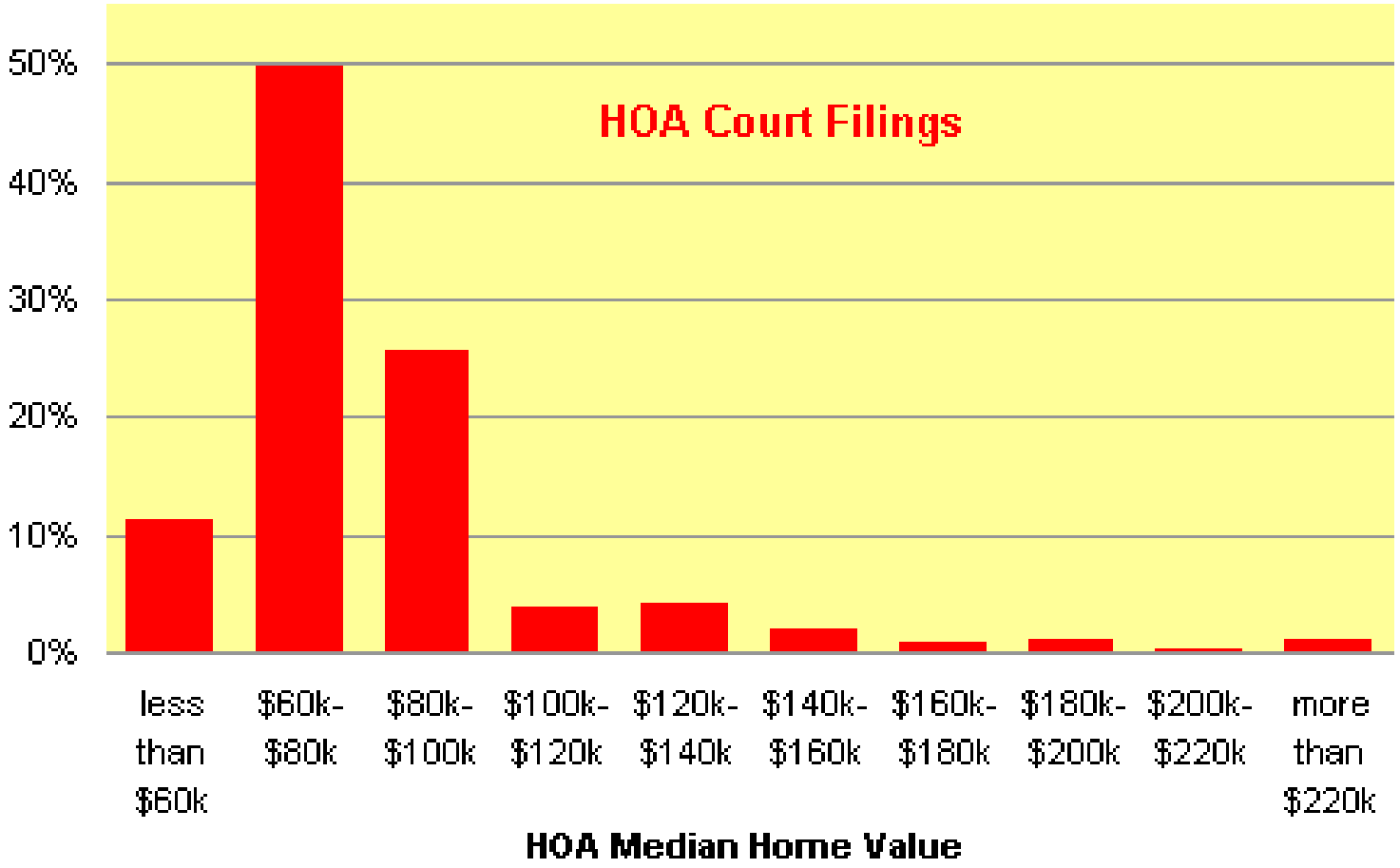
and an essentially steady rate from 1995–2001



A "histogram" of HOA foreclosure filings across neighborhoods



Logging the vertical axis reveals interesting variation overshadowed by the 0s



Omitting the zeros and binning the data

reveals a relationship between home value and filing rates

Source: `www.HOAdata.org`

Example: Homeowner Associations

Let's parameterize the relationship in the last plot using a Poisson probability model

Why? We want to know the uncertainty in this relationship,
and we want to be able to include controls

We fit the following Poisson regression model (note the offset)

$$
\begin{aligned}
\text{forclosure filings}_i &\sim \text{Poisson}(\lambda_i) \\
\log \lambda_i &= \beta_0 + \beta_1 \log(\text{median valuation}_i) \\
&\quad + \beta_2 \text{post-1973 neighborhood}_i \\
&\quad + \log\left(N_{\text{homes}} \times N_{\text{years}}\right)
\end{aligned}
$$

The key variable of interest, home value, may proxy

- income
- access to legal counsel
- ability to pay fees
- potential money collectable in foreclosure auction

	1
log median valuation	-0.90
	(0.03)
Post-1975 neighborhood	2.71
	(0.04)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00
	–
Constant	1.84
	(0.31)
Model	Poisson
N	1417
AIC	15605
In-sample mean absolute error (MAE)	5.28
5-fold cross-validated MAE	5.31

We fit the Poisson model to the full dataset, including the many zeros

How do we interpret these coefficients?

Hint: Are there any tricks for interpreting log-log relationships?

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y}$$

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}}$$

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

Because the above gives different answers depending on the size of Δy and Δx, we usually work with the point elasticity:

$$\eta_{y,x} = \frac{\partial y}{\partial x}\frac{x}{y}$$

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

Because the above gives different answers depending on the size of Δy and Δx, we usually work with the point elasticity:

$$\eta_{y,x} = \frac{\partial y}{\partial x}\frac{x}{y} = \frac{\partial \log y}{\partial \log x}$$

Elasticities

What is an "elasticity," and how can we use the concept to interpret log-linear links?

Elasticity is a commonly used concept in economics.

It is easily understood using an economics analogy,
but is mathematically simple and general

Definition: elasticity of y to x, $\eta_{y,x}$, is the % change in y for a 1% change in x

$$\eta_{y,x} = \frac{\%\Delta x}{\%\Delta y} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

Because the above gives different answers depending on the size of Δy and Δx, we usually work with the point elasticity:

$$\eta_{y,x} = \frac{\partial y}{\partial x}\frac{x}{y} = \frac{\partial \log y}{\partial \log x} \approx \frac{\%\Delta x}{\%\Delta y}$$

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"
- large swings in quantity demanded follow small price changes

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"
- large swings in quantity demanded follow small price changes
- Ex. iPhones, books, movie tickets . . .

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"
- large swings in quantity demanded follow small price changes
- Ex. iPhones, books, movie tickets . . .

If $\eta_{\text{quantity,price}} < 1$

- demand is "inelastic"

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"
- large swings in quantity demanded follow small price changes
- Ex. iPhones, books, movie tickets . . .

If $\eta_{\text{quantity,price}} < 1$

- demand is "inelastic"
- small swings in quantity demanded follow even large price changes

Elasticities

A standard example in economics: the elasticity of demand for a good with respect to price ("price elasticity of demand")

If $\eta_{\text{quantity,price}} > 1$

- demand is "elastic"
- large swings in quantity demanded follow small price changes
- Ex. iPhones, books, movie tickets . . .

If $\eta_{\text{quantity,price}} < 1$

- demand is "inelastic"
- small swings in quantity demanded follow even large price changes
- Ex. grains, cigarettes, rental housing . . .

(NB: this is how economists distinguish "necessities" from "luxury goods")

Elasticities

What does this have to do with statistics?

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities are another, equally valid frame of reference
(in the linear regression case, the elasticity of y wrt x is $\beta x/y$)

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities are another, equally valid frame of reference
(in the linear regression case, the elasticity of y wrt x is $\beta x/y$)

Sometimes, the elasticity is as easy or easier to calculate mentally

Sometimes the elasticity is constant when the slope is not

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities are another, equally valid frame of reference
(in the linear regression case, the elasticity of y wrt x is $\beta x/y$)

Sometimes, the elasticity is as easy or easier to calculate mentally

Sometimes the elasticity is constant when the slope is not

		Slope	Elasticity
Linear model	$y = x\beta$	β	$\beta x/y$

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities are another, equally valid frame of reference
(in the linear regression case, the elasticity of y wrt x is $\beta x/y$)

Sometimes, the elasticity is as easy or easier to calculate mentally

Sometimes the elasticity is constant when the slope is not

		Slope	Elasticity
Linear model	$y = x\beta$	β	$\beta x/y$
Log-linear models	$\log(y) = x\beta$	βy	βx

Elasticities

What does this have to do with statistics?

We commonly understand regression results through the frame of slopes, or partial derivatives

Elasticities are another, equally valid frame of reference
(in the linear regression case, the elasticity of y wrt x is $\beta x/y$)

Sometimes, the elasticity is as easy or easier to calculate mentally

Sometimes the elasticity is constant when the slope is not

		Slope	Elasticity
Linear model	$y = x\beta$	β	$\beta x/y$
Log-linear models	$\log(y) = x\beta$	βy	βx
Log-log models	$\log(y) = \log(x)\beta$	$\beta y/x$	β

If we have logs on both sides, the estimated parameter *is* a point elasticity

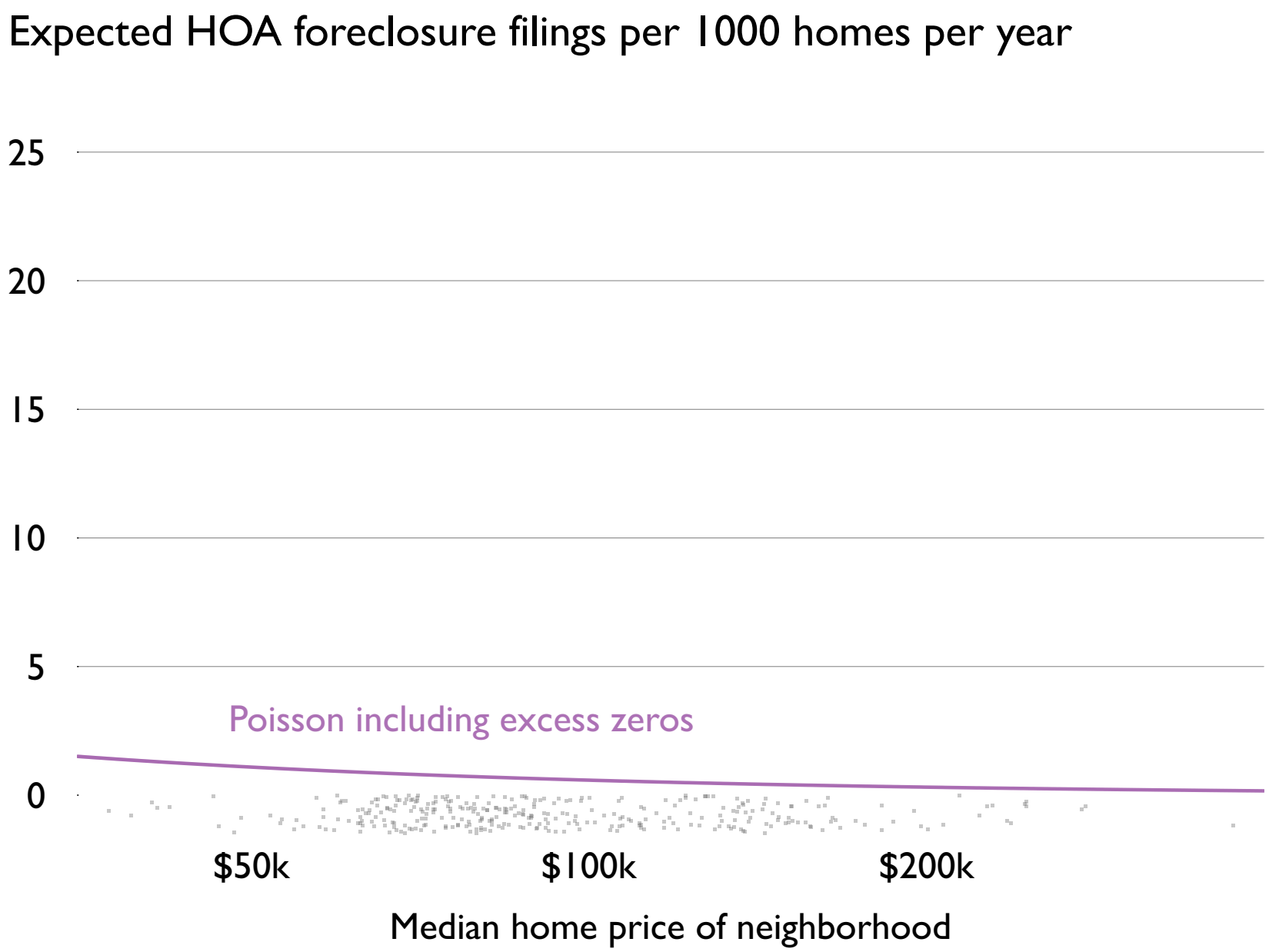
	1
log median valuation	-0.90
	(0.03)
Post-1975 neighborhood	2.71
	(0.04)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00
	–
Constant	1.84
	(0.31)
Model	Poisson
N	1417
AIC	15605
In-sample mean absolute error (MAE)	5.28
5-fold cross-validated MAE	5.31

Validity of inferences depends on the plausibility of our (Poisson) assumptions

Are filings within a neighborhood independent events?

Are all the data even counts? What if some HOAs are structual "non-filers"?

E.g., is the count of cigarettes smoked by a non-smoker really a count?



As usual, a plot of Expected Values is more useful than the table of coefficients

Note the miniscule confidence intervals – precise, tiny effect?

Example: Homeowner Associations

Is this a good model? How could we tell?

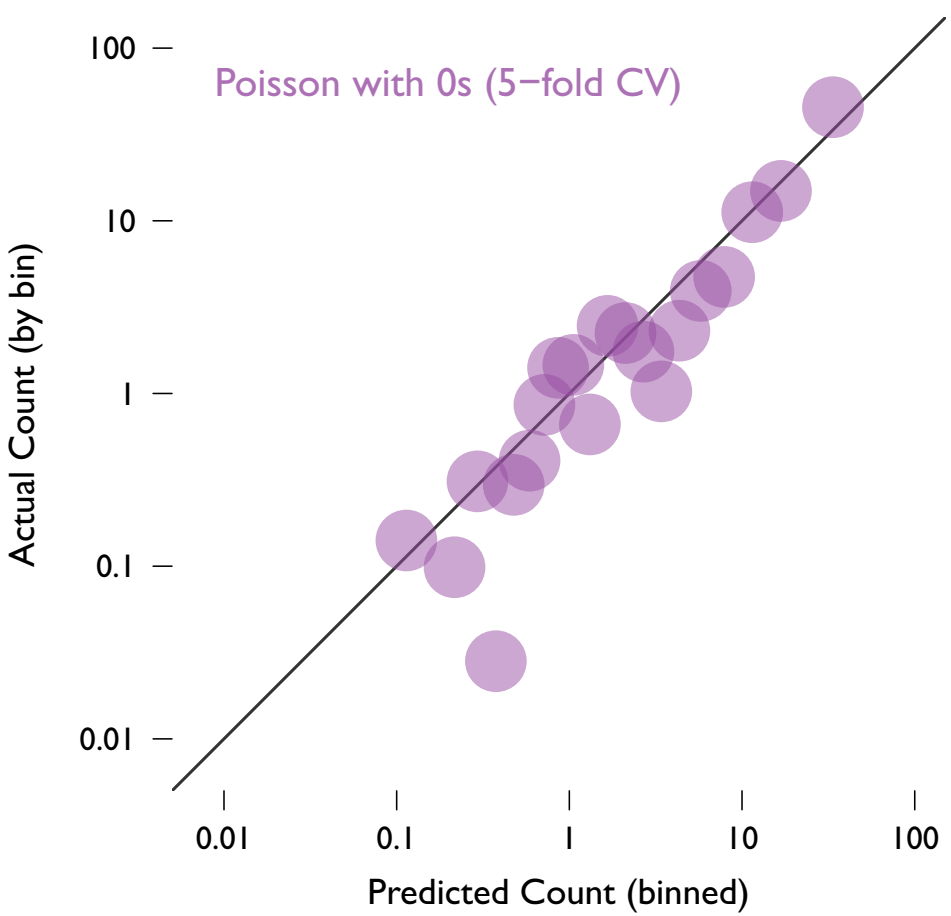
Example: Homeowner Associations

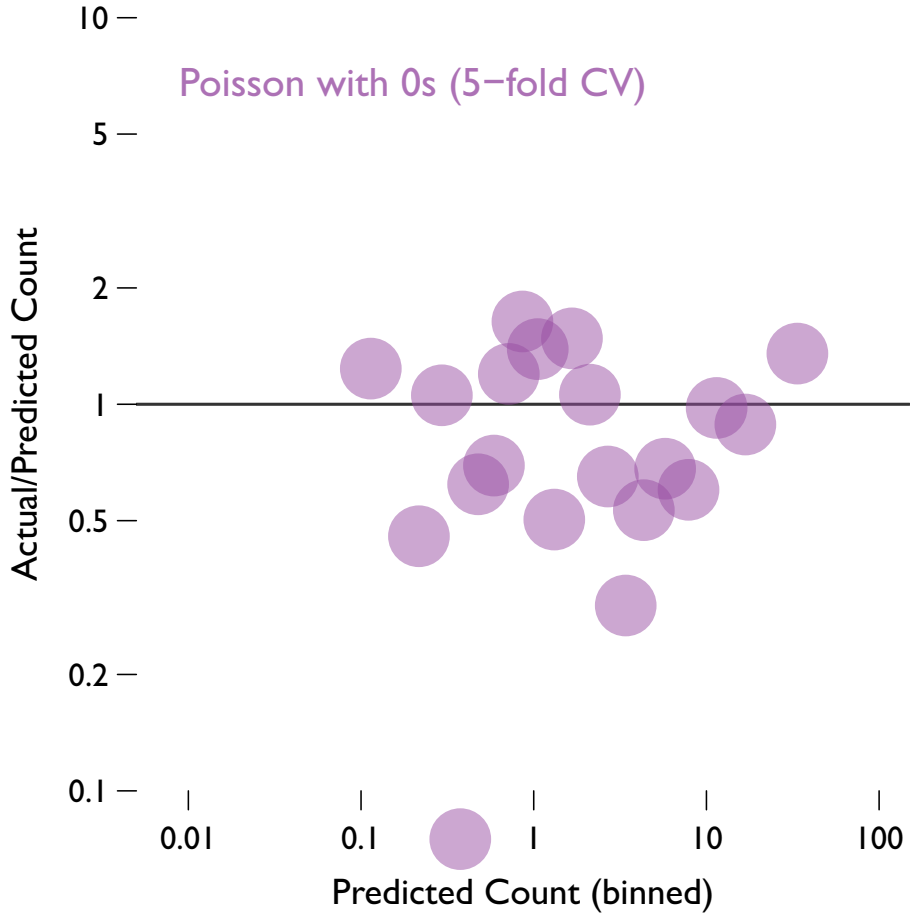
Is this a good model? How could we tell?

Let's look at Actual vs. Predicted plots and residuals. . .

NB: I use the GLM residuals to create a plot of studentized residuals versus standardized leverage

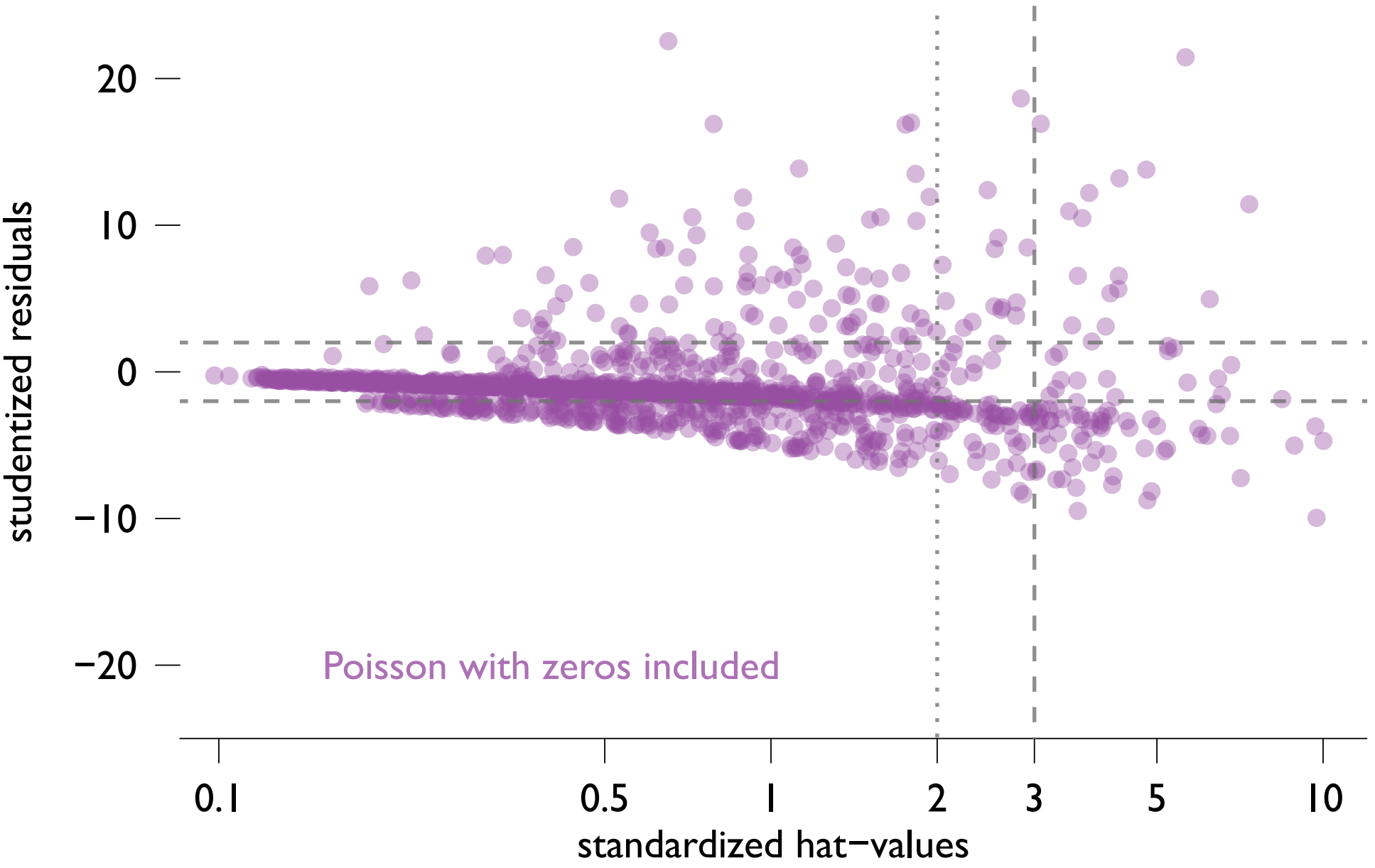
See Topic 6 lecture notes for POLS/CSSS 503:
`faculty.washington.edu/cadolph/503/topic6.p.pdf`





Fit looks surprisingly good

Still, systematically overpredicting filings, especially in the middle range



Yikes! Many high leverage outliers – large hat-values & $> \pm 2$ studentized residuals

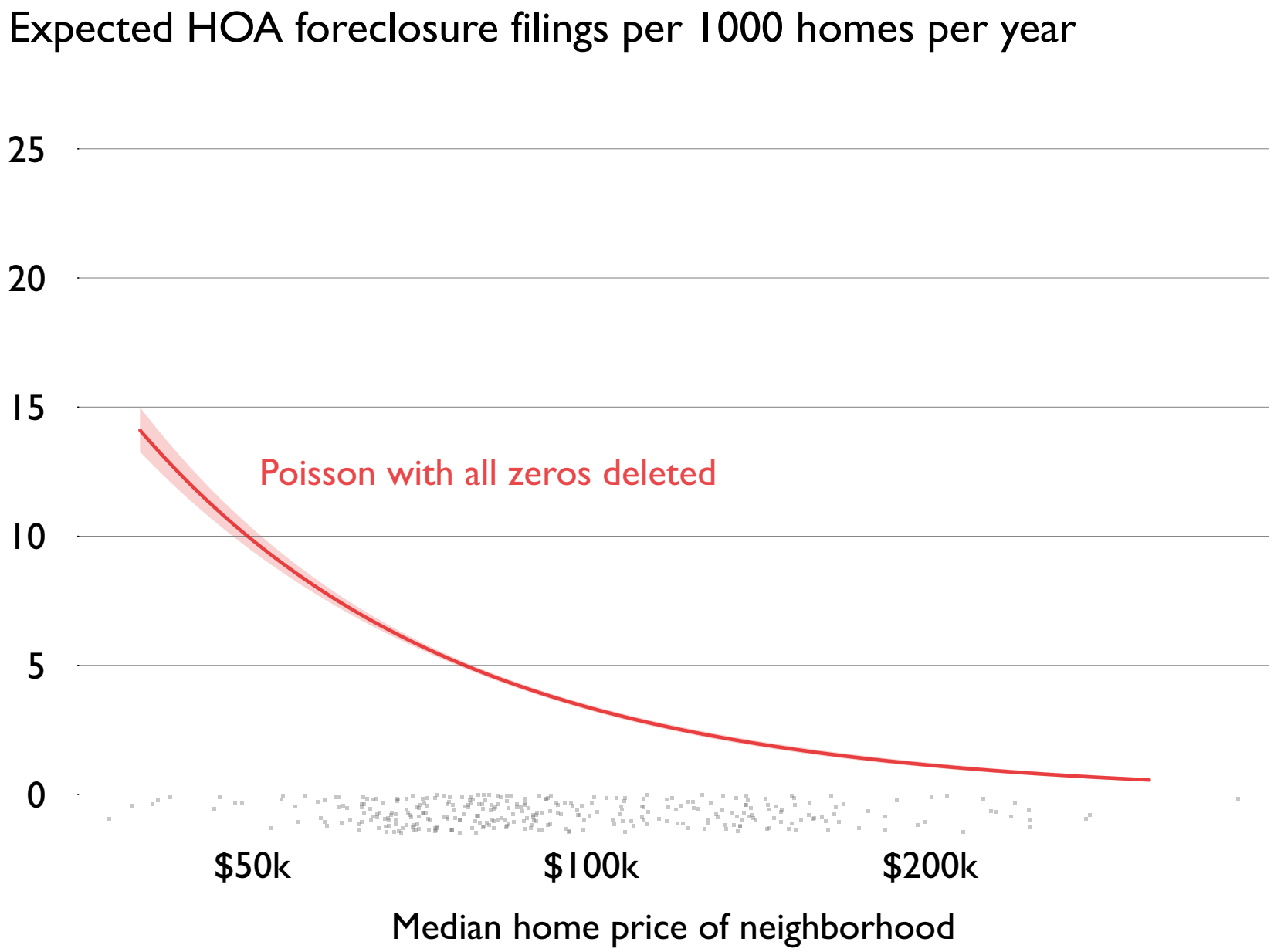
Distributional assumptions of Poisson do not fit the data

	1	2
log median valuation	-0.90	-1.56
	(0.03)	(0.04)
Post-1975 neighborhood	2.71	0.74
	(0.04)	(0.04)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00	1.00
	–	–
Constant	1.84	11.65
	(0.31)	(0.48)
Model	Poisson	Poisson
Exclude zeros?		•
N	1417	326
AIC	15605	6057
In-sample mean absolute error (MAE)	5.28	13.62
5-fold cross-validated MAE	5.31	13.8

Let's do something about the zeros

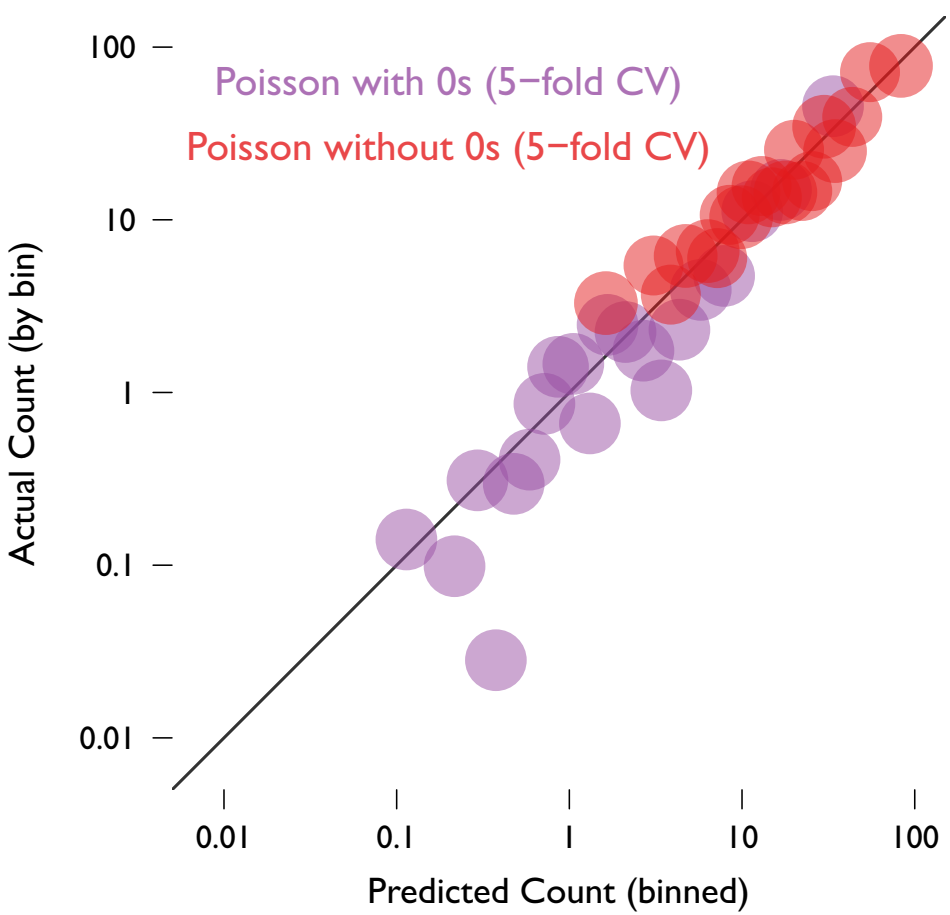
The crudest approach is to discard them; this introduces sampling bias

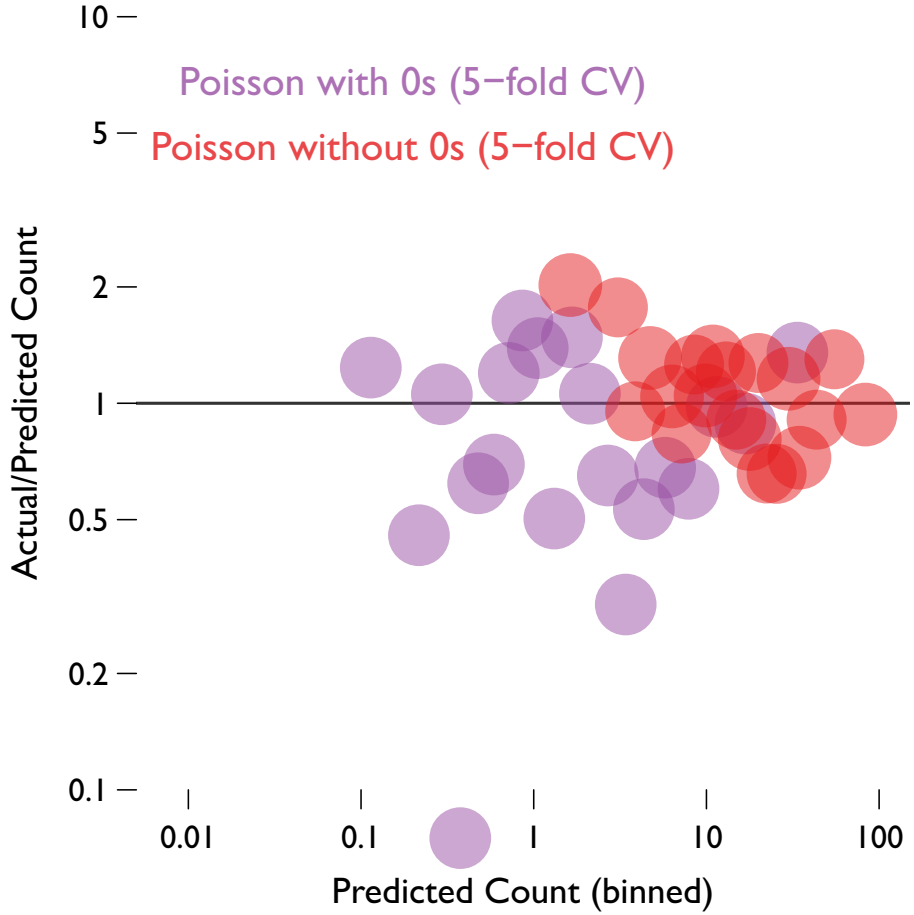
How do we interpret this table?



Excluding all zeros, home prices and foreclosure rates are negatively related

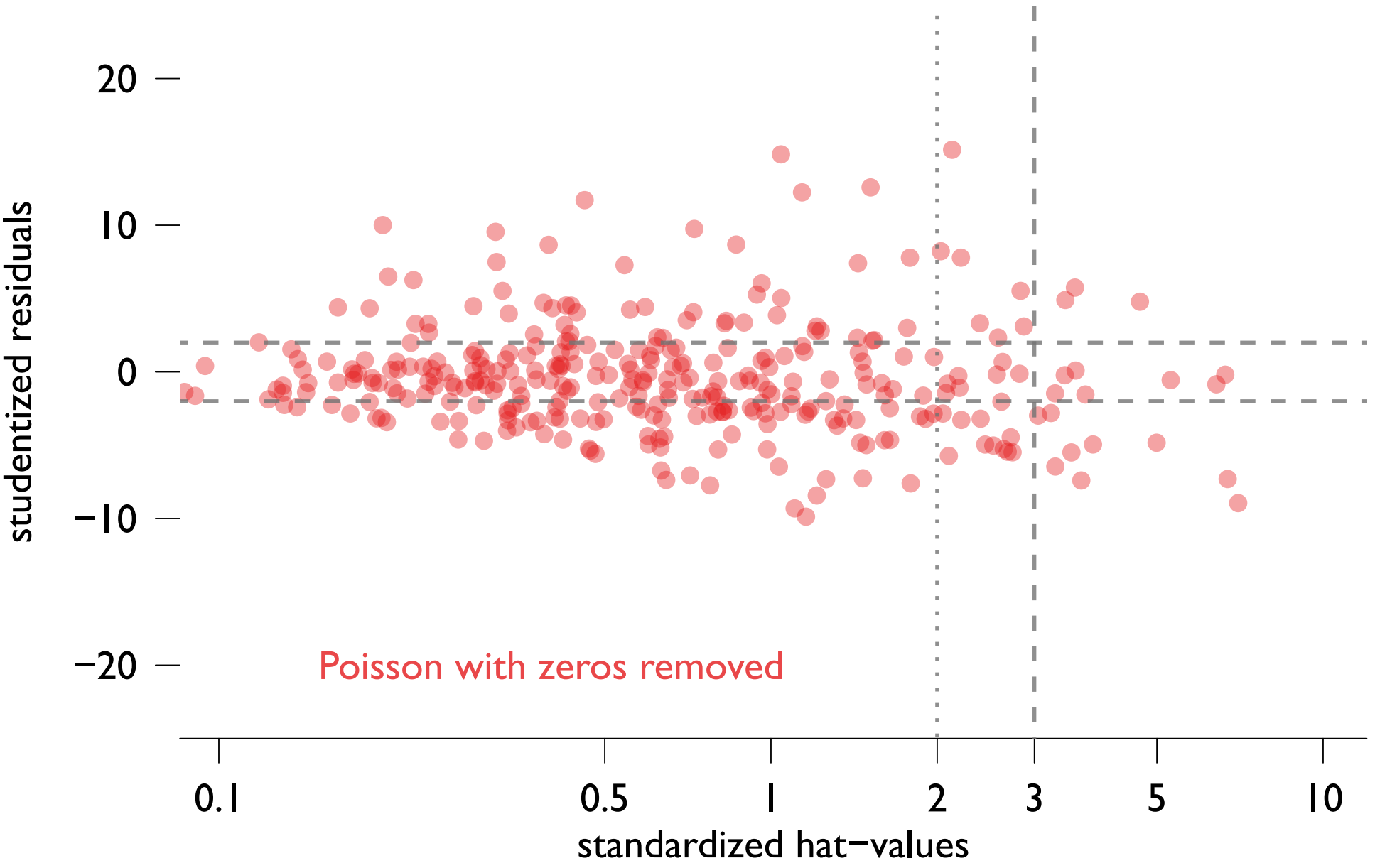
. . . with suspiciously tiny 95% confidence intervals





The fit of our model is significantly improved

at the cost of ignoring much of the data



There are far fewer extreme or high leverage residuals – still too many

Distributional assumptions of Poisson not quite right, even deleting zeros

Example: Homeowner Associations

The t-statistic for Median home value is 37.6.
With 326 observations, this is on the edge of plausibility at best

Example: Homeowner Associations

The t-statistic for Median home value is 37.6.
With 326 observations, this is on the edge of plausibility at best

Should we suspect that $\mathrm{var}(\mathrm{Filings}|\lambda) > \mathbb{E}(\mathrm{Filings}|\lambda)$,
leading to a downward bias in Poisson standard errors?

Example: Homeowner Associations

The t-statistic for Median home value is 37.6.
With 326 observations, this is on the edge of plausibility at best

Should we suspect that $\mathrm{var}(\mathrm{Filings}|\lambda) > \mathbb{E}(\mathrm{Filings}|\lambda)$,
leading to a downward bias in Poisson standard errors?

Based on simple descriptive statistics, you bet

Means and variances of filing count by neighborhood price bracket:

Top of price bracket	225k	175k	125k	75k
Variance of filings	2.5	6.1	27.3	36.9
Mean of filings	1.2	2.4	4.9	6.5
Var/Mean	2.1	2.5	5.6	5.7

The empirical *ratio* of the variance to mean rises as the average price falls.
Poisson probably a poor choice of model.

Once we have a model for overdispersed data,
we'll also have more sophisticated tests for overdispersion

Overdispersion in Unbounded Counts

We want to break the iron link between $\mathbb{E}(y)$ and $\mathrm{var}(y)$ in the Poisson

Overdispersion in Unbounded Counts

We want to break the iron link between $\mathbb{E}(y)$ and $\mathrm{var}(y)$ in the Poisson

One way is to start with the Poisson

$$y_i \sim \mathrm{Poisson}(y_i|\lambda_i)$$

but build in variance by letting λ_i be randomly distributed

Overdispersion in Unbounded Counts

We want to break the iron link between $\mathbb{E}(y)$ and $\mathrm{var}(y)$ in the Poisson

One way is to start with the Poisson

$$y_i \sim \mathrm{Poisson}(y_i|\lambda_i)$$

but build in variance by letting λ_i be randomly distributed

The customary approach is to let λ_i follow the Gamma distribution:

$$\lambda_i \sim \mathrm{Gamma}(\lambda_i|\mu_i, \phi)$$

Overdispersion in Unbounded Counts

We want to break the iron link between $\mathbb{E}(y)$ and $\mathrm{var}(y)$ in the Poisson

One way is to start with the Poisson

$$y_i \sim \mathrm{Poisson}(y_i|\lambda_i)$$

but build in variance by letting λ_i be randomly distributed

The customary approach is to let λ_i follow the Gamma distribution:

$$\lambda_i \sim \mathrm{Gamma}(\lambda_i|\mu_i, \phi)$$

Putting these two distributions together leads to a compound distribution

Overdispersion in Unbounded Counts

We want to break the iron link between $\mathbb{E}(y)$ and $\mathrm{var}(y)$ in the Poisson

One way is to start with the Poisson

$$y_i \sim \mathrm{Poisson}(y_i|\lambda_i)$$

but build in variance by letting λ_i be randomly distributed

The customary approach is to let λ_i follow the Gamma distribution:

$$\lambda_i \sim \mathrm{Gamma}(\lambda_i|\mu_i, \phi)$$

Putting these two distributions together leads to a compound distribution

Before we look at this compound distribution, let's review the Gamma distribution

The Gamma Distribution

Suppose λ_i follows the Gamma distribution

Then the probability of a particular value for λ_i is given by the Gamma pdf, $f_{\mathcal{G}}$:

$$f_{\mathcal{G}}(\lambda_i|\mu_i,\alpha) = \frac{\lambda_i^{\mu_i/\alpha-1}\exp(-\lambda/\alpha)}{\Gamma(\mu_i/\alpha)\alpha^{\mu_i/\alpha}}$$

$$\mathbb{E}(\lambda_i) = \mu_i, \qquad \text{var}(\lambda_i) = \mu_i\alpha$$

Recall that $\Gamma(\cdot)$ is the Gamma *function*, or an interpolated factorial of $x-1$

The Gamma *distribution*, $f_{\mathcal{G}}(\cdot)$ has positive mass on the positive real numbers

As the shape parameter μ increases, the Gamma approximates the Normal

The Gamma Distribution

Suppose λ_i follows the Gamma distribution

Then the probability of a particular value for λ_i is given by the Gamma pdf, $f_{\mathcal{G}}$:

$$f_{\mathcal{G}}(\lambda_i|\mu_i,\alpha) = \frac{\lambda_i^{\mu_i/\alpha-1}\exp(-\lambda/\alpha)}{\Gamma(\mu_i/\alpha)\alpha^{\mu_i/\alpha}}$$

$$\mathbb{E}(\lambda_i) = \mu_i, \qquad \text{var}(\lambda_i) = \mu_i\alpha$$

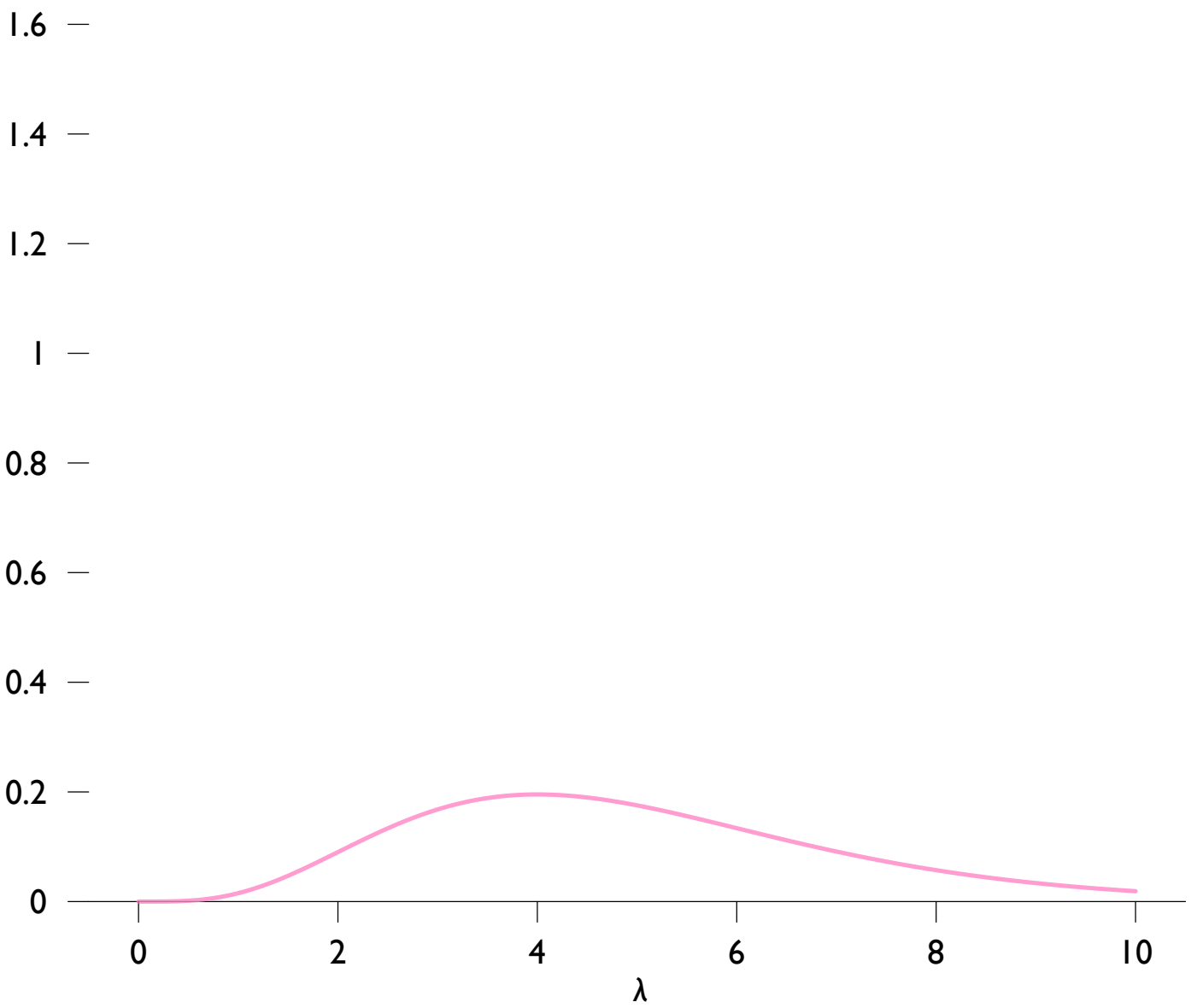
Recall that $\Gamma(\cdot)$ is the Gamma *function*, or an interpolated factorial of $x-1$

The Gamma *distribution*, $f_{\mathcal{G}}(\cdot)$ has positive mass on the positive real numbers

As the shape parameter μ increases, the Gamma approximates the Normal

Gamma is the obvious choice to compound with the Poisson because

- λ must be positive and continuous
- the Gamma and Poisson compound to a distribution with closed form

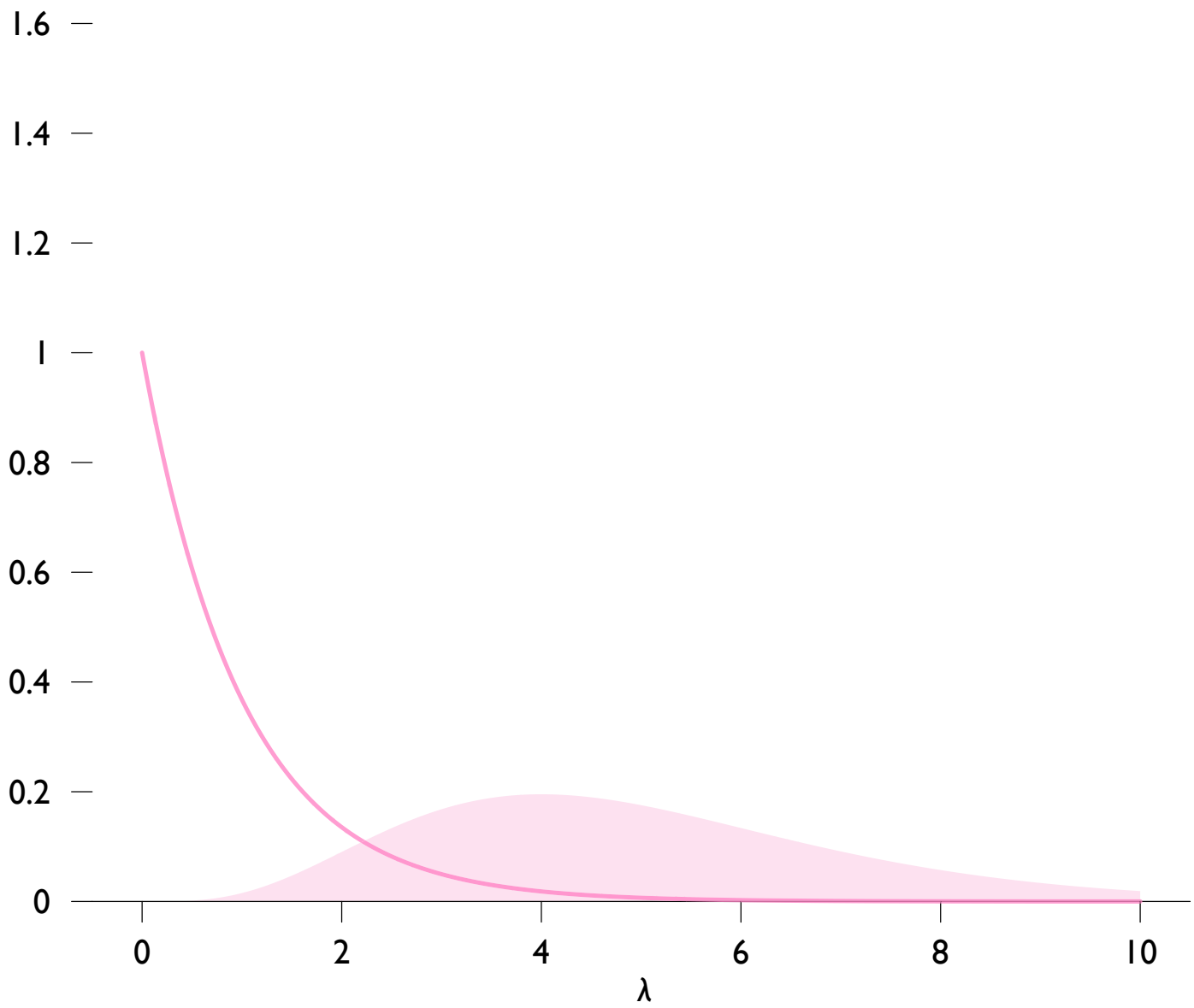


Gamma($\mu = 5$, $a = 1$)

$E(\lambda) = \mu = 5.000$

$\text{var}(\lambda) = \mu a = 5.000$

Let's explore the Gamma distribution, starting with a Gamma($\mu = 5$, $\alpha = 1$):
with a relatively large mean and small variance,
the Gamma looks like the Normal, but a bit asymmetric



Gamma($\mu = 1$, $a = 1$)

$E(\lambda) = \mu = 1.000$

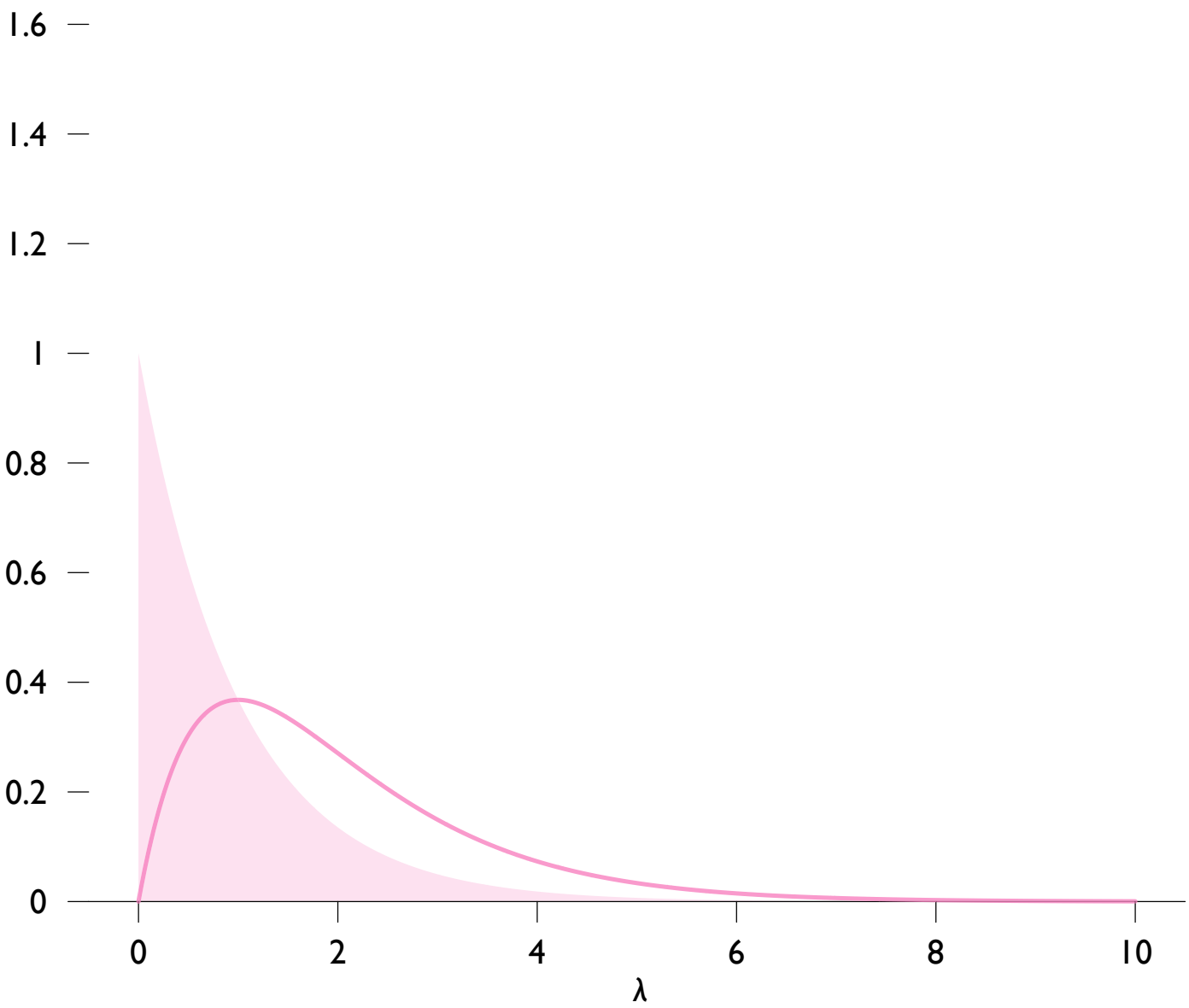
$\text{var}(\lambda) = \mu a = 1.000$

Gamma($\mu = 5$, $a = 1$)

$E(\lambda) = \mu = 5.000$

$\text{var}(\lambda) = \mu a = 5.000$

Lowering μ from 5 to 1 shifts the distribution to the left:
now the Gamma is very asymmetric,
because it has support only over the positive real numbers



Gamma($\mu = 2$, $a = 1$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 2.000$

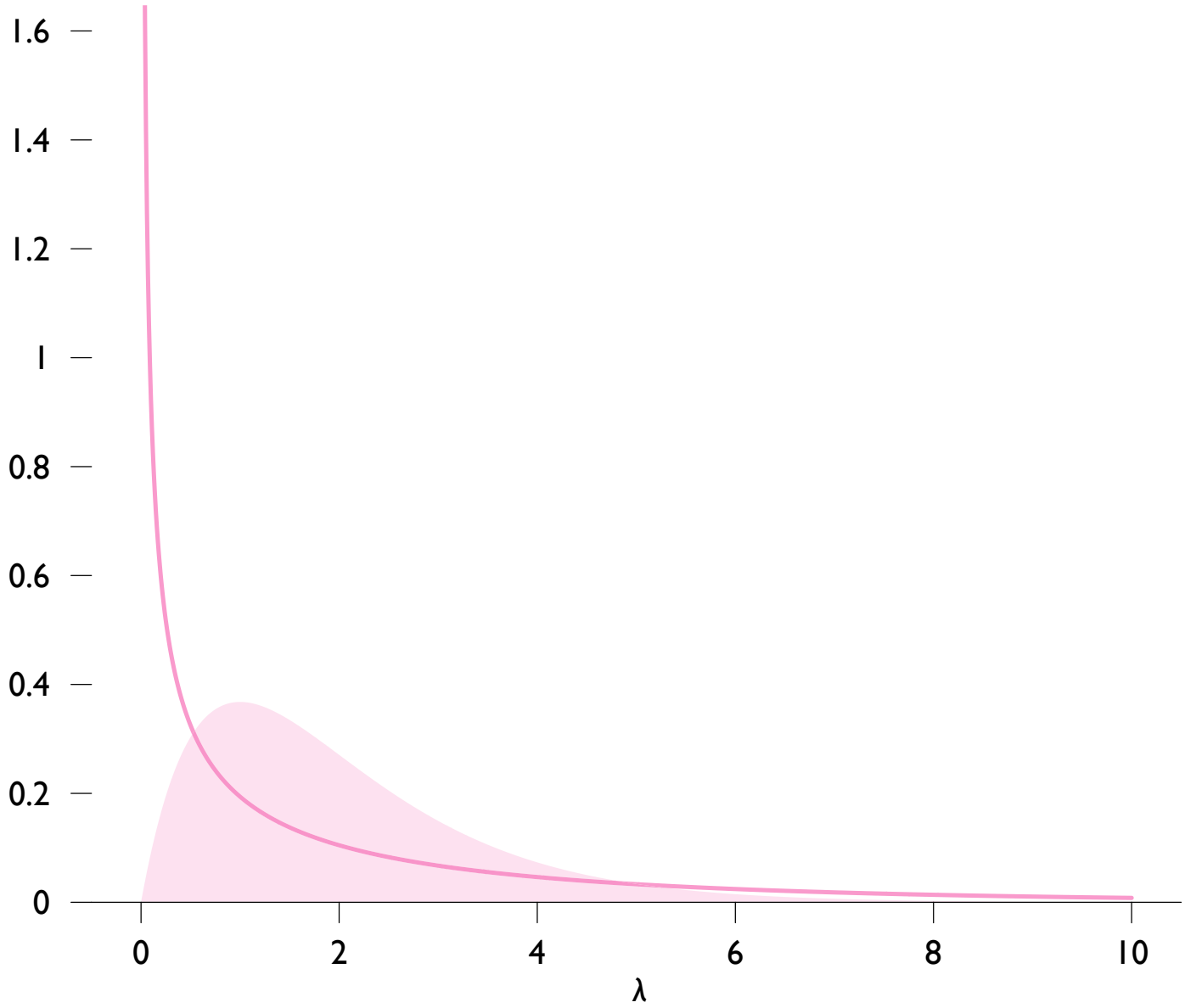
Gamma($\mu = 1$, $a = 1$)

$E(\lambda) = \mu = 1.000$

$\text{var}(\lambda) = \mu a = 1.000$

Let's set μ to 2: some asymmetry, but not too much

What happens as we shift the dispersion parameter α?



Gamma($\mu = 2$, $a = 5$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 10.000$

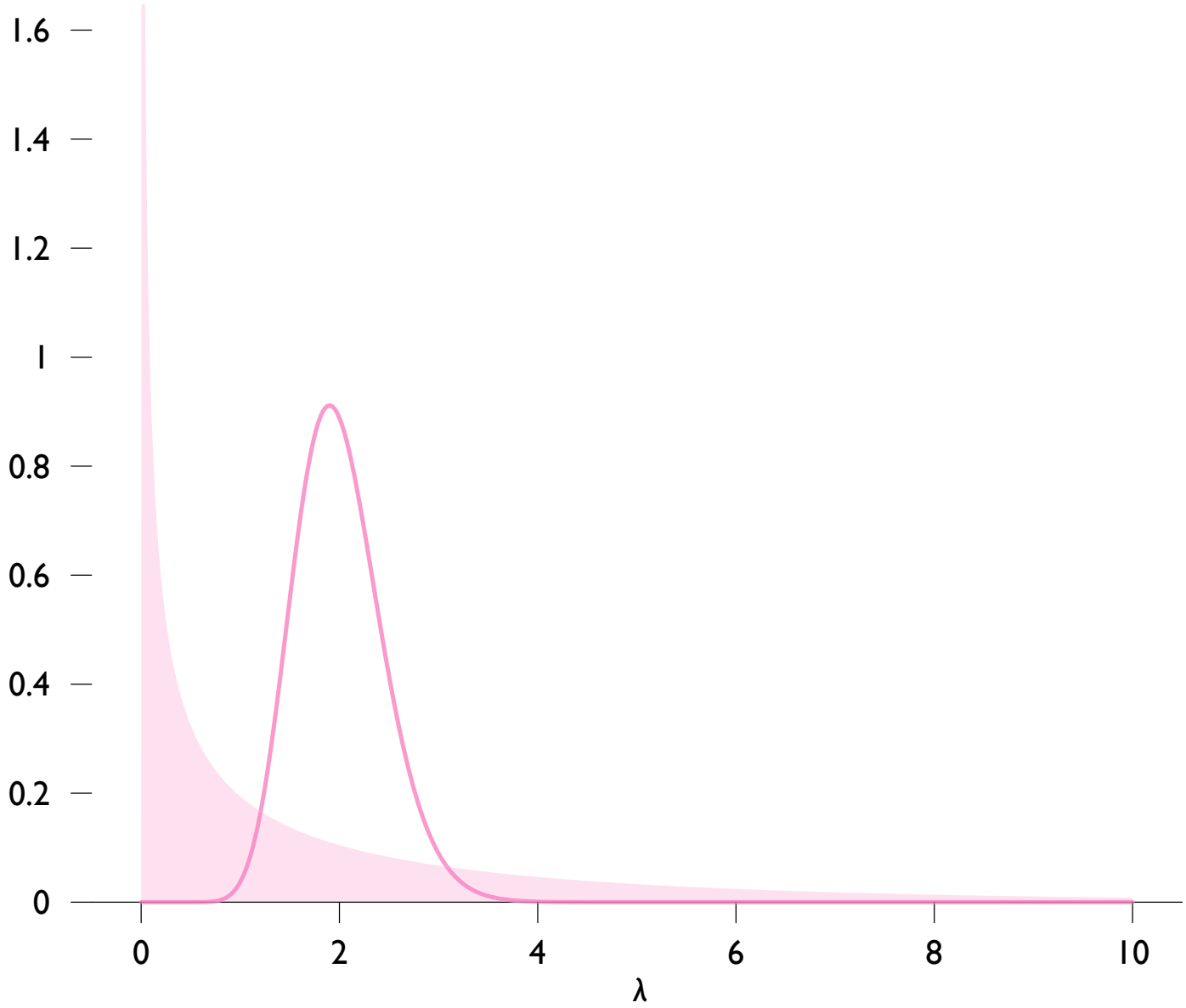
Gamma($\mu = 2$, $a = 1$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 2.000$

Raising α to 5 increases dispersion, but keeps the mean at μ

Because the dispersion hits the zero-bound, the Gamma is more asymmetric



Gamma($\mu = 2$, $a = 0.1$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 0.200$

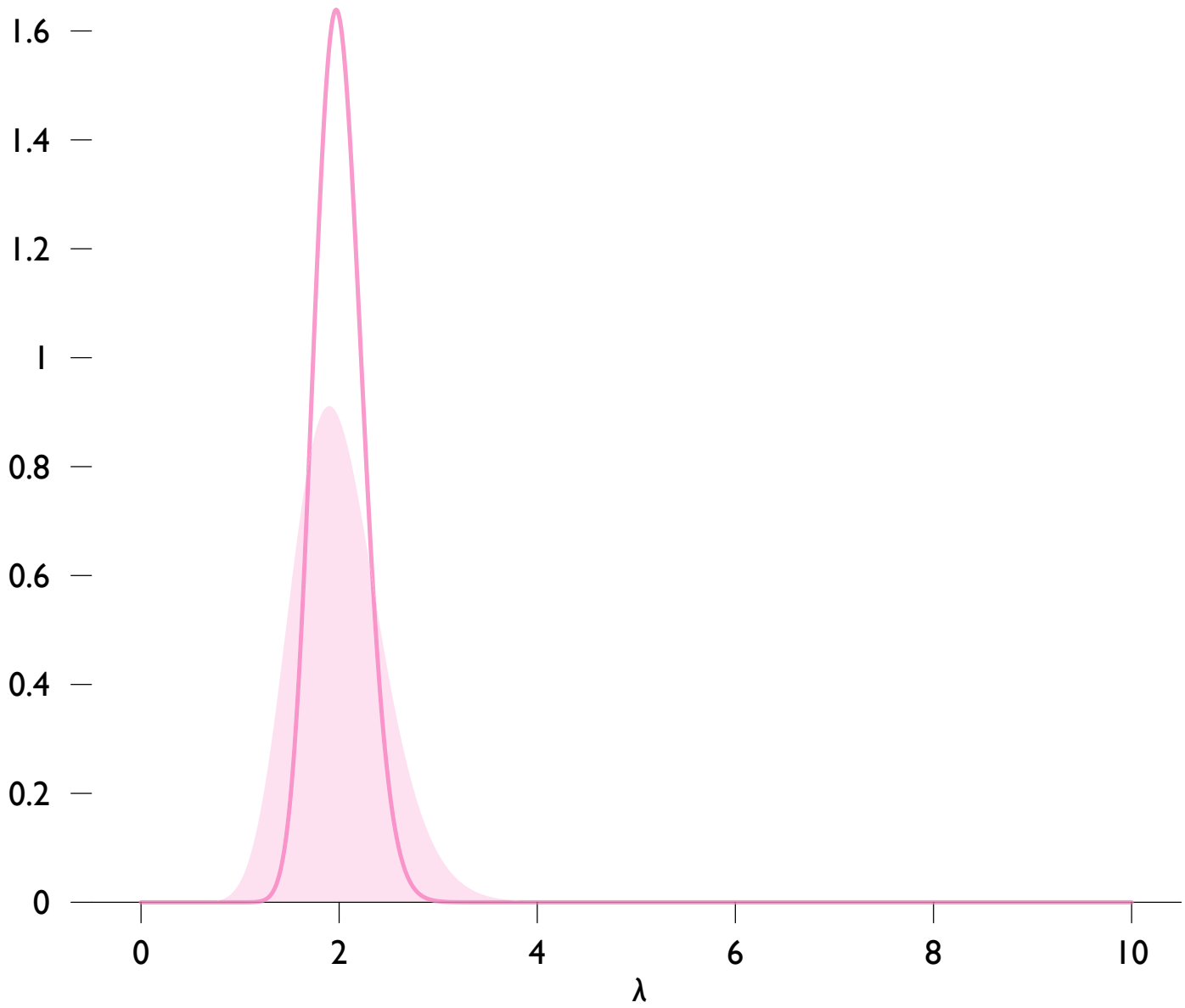
Gamma($\mu = 2$, $a = 5$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 10.000$

Smaller α's, like 0.1, reduce dispersion

Now the Gamma is clustered around the mean & much more symmetric



Gamma($\mu = 2$, $a = 0.03$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 0.060$

Gamma($\mu = 2$, $a = 0.1$)

$E(\lambda) = \mu = 2.000$

$\text{var}(\lambda) = \mu a = 0.200$

As $\alpha \to 0$, the Gamma collapses to a spike at the mean μ

This corresponds to increasing certainty that λ is constant

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i)$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i)$$

Rewrite, to reveal this is a multiplicative term

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta}) \exp(\varepsilon_i)$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i)$$

Rewrite, to reveal this is a multiplicative term

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta})\exp(\varepsilon_i) \qquad = \lambda_i \exp(\varepsilon_i)$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i \boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i)$$

Rewrite, to reveal this is a multiplicative term

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i \boldsymbol{\beta}) \exp(\varepsilon_i) \qquad = \lambda_i \exp(\varepsilon_i) \qquad = \lambda_i \delta_i$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i)$$

Rewrite, to reveal this is a multiplicative term

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta})\exp(\varepsilon_i) \qquad = \lambda_i \exp(\varepsilon_i) \qquad = \lambda_i\delta_i$$

Assume $\mathbb{E}(\varepsilon_i) = 0$, so $\mathbb{E}(\delta_i) = 1$ and

$$\mathbb{E}(y_i) = \lambda_i\mathbb{E}(\delta_i)$$

The Negative Binomial Distribution

To make the Negative Binomial, start with the systematic component of the Poisson

$$\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

Add a random component to account for unexplained variance

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i)$$

Rewrite, to reveal this is a multiplicative term

$$\tilde{\lambda}_i = \exp(\mathbf{x}_i\boldsymbol{\beta})\exp(\varepsilon_i) \qquad = \lambda_i \exp(\varepsilon_i) \qquad = \lambda_i\delta_i$$

Assume $\mathbb{E}(\varepsilon_i) = 0$, so $\mathbb{E}(\delta_i) = 1$ and

$$\mathbb{E}(y_i) = \lambda_i\mathbb{E}(\delta_i) \qquad = \lambda_i$$

as in the Poisson. The variance will be different, as we see below

The Negative Binomial Distribution

Substituting $\tilde{\lambda}_i = \lambda_i \delta_i$ into the Poisson pdf, we get

$$\Pr(y_i|\lambda_i, \delta_i) = \frac{\exp(-\lambda_i \delta_i)(\lambda_i \delta_i)^{y_i}}{y_i!}$$

The Negative Binomial Distribution

Substituting $\tilde{\lambda}_i = \lambda_i \delta_i$ into the Poisson pdf, we get

$$\Pr(y_i|\lambda_i, \delta_i) = \frac{\exp(-\lambda_i \delta_i)(\lambda_i \delta_i)^{y_i}}{y_i!}$$

Next we integrate out δ_i, effectively averaging over the δ's

$$\Pr(y_i|\lambda_i) = \int_0^\infty \Pr(y_i|\lambda_i, \delta_i) g(\delta_i) \mathrm{d}\delta_i$$

The Negative Binomial Distribution

Substituting $\tilde{\lambda}_i = \lambda_i \delta_i$ into the Poisson pdf, we get

$$\Pr(y_i|\lambda_i, \delta_i) = \frac{\exp(-\lambda_i \delta_i)(\lambda_i \delta_i)^{y_i}}{y_i!}$$

Next we integrate out δ_i, effectively averaging over the δ's

$$\Pr(y_i|\lambda_i) = \int_0^\infty \Pr(y_i|\lambda_i, \delta_i) g(\delta_i) \mathrm{d}\delta_i$$

As Long points out, an intuitive way to understand this integral is to imagine $\delta_i \in \{1, 2\}$, in which case we would have

$$\begin{aligned} \Pr(y_i|\lambda_i) &= [\Pr(y_i|\lambda_i, \delta_i = 1) \times g(\delta_i = 1)] \\ &\quad + [\Pr(y_i|\lambda_i, \delta_i = 2) \times g(\delta_i = 2)] \end{aligned}$$

δ_i could take on any positive value, hence the integral above

The Negative Binomial Distribution

Next we need to substitute for $g(\delta_i)$

$g(\cdot)$ is a continuous pdf with mass ("support") on the positive real line only

The Negative Binomial Distribution

Next we need to substitute for $g(\delta_i)$

$g(\cdot)$ is a continuous pdf with mass ("support") on the positive real line only

The most convenient candidate is the Gamma distribution

We make the following substitution

$$\Pr(y_i|\lambda_i, \mu_i, \alpha) = \int_0^\infty \Pr(y_i|\lambda_i, \delta_i)\mathrm{P}(\delta_i|\mu_i, \alpha)\mathrm{d}\delta_i$$

The Negative Binomial Distribution

Next we need to substitute for $g(\delta_i)$

$g(\cdot)$ is a continuous pdf with mass ("support") on the positive real line only

The most convenient candidate is the Gamma distribution

We make the following substitution

$$\Pr(y_i|\lambda_i, \mu_i, \alpha) = \int_0^\infty \Pr(y_i|\lambda_i, \delta_i)\mathrm{P}(\delta_i|\mu_i, \alpha)\mathrm{d}\delta_i$$

A great deal of algebra results in this closed form expression,

$$f_{\mathcal{NB}}(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\mu_i + \alpha^{-1}}\right)^{y_i}$$

known as the Negative Binomial distribution,
which is parameterized by the Gamma parameters, rather than the Poisson

The Negative Binomial Distribution

The Negative Binomial with shape parameter μ and scale parameter α

$$f_{\mathcal{NB}}(y_i|\mu_i,\alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{\alpha^{-1}}{\mu_i+\alpha^{-1}}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\mu_i+\alpha^{-1}}\right)^{y_i}$$

Many sources use an alternative parameterization using μ and rate parameter ϕ, where $\phi = 1/\alpha$

$$f_{\mathcal{NB}}(y_i|\mu_i,\phi) = \frac{\Gamma(y_i+\phi)}{\Gamma(\phi)\Gamma(y_i+1)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(1-\frac{\phi}{\mu_i+\phi}\right)^{y_i}$$

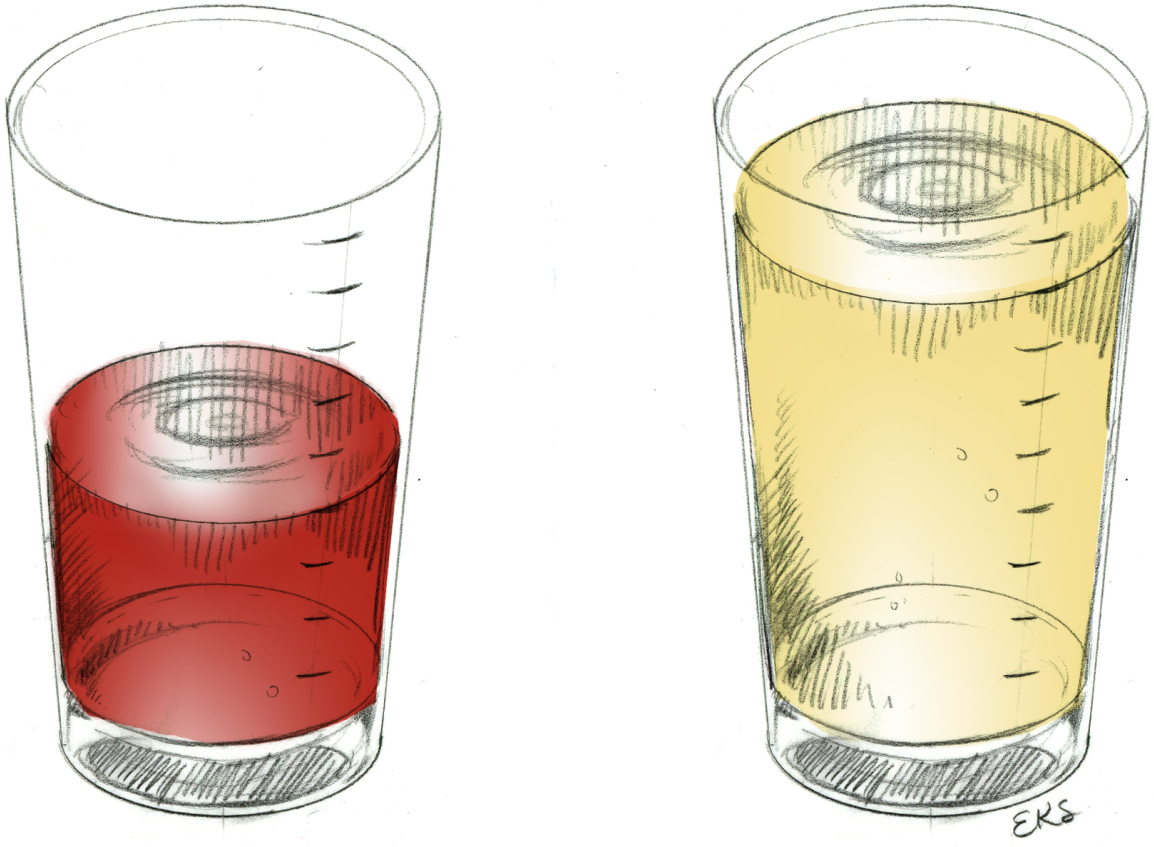
The choice between α and ϕ is arbitrary and will lead to identical conclusions

We choose one parameterization over another for convenience. . .

A quick thought experiment reminds us of the mechanics and pitfalls of reparameterization

Suppose we ask a friend – who is drinking white wine – to pour us a red wine

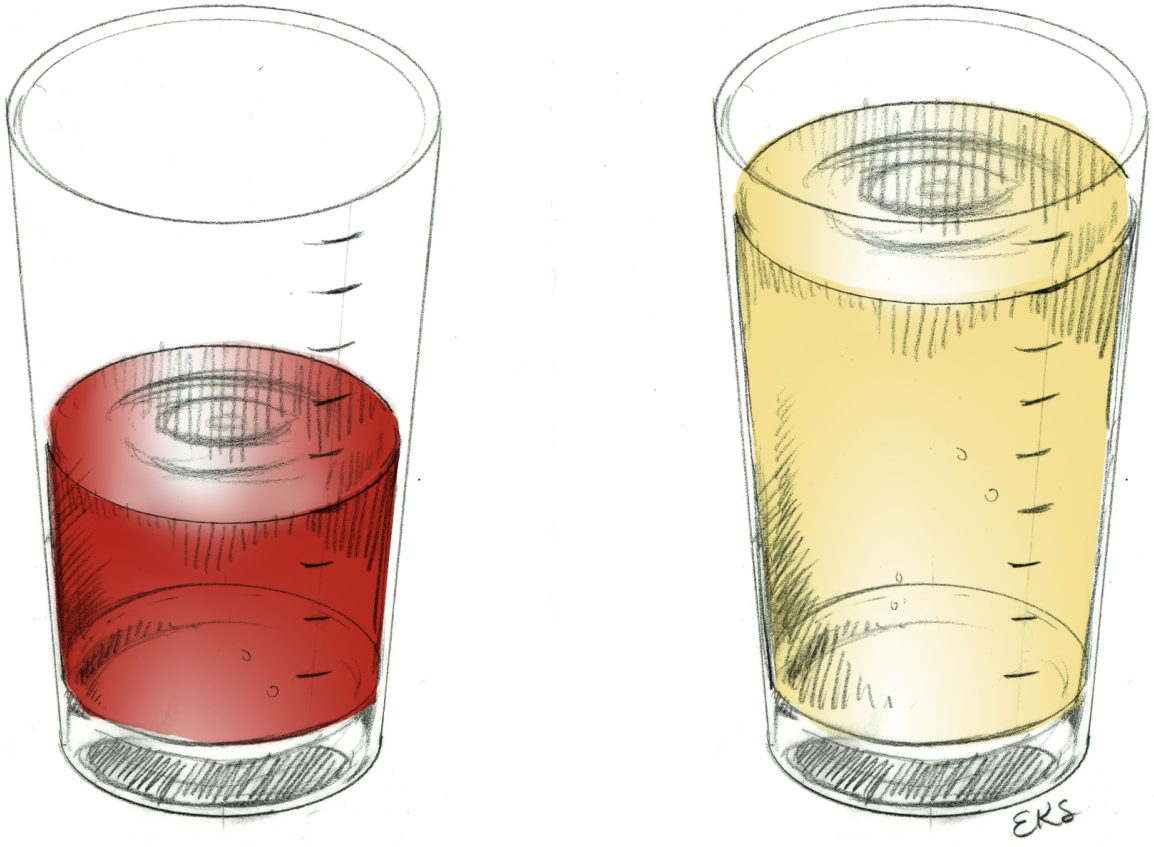
We don't want as much wine as he is drinking



We could say "I'll half as much as you're having" *or* $\alpha = 0.5$

Or we could say "Your glass is double what I want" *or* $\phi = 2$

These are equivalent requests – just as α and ϕ are equivalent parameters



Parameterization choices are usually arbitrary and often opaque:
to figure out what a textbook or stat package means by α or ϕ,
don't count on the letters matching this lecture

Work through the math and confirm the meaning of the parameter in context

The Negative Binomial Distribution

Return to my preferred parameterization of the Negative Binomial, in which $\alpha > 0$ is overdispersion relative to the Poisson:

$$f_{\mathcal{NB}}(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\mu_i + \alpha^{-1}}\right)^{y_i}$$

The Negative Binomial Distribution

Return to my preferred parameterization of the Negative Binomial, in which $\alpha > 0$ is overdispersion relative to the Poisson:

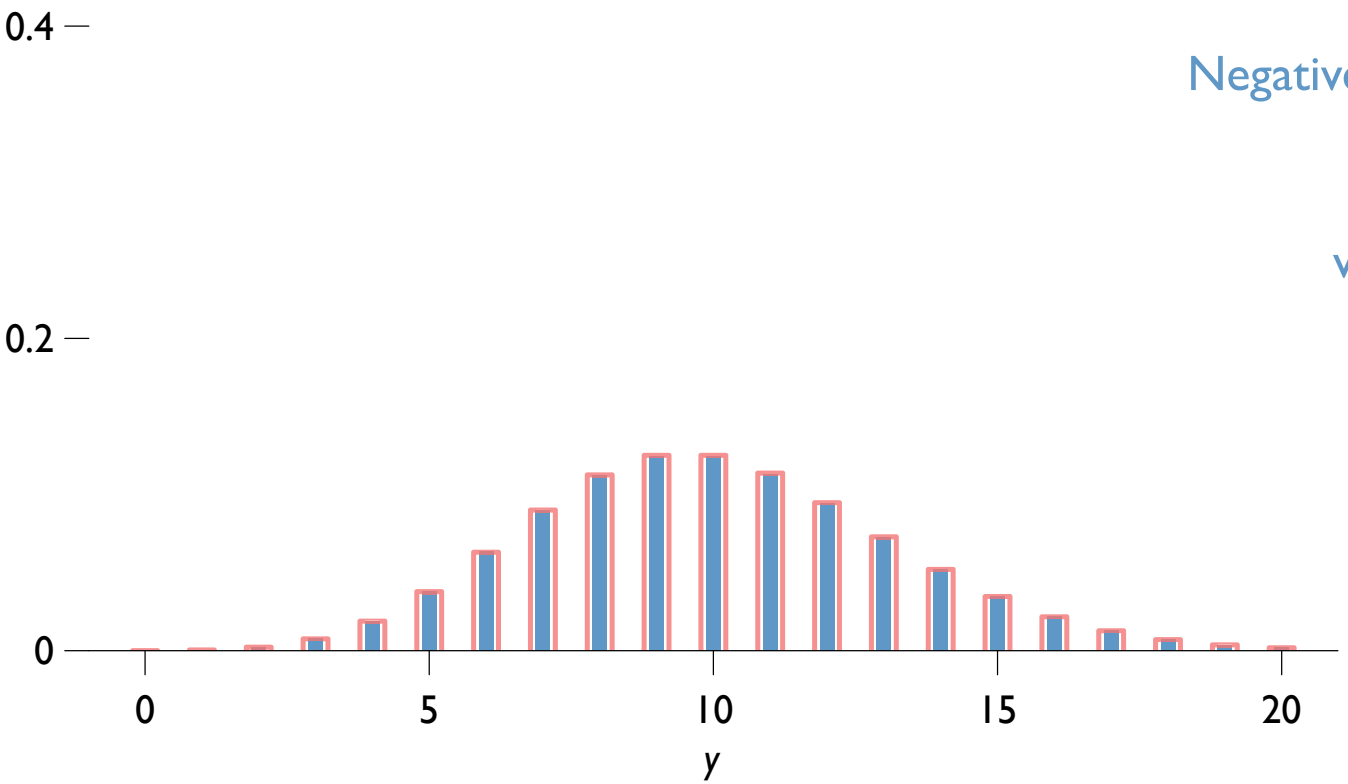
$$f_{\mathcal{NB}}(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\mu_i + \alpha^{-1}}\right)^{y_i}$$

The first two moments are

$$\mathbb{E}(y_i) = \mu_i, \qquad \text{var}(y_i) = \mu_i + \mu_i^2\alpha$$

Higher α means more dispersion (i.e., contagion)

As $\alpha \to 0$, $f_{\mathcal{NB}} \to f_{\mathcal{P}}$ But $\alpha = 0$ exactly is not allowed in the NB



Negative Binomial($\mu = 10$, $a = 0.001$)

$E(y) = \mu = 10$

$\text{var}(y) = \mu + \mu^2 a = 10.1$

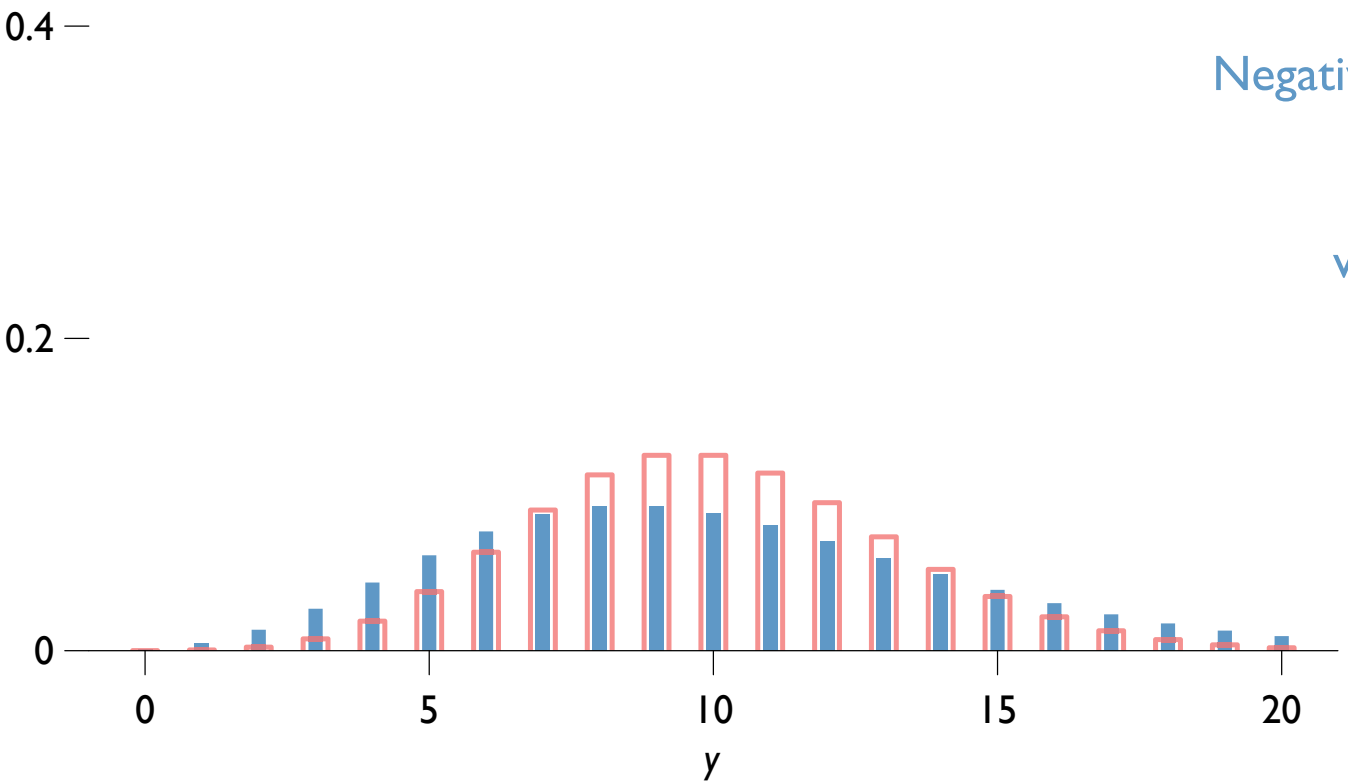
Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$\text{var}(y) = \lambda = 10$

As $\alpha \to 0$, $f_{\mathcal{NB}} \to f_{\mathcal{P}}$

But $\alpha = 0$ exactly is not allowed in the NB



Negative Binomial($\mu = 10$, $a = 0.1$)

$E(y) = \mu = 10$

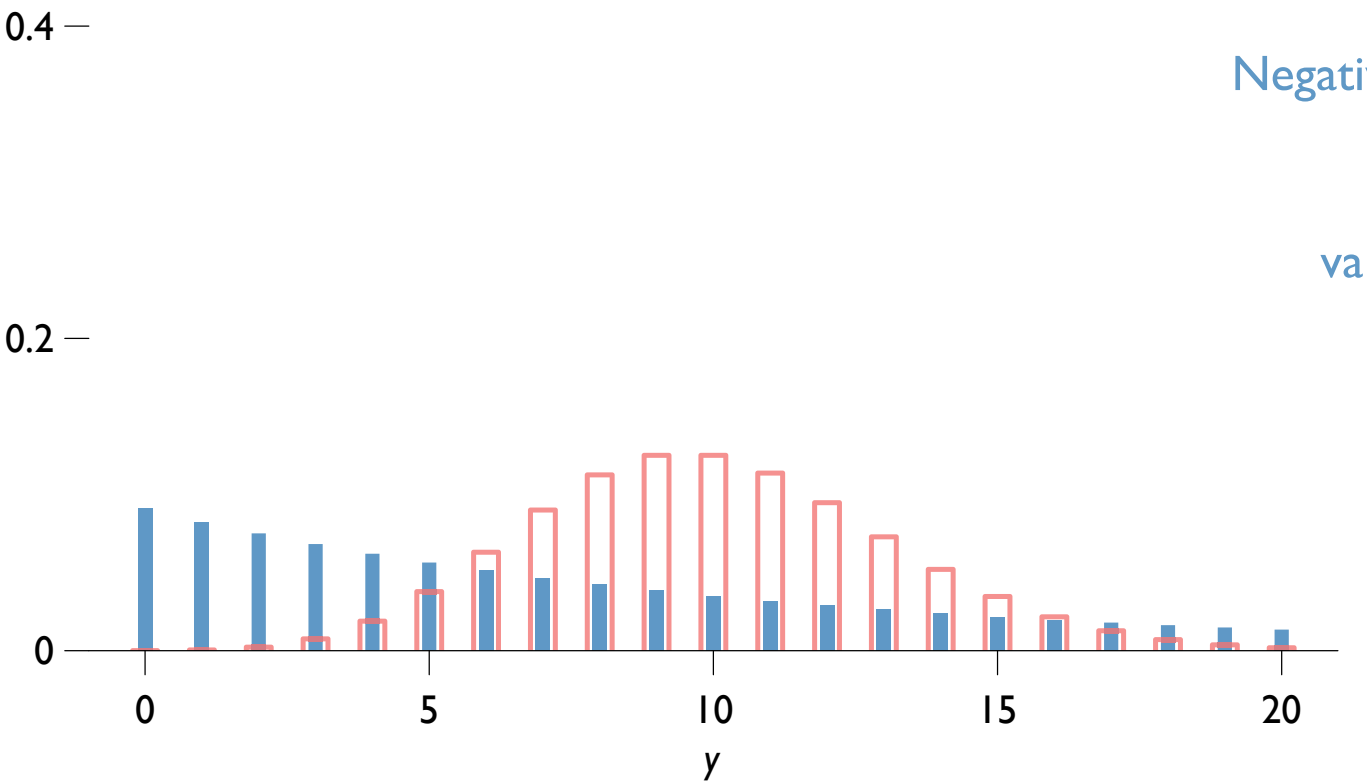
$var(y) = \mu + \mu^2 a = 20.0$

Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$var(y) = \lambda = 10$

Higher α means more dispersion (i.e., contagion)



Negative Binomial($\mu = 10$, $a = 1$)

$E(y) = \mu = 10$

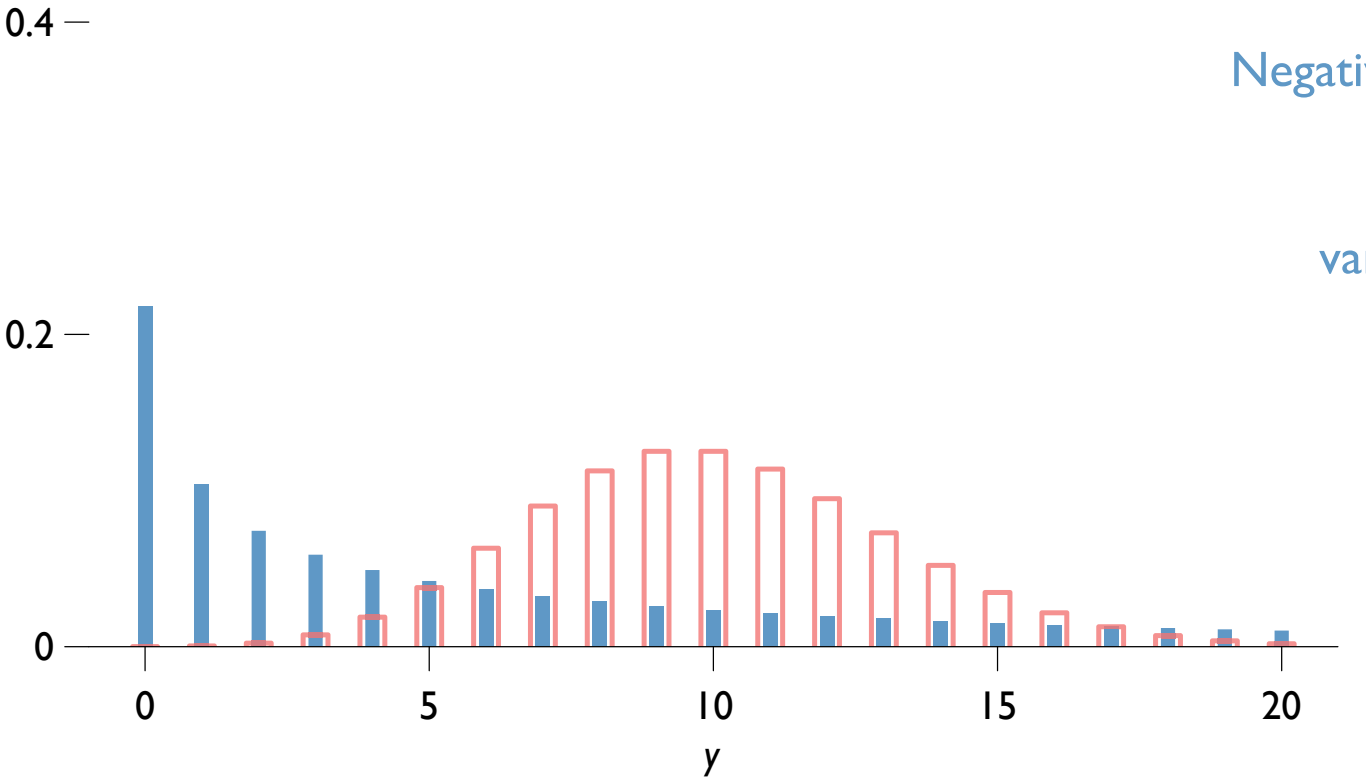
$\text{var}(y) = \mu + \mu^2 a = 110.0$

Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$\text{var}(y) = \lambda = 10$

Does this remind you of another distribution?



Negative Binomial($\mu = 10$, $a = 2$)

$E(y) = \mu = 10$

$\text{var}(y) = \mu + \mu^2 a = 210.0$

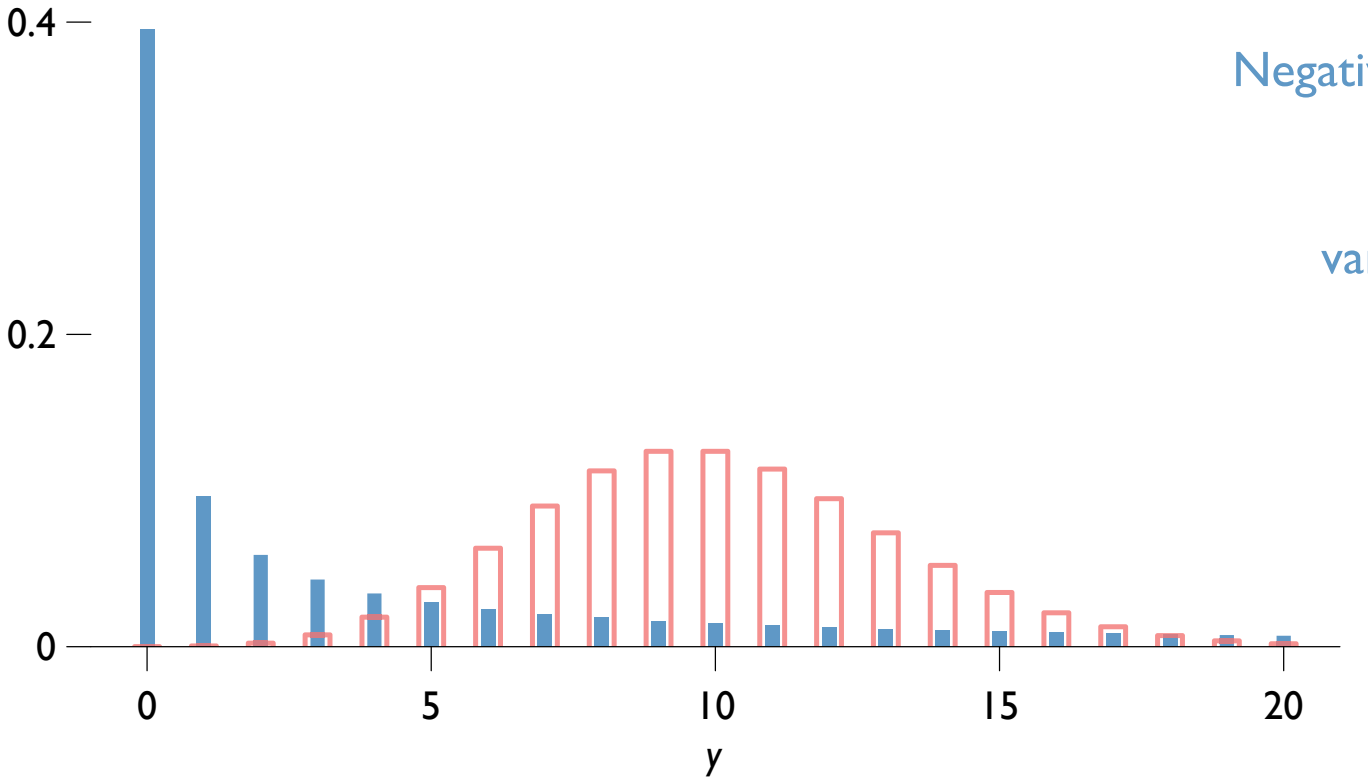
Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$\text{var}(y) = \lambda = 10$

Hopefully, it reminds you of three:

1. The Negative Binomial looks like a discretized Gamma distribution
2. Also similar to the Beta-Binomial, but only bounded on one side
3. Looks a lot like the distribution of filings after removing zeros



Negative Binomial($\mu = 10$, $a = 4$)

$E(y) = \mu = 10$

$\text{var}(y) = \mu + \mu^2 a = 410.0$

Poisson($\lambda = 10$)

$E(y) = \lambda = 10$

$\text{var}(y) = \lambda = 10$

Lots (most?) social science unbounded counts have histograms like this

Negative Binomial: plausible starting point for models of unbounded events

The Negative Binomial Distribution

Trivia: Why is it called the Negative Binomial?

There are *many* ways to derive the NB and it has many uses

Another use for the NB distribution is in quality control:

It is the expected number of successes of a Bernoulli process that occur before a specific number of failures are observed, where

μ is the number of failures to observe

α is the odds of failure in a given trial

The Bernoulli waiting time derivation can lead to a formula that looks like the binomial probability of a "negative" event – hence, "negative" binomial

A shame the Negative Binomial wasn't named the Gamma-Poisson distribution

Sorry you asked?

The Negative Binomial Regression Model

We can use the Negative Binomial as the basis for a regression model

The systematic component is just like the Poisson:

$$\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

But we also have a dispersion parameter to estimate, α

The Negative Binomial Regression Model

We can use the Negative Binomial as the basis for a regression model

The systematic component is just like the Poisson:

$$\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

But we also have a dispersion parameter to estimate, α

Coefficients and expected values will depend on μ only and can be interpreted exactly as in the Poisson

- Coefficients are factor changes in $\mathbb{E}(y)$ for level changes in $\mathbf{x}$
- Expected values are just $\exp(\mathbf{x}_c\boldsymbol{\beta})$ for some hypothetical $\mathbf{x}_c$

The Negative Binomial Regression Model

We can use the Negative Binomial as the basis for a regression model

The systematic component is just like the Poisson:

$$\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

But we also have a dispersion parameter to estimate, α

Coefficients and expected values will depend on μ only and can be interpreted exactly as in the Poisson

- Coefficients are factor changes in $\mathbb{E}(y)$ for level changes in $\mathbf{x}$
- Expected values are just $\exp(\mathbf{x}_c\boldsymbol{\beta})$ for some hypothetical $\mathbf{x}_c$

Predicted values will depend on μ and α

- Draw from the NB distribution (`rnbinom()` in the `stats` library)

The Negative Binomial Likelihood

We form the likelihood in the usual way, starting with the probability density

$$\mathcal{L}(\mu_i, \alpha | y_i, \mathbf{x}_i) = \prod_{i=1}^{N} \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\mu_i + \alpha^{-1}} \right)^{\alpha^{-1}} \left(\frac{\mu_i}{\mu_i + \alpha^{-1}} \right)^{y_i}$$

Next we replace the Γ functions, which will give R a headache, with something more numerically tractable

Cameron and Trivedi note that for integer values of y,

$$\log \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} = \sum_{j=0}^{y_i - 1} \log(j + \alpha^{-1})$$

The Negative Binomial Likelihood

Substituting and reducing to sufficient statistics, we get

$$\log \mathcal{L}(\boldsymbol{\beta}, \alpha | y_i, \mathbf{x}_i) = \sum_{i=1}^{N} \left\{ \left(\sum_{j=0}^{y_i - 1} \log(j + \alpha^{-1}) \right) - (y_i + \alpha^{-1}) \log \left[1 + \alpha \exp(\mathbf{x}_i \boldsymbol{\beta}) \right] + y_i \log \alpha + y_i \mathrm{x}_i \boldsymbol{\beta} \right\}$$

which must be maximized numerically

Negative Binomial Estimation Issues

GLM encompasses the exponential family of distributions.
NB is exponetial only for fixed α,
The usual goal is estimating α, so GLM by itself insufficient

Two-step estimation is available; see `MASS`, `glm.nb()`

Estimation by ML is also relatively easy, using `optim()` as usual

Negative Binomial Estimation Issues

GLM encompasses the exponential family of distributions.
NB is exponetial only for fixed α,
The usual goal is estimating α, so GLM by itself insufficient

Two-step estimation is available; see `MASS`, `glm.nb()`

Estimation by ML is also relatively easy, using `optim()` as usual

Need to be careful to restrict numerical searches to positive α.
Reparameterize: e.g., $\alpha' = \exp(\alpha)$.

Memory constraints prevent `R` from taking the log of $\Gamma(\cdot)$ directly
Usually can find or code an alternative, e.g., `lgamma`, or the substitution above

Negative Binomial Estimation Issues

GLM encompasses the exponential family of distributions.
NB is exponetial only for fixed α,
The usual goal is estimating α, so GLM by itself insufficient

Two-step estimation is available; see `MASS`, `glm.nb()`

Estimation by ML is also relatively easy, using `optim()` as usual

Need to be careful to restrict numerical searches to positive α.
Reparameterize: e.g., $\alpha' = \exp(\alpha)$.

Memory constraints prevent `R` from taking the log of $\Gamma(\cdot)$ directly
Usually can find or code an alternative, e.g., `lgamma`, or the substitution above

Regardless of the method, we can test whether the Poisson would be adequate by rejecting the null that $\alpha = 0$ (or that $\theta = \infty$)

Because these tests involve a null on the boundary of the parameter space, the usual two-sided LR or t-tests are not quite right

Usually, a formal test is unnecessary; otherwise, use a 1-sided test ($0.5 \times p$)

The Quasipoisson

We have multiple strategies for dealing with overdispersion in unbounded counts

The Negative Bionomial builds a probability model of overdispersion from the Gamma distribtion

The Quasipoisson instead allows a Poisson-like relationship – $\mathbb{E}(y_i) = \exp(\mathbf{x}_{\boldsymbol{\beta}})$ – and multiplicatively re-scales the variance up or down as needed

This is a bit more flexible than Negative Binomial, which only scales up – but underdispersion is very rare

Quasipoisson is not a true probability or likelihood model:
only the mean and variance are specified

More robust to misspecification than incorrect probability models;
less efficient than the correct probability model

Scaling of variance is linear, versus quadratic in the Negative Binomial

This difference in assumptions can affect $\hat{\boldsymbol{\beta}}$ and $\mathrm{se}(\hat{\boldsymbol{\beta}})$

HOA Filings Revisited

The HOA filings data appear to be overdispersed

Unobserved characteristics (attorneys, management styles, etc)
may lead to "contagion" within neighborhoods

Seems like a good candidate for the Negative Binomial or Quasipoisson

	1	2	3	4
log median valuation	-0.90	-1.56	-1.56	-1.28
	(0.03)	(0.04)	(0.17)	(0.14)
Post-1975 neighborhood	2.71	0.74	0.74	0.59
	(0.04)	(0.04)	(0.15)	(0.13)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00	1.00	1.00	1.00
	(—)	(—)	(—)	(—)
Constant	1.84	11.65	11.65	8.63
	(0.31)	(0.48)	(1.98)	(1.58)
"theta"				1.38
				(0.11)
Model	Poisson	Poisson	Quasi-Poisson	Negative Binomial
Exclude zeros?		•	•	•
N	1417	326	326	326
AIC	15605	6057	—	2417
In-sample mean absolute error (MAE)	5.28	13.62	13.62	13.89
5-fold cross-validated MAE	5.31	13.80	13.80	14.01

What do these coefficients & se's suggest? *And what is "theta"?*

	1	2	3	4
log median valuation	-0.90	-1.56	-1.56	-1.28
	(0.03)	(0.04)	(0.17)	(0.14)
Post-1975 neighborhood	2.71	0.74	0.74	0.59
	(0.04)	(0.04)	(0.15)	(0.13)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00	1.00	1.00	1.00
	(—)	(—)	(—)	(—)
Constant	1.84	11.65	11.65	8.63
	(0.31)	(0.48)	(1.98)	(1.58)
Dispersion (α)				0.73
				(0.06)
Model	Poisson	Poisson	Quasi-Poisson	Negative Binomial
Exclude zeros?		•	•	•
N	1417	326	326	326
AIC	15605	6057	—	2417
In-sample mean absolute error (MAE)	5.28	13.62	13.62	13.89
5-fold cross-validated MAE	5.31	13.80	13.80	14.01

I reparameterized to α and simulated $\text{se}(\alpha)$ *how did I (know to) do this?*

Interpreting "theta"

The `glm.nb()` function reports a parameter it calls "theta"

From the documentation, it is unclear what "theta" represents

Not safe to assume one person's use of the letter θ corresponds to another's

Most likely, it is either our α or $1/\alpha$

How can we figure out what it is?

Interpreting "theta"

The `glm.nb()` function reports a parameter it calls "theta"

From the documentation, it is unclear what "theta" represents

Not safe to assume one person's use of the letter θ corresponds to another's

Most likely, it is either our α or $1/\alpha$

How can we figure out what it is?

(1) looking around the internet *but what if you can't find anything?*

(2) looking at the glm.nb() code itself *what if it's opaque?*

(3) simulating your own data with known α and run `glm.nb()` on it

Option 3 suggests $\theta = 1/\alpha$

How do we convert $\hat{\theta}$ *and* $\mathrm{se}(\hat{\theta})$ *to* $\hat{\alpha}$ *and* $\mathrm{se}(\hat{\alpha})$*?*

Interpreting "theta"

How do we convert $\hat{\theta}$ *and* $\text{se}(\hat{\theta})$ *to* $\hat{\alpha}$ *and* $\text{se}(\hat{\alpha})$*?*

Although you can invert $\hat{\theta}$ to make $\hat{\alpha} = 1/\hat{\theta}$, you can't simply invert the se's

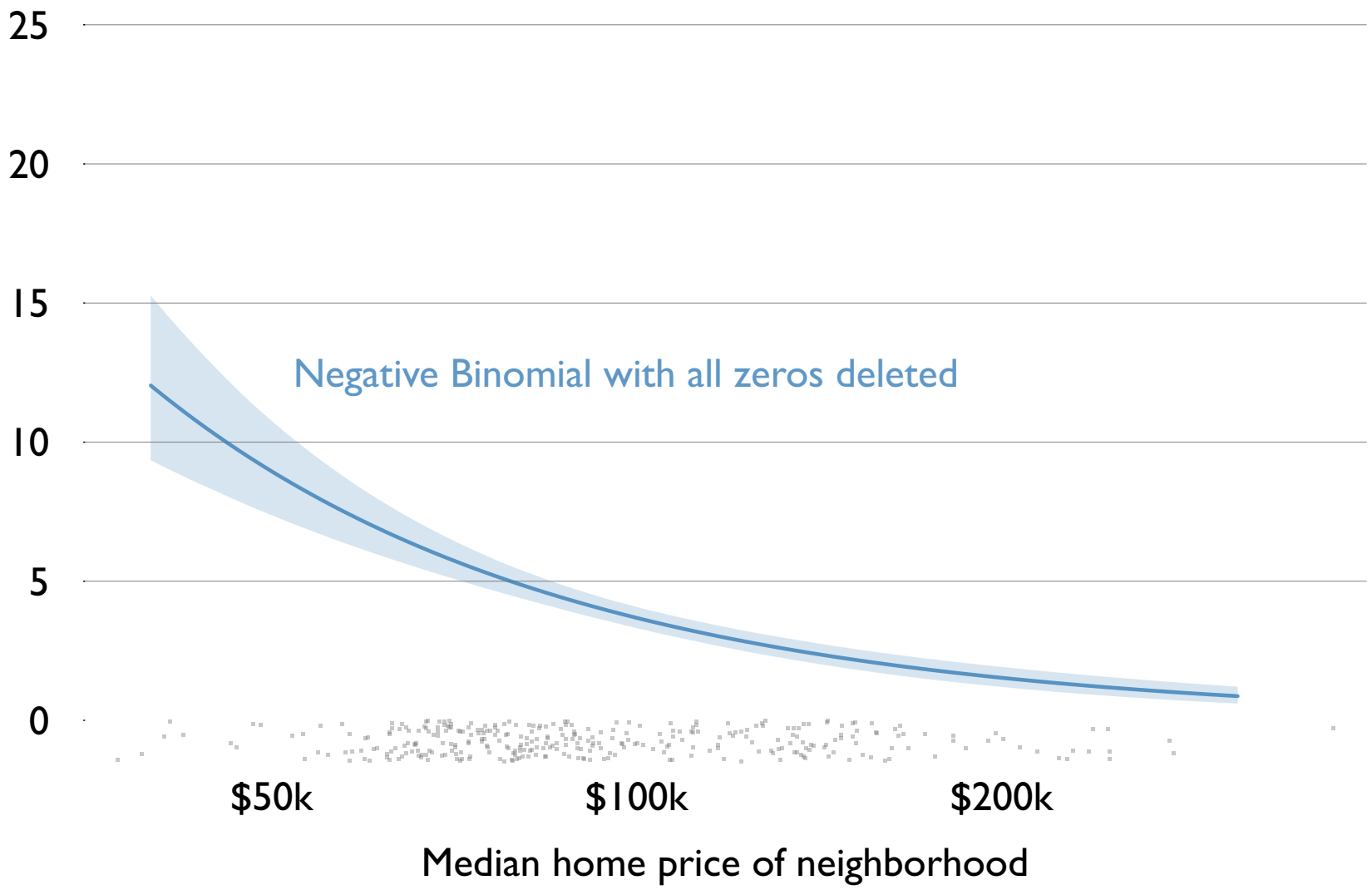
Simulation is very helpful for reparameterizing estimated models:

1. Draw 10,000 simulates from $\tilde{\boldsymbol{\theta}} \sim \text{Normal}(\hat{\theta}, \text{se}(\hat{\theta}))$
2. Compute $\tilde{\boldsymbol{\alpha}} = 1/\tilde{\boldsymbol{\theta}}$
3. Summarize using $\hat{\alpha} = \text{mean}(\tilde{\boldsymbol{\alpha}})$ and $\text{se}(\hat{\alpha}) = \text{sd}(\tilde{\boldsymbol{\alpha}})$

Instead of wondering whether "theta"=1.38 (se=0.11) is "close" to ∞

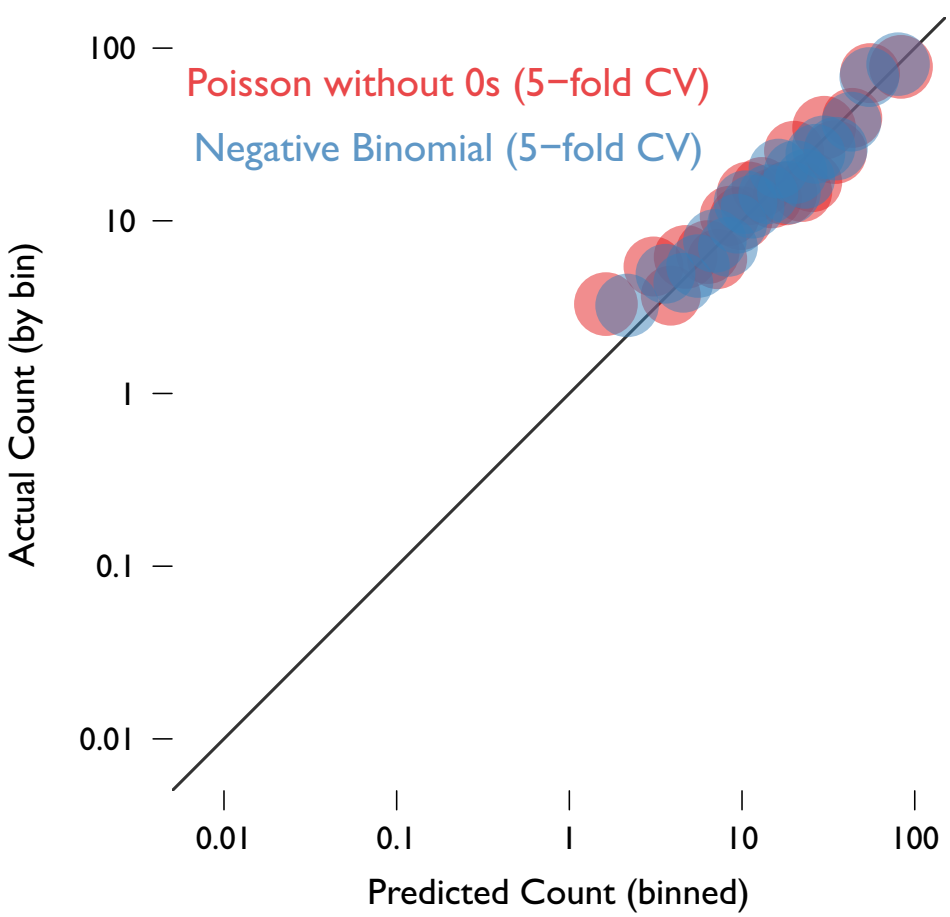
We test whether $\hat{\alpha} = 0.73$ (se=0.06) is close to 0 (obviously not)

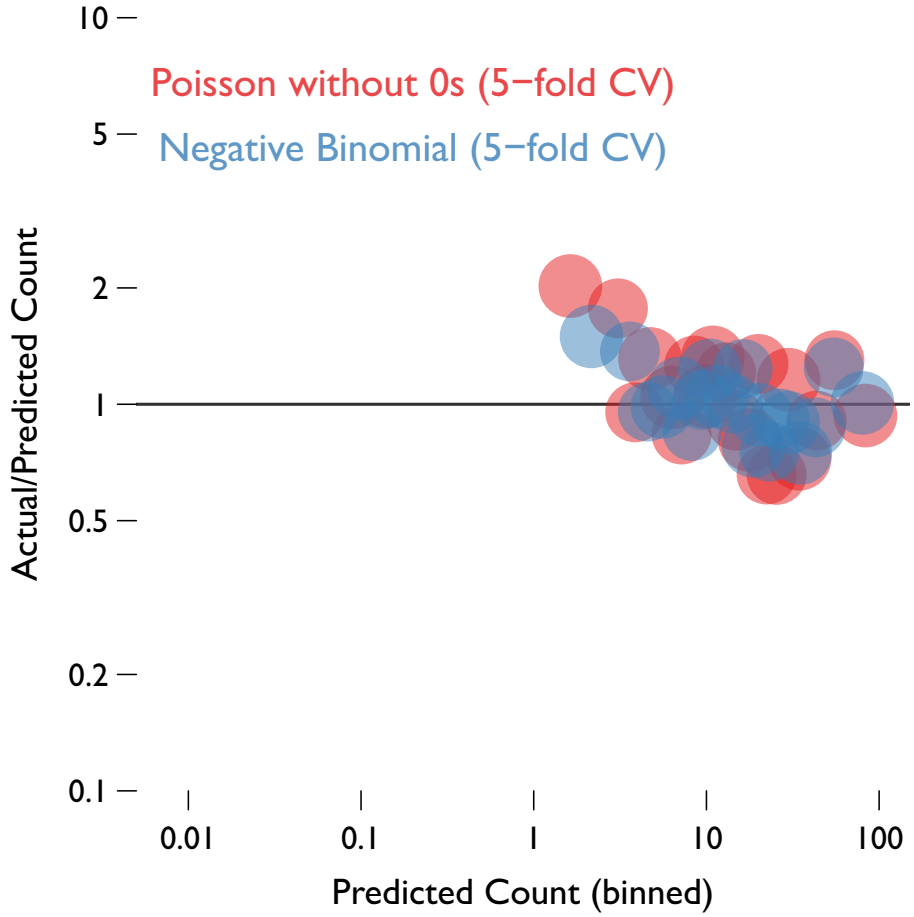
Expected HOA foreclosure filings per 1000 homes per year



Excluding all zeros, home prices and foreclosure rates still negatively related

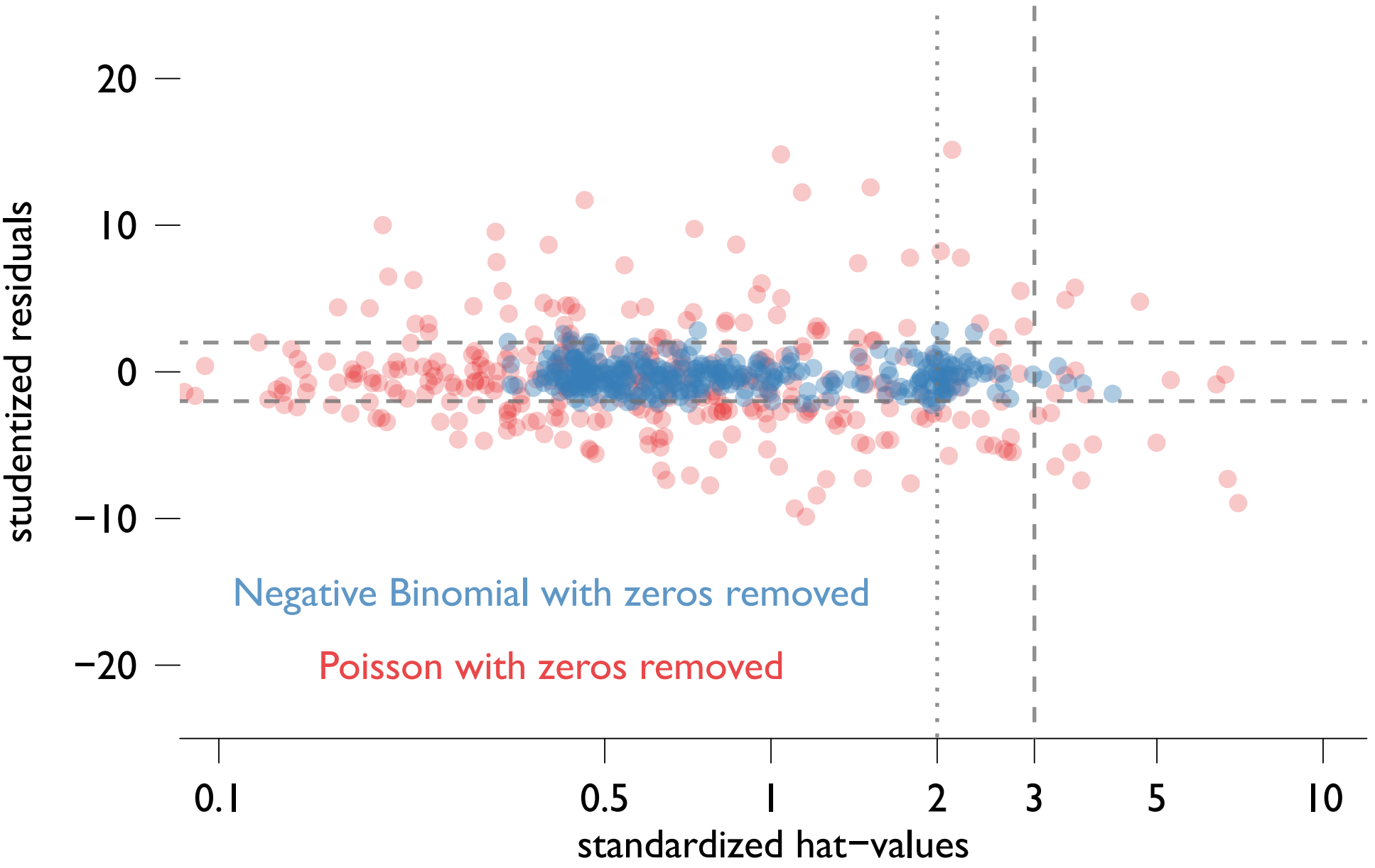
The confidence intervals are now much more plausible





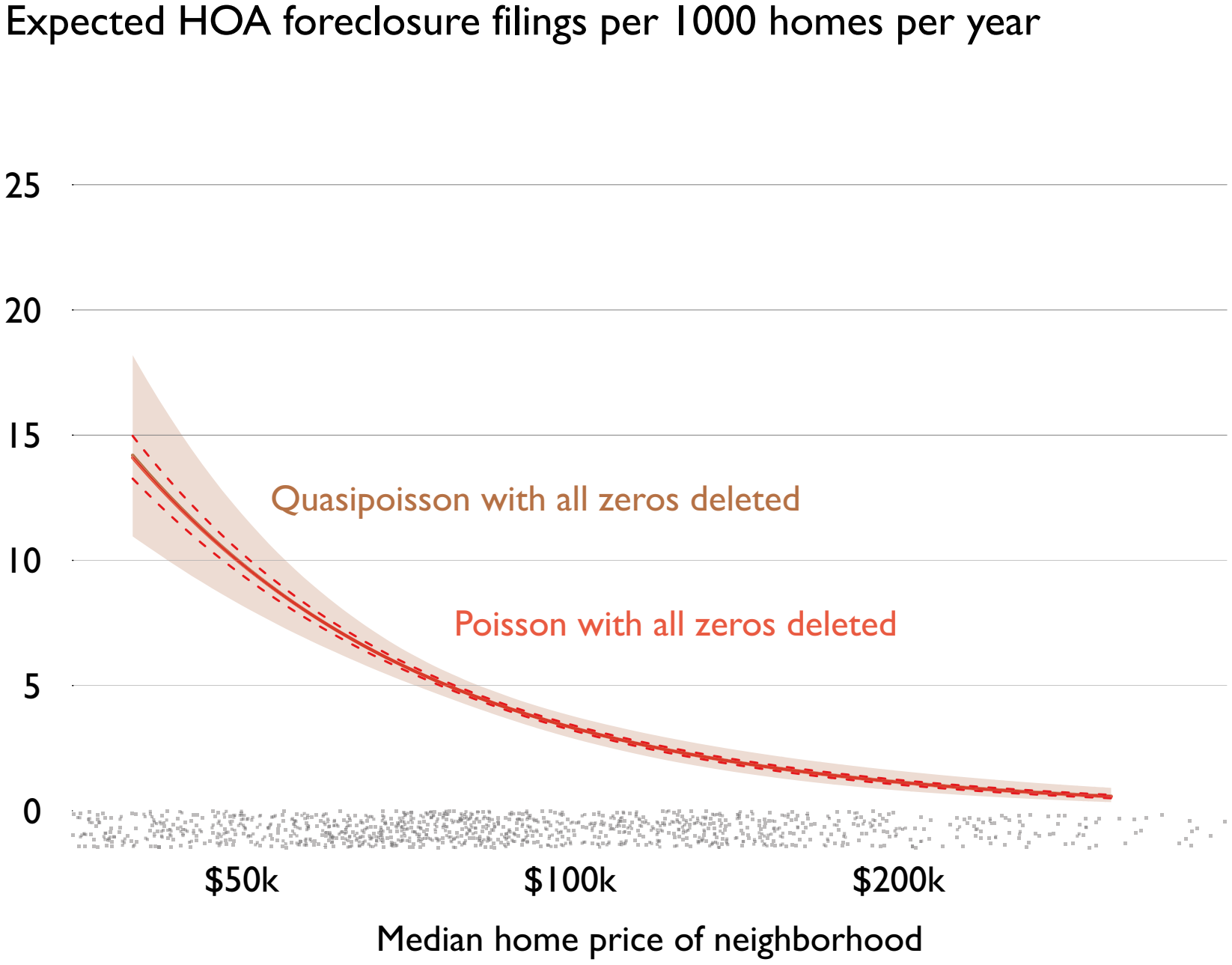
Slight improvement – perhaps not much should be expected

Main point of Negative Binomial is improving estimates of se($\hat{\beta}$)

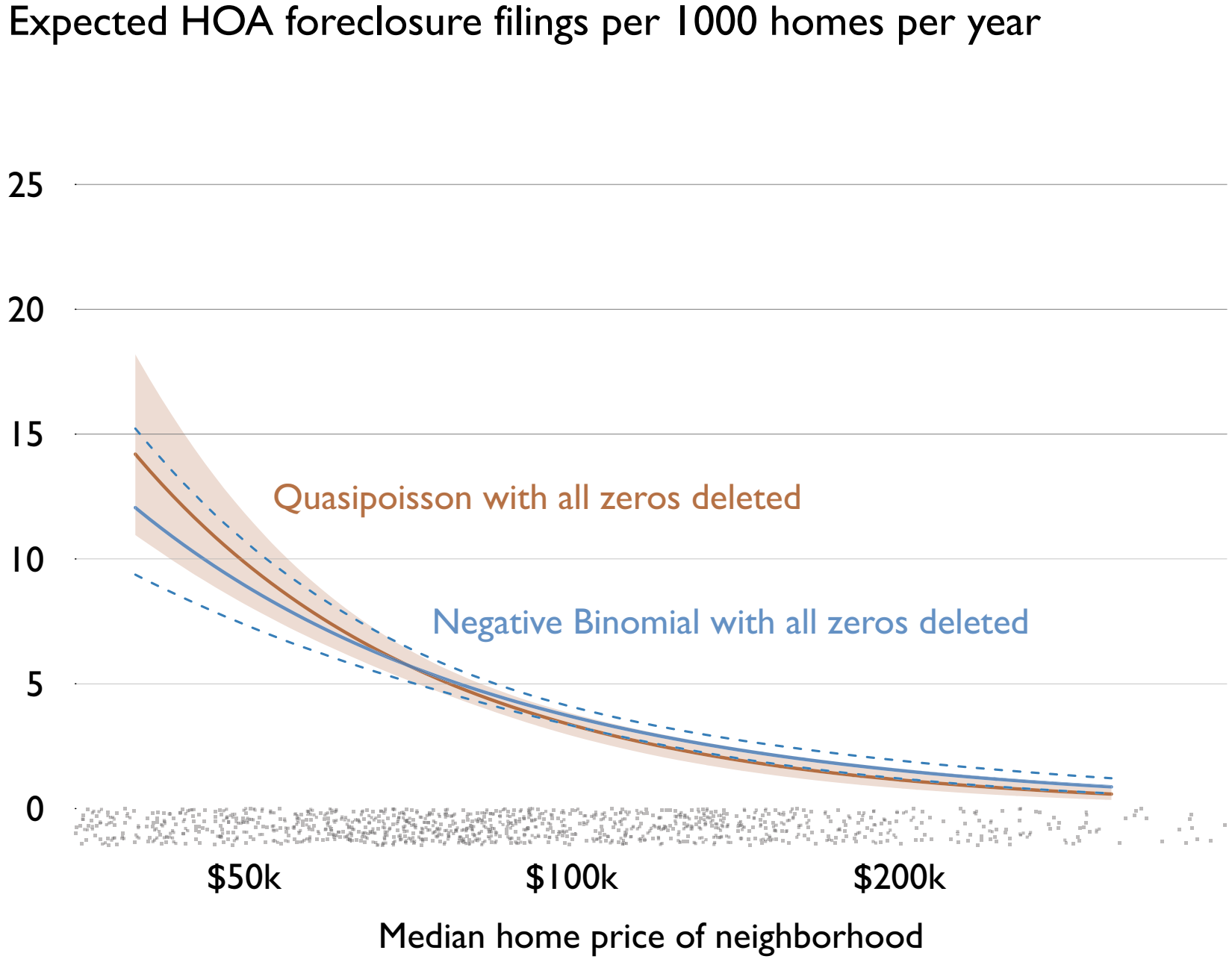


However, the studentized residuals are much better behaved

Suggests Negative Binomial is a more appropriate distribution for these data

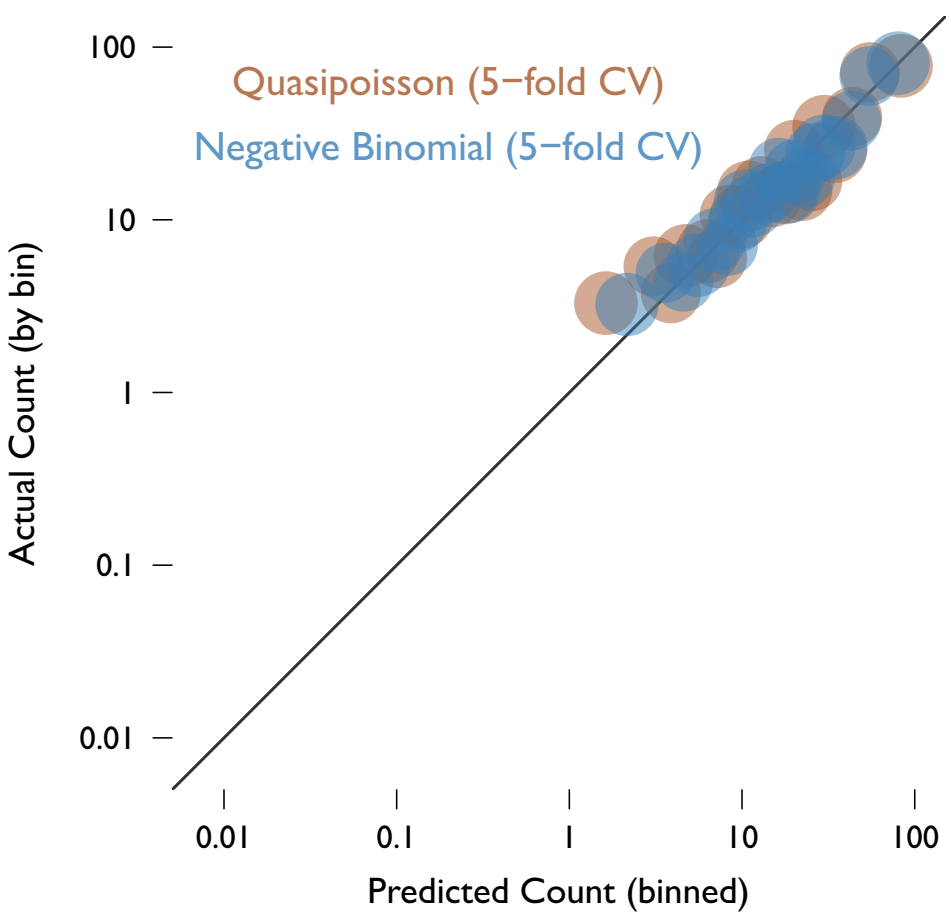


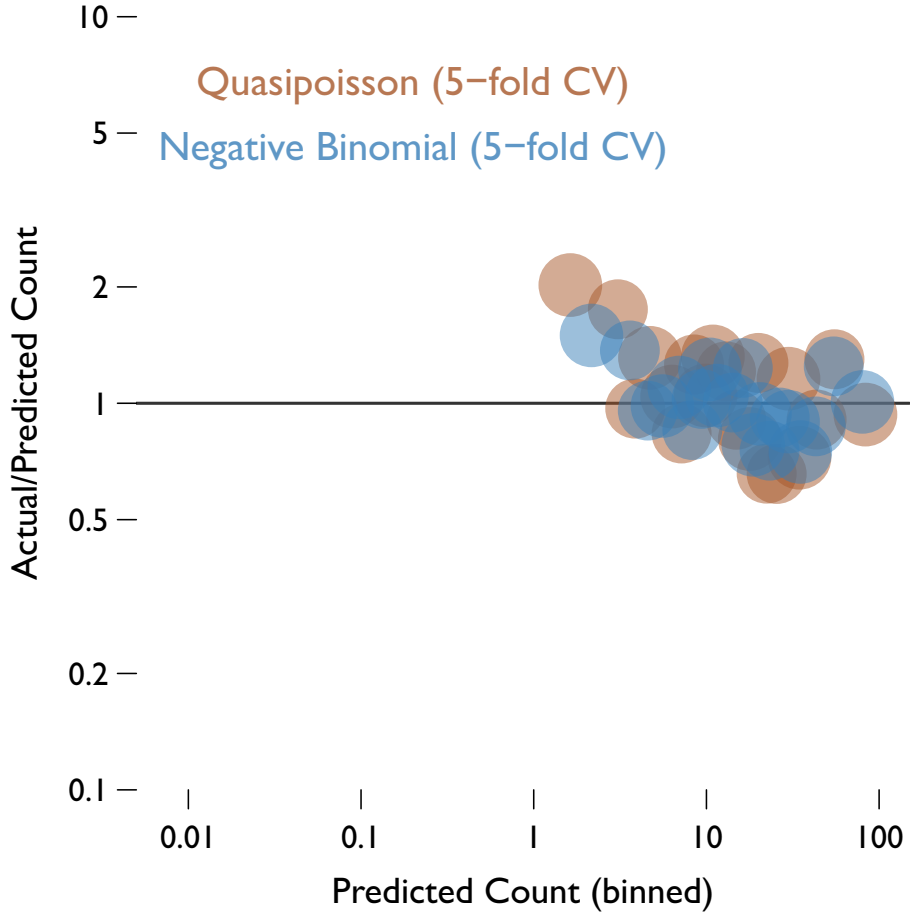
Quasipoisson finds *exactly* the same relationship as the Poisson, but confidence intervals have been adjusted for overdispersion – much like "robust" standard errors in linear regression



Quasipoisson a plausible alternative to NB: similar CIs & mostly similar findings

How do we decide between these models? What should we report?

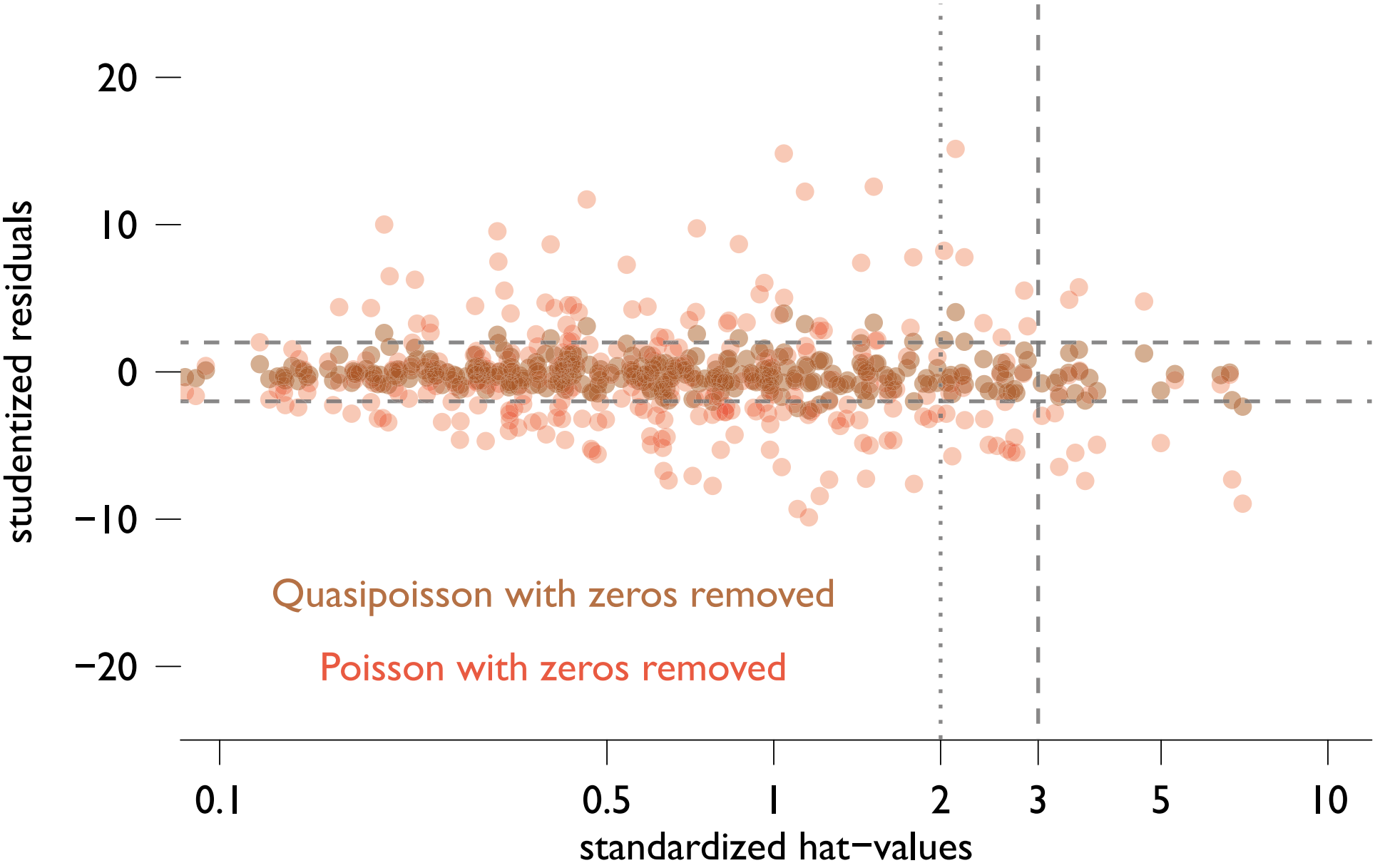




Quasipoisson had slightly smaller in-sample & cross-validated error than NB

Cross-validated Actual vs Predicted plots provide a more nuanced view

While the Negative Binomial looks slightly better to me –
at least at predicting low rates – very little difference overall



The quasipoisson solves the excessive GLM residual problem,
but does leave behind more high-leverage residuals than the NB

Zeros Models

Now, the outstanding problem is the extra zeros we deleted

Did we introduce bias?

Can we somehow estimate a model with the zeros?

Zeros Models

Now, the outstanding problem is the extra zeros we deleted

Did we introduce bias?

Can we somehow estimate a model with the zeros?

Two basic approaches: Hurdle or Zero-Inflation.

The Hurdle: Going from 0 to 1 is a qualitatively different decision compared to getting from 1 to any other positive count

Zeros Models

Now, the outstanding problem is the extra zeros we deleted

Did we introduce bias?

Can we somehow estimate a model with the zeros?

Two basic approaches: Hurdle or Zero-Inflation.

The Hurdle: Going from 0 to 1 is a qualitatively different decision compared to getting from 1 to any other positive count

- Eating potato chips? (according to the slogan)

Zeros Models

Now, the outstanding problem is the extra zeros we deleted

Did we introduce bias?

Can we somehow estimate a model with the zeros?

Two basic approaches: Hurdle or Zero-Inflation.

The Hurdle: Going from 0 to 1 is a qualitatively different decision compared to getting from 1 to any other positive count

- Eating potato chips? (according to the slogan)
  (The hurdle is deciding to open the bag;
  the event count is what happens to your cholesterol count afterwards)
- Committing violent crimes?

Zeros Models

Now, the outstanding problem is the extra zeros we deleted

Did we introduce bias?

Can we somehow estimate a model with the zeros?

Two basic approaches: Hurdle or Zero-Inflation.

The Hurdle: Going from 0 to 1 is a qualitatively different decision compared to getting from 1 to any other positive count

- Eating potato chips? (according to the slogan)
  (The hurdle is deciding to open the bag;
  the event count is what happens to your cholesterol count afterwards)
- Committing violent crimes?
- Anything where the "capital" costs are large relative to the marginal cost?

Zeros Models

Zero-inflation: Some observations are structurally zero, others are unrestricted counts.

Zeros Models

Zero-inflation: Some observations are structurally zero,
others are unrestricted counts.

- Ex. of structural zero: number of cigarettes smoked by a non-smoker

Zeros Models

Zero-inflation: Some observations are structurally zero,
others are unrestricted counts.

- Ex. of structural zero: number of cigarettes smoked by a non-smoker
  But even a smoker could have a zero count for some finite period

Zeros Models

Zero-inflation: Some observations are structurally zero,
others are unrestricted counts.

- Ex. of structural zero: number of cigarettes smoked by a non-smoker
  But even a smoker could have a zero count for some finite period

- More structural zeros: number of wars started by a country with no army;
  the number of political donations given in a year by an apathetic citizen

Zeros Models

Zero-inflation: Some observations are structurally zero,
others are unrestricted counts.

- Ex. of structural zero: number of cigarettes smoked by a non-smoker
  But even a smoker could have a zero count for some finite period

- More structural zeros: number of wars started by a country with no army;
  the number of political donations given in a year by an apathetic citizen

- Some people/countries just don't do certain things;
  other people do them with some positive rate.

- That rate could still produce a zero.

In either hurdle or ZI models, what gets you into the count is a different process,
with potentially different covariates, from the count itself.

Zeros Models

You can add Hurdles or Zero-Inflation to any count model

Hence, you could use ML to estimate

Zero-inflated Poisson models

Zero-inflated Negative Binomial models

Hurdle Poisson models

Hurdle Negative Binomial models

Zero-inflated Binomial models

Zero-inflated Beta-Binomial models

Hurdle Binomial models

Hurdle Beta-Binomial models

Hurdle or zero-inflated quasilikelihood models also possible

Mixture Models

We say the data generating process is a mixture
when observations may come from different probability distributions

Example Model	Mixture of
Tobit	Bernoulli and censored Normal
Zero-inflated Poisson	Bernoulli and Poisson
Zero-inflated Negative Binomial	Bernoulli and Negative Binomial
Hurdle Poisson	Bernoulli and truncated Poisson
Hurdle Negative Binomial	Bernoulli and truncated Negative Binomial
Zero-inflated Binomial	Bernoulli and Binomial
Zero-inflated Beta-Binomial	Bernoulli and Beta-Binomial
Hurdle Binomial	Bernoulli and truncated Binomial
Hurdle Beta-Binomial	Bernoulli and truncated Beta-Binomial

Today: Zero-inflated Poisson (ZIP) & Zero-inflated Negative Binomial (ZINB)

Zero-inflated Poisson

Process 1: Structural zeros arise from the Bernoulli with probability ψ_i:

$$\psi_i = \frac{1}{1 + \exp(-\mathbf{z}\gamma)}$$

Process 2: Incidental zeros and non-zero counts arise from the Poisson

$$\Pr(y_i|\lambda_i, \psi_i = 0) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

Zero-inflated Poisson

Process 1: Structural zeros arise from the Bernoulli with probability ψ_i:

$$\psi_i = \frac{1}{1 + \exp(-\mathbf{z}\gamma)}$$

Process 2: Incidental zeros and non-zero counts arise from the Poisson

$$\Pr(y_i|\lambda_i, \psi_i = 0) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$$

Think of the DGP as drawing two numbers simultaneously:

1. The first is a Bernoulli coin-flip d: if $d = 1$, we record $y_i = 0$
2. The second is a Poisson count c: if the Bernoulli produced $d = 0$, we record $y_i = c$ as this count, regardless of whether $c = 0$ or not

Zero-inflated Poisson

Combining the zeros from both processes, we find

$$\Pr(y_i = 0|\psi_i, \lambda_i) = \psi_i$$

Zero-inflated Poisson

Combining the zeros from both processes, we find

$$\Pr(y_i = 0|\psi_i, \lambda_i) = \psi_i + (1 - \psi_i) \exp(-\lambda_i)$$

Zero-inflated Poisson

Combining the zeros from both processes, we find

$$\Pr(y_i = 0|\psi_i, \lambda_i) = \psi_i + (1 - \psi_i)\exp(-\lambda_i)$$

and for all positive integers y,

$$\Pr(y_i = y|\psi_i, \lambda_i) = (1 - \psi_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

In words, ψ_i is the probability of an excess zero;
if this probability is not met, the DGP defaults to a Poisson

Zero-inflated Poisson

The likelihood has two pieces,
corresponding to the two pieces of the probability function

$$\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{y_i=0} \{\psi_i + (1-\psi_i)\exp(-\lambda_i)\} \prod_{y_i>0} \left\{(1-\psi_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right\}$$

Zero-inflated Poisson

The likelihood has two pieces,
corresponding to the two pieces of the probability function

$$\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{y_i=0} \{\psi_i + (1-\psi_i)\exp(-\lambda_i)\} \prod_{y_i>0} \left\{(1-\psi_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right\}$$

Substituting for ψ_i and λ_i and taking logs yields

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y}, \mathbf{Z}, \mathbf{X}) &= \sum_{y_i=0} \log\left\{\frac{1}{1+\exp(-\mathbf{z}_i\boldsymbol{\gamma})} + \frac{1}{1+\exp(\mathbf{z}_i\boldsymbol{\gamma})}\exp(-\mathbf{x}_i\boldsymbol{\beta})\right\} \\ &\quad + \sum_{y_i>0}\left\{\log\left(\frac{1}{1+\exp(\mathbf{z}_i\boldsymbol{\gamma})}\right) - \exp(\mathbf{x}_i\boldsymbol{\beta}) + \mathbf{x}_i\boldsymbol{\beta}y_i\right\}\end{aligned}$$

which can be optimized numerically, as with `optim()`

Zero-Inflated Count Model Parameters

We have two sets of regressors, which may be overlapping

γ's govern the excess zeros, and can be interpreted in logit fashion.

β's govern the expected count given no excess zeros,
and can be interpreted in Poisson fashion.

Zero-Inflated Count Model Parameters

We have two sets of regressors, which may be overlapping

γ's govern the excess zeros, and can be interpreted in logit fashion.

β's govern the expected count given no excess zeros,
and can be interpreted in Poisson fashion.

These parameters can be combined to find a variety of quantities of interest

For expected values alone,
there are four obvious quantities of interest

Which one(s) you should show depends on your research question

Zero-Inflated Count Model Quantities of Interest

QoI 1 *The probability of a structural zero*

$$\Pr(y \text{ is a structural zero} | \mathbf{x}_c, \mathbf{z}_c) \;=\; \mathbb{E}(\psi | \mathbf{z}_c) \;=\; [1 + \exp(-\mathbf{z}_c \boldsymbol{\gamma})]^{-1}$$

Zero-Inflated Count Model Quantities of Interest

QoI 1 *The probability of a structural zero*

$$\Pr(y \text{ is a structural zero}|\mathbf{x}_c, \mathbf{z}_c) \;=\; \mathbb{E}(\psi|\mathbf{z}_c) \;=\; [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1}$$

QoI 2 *The probability of either a structural or incidental zero*

$$\begin{aligned}\Pr(y=0|\mathbf{x}_c, \mathbf{z}_c) &= \mathbb{E}(\psi|\mathbf{z}_c) - \mathbb{E}(1-\psi|\mathbf{z}_c)\Pr(y=0|\psi=0,\lambda) \\ &= [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1} + [1+\exp(\mathbf{z}_c\boldsymbol{\gamma})]^{-1}\exp(-\mathbf{x}_c\boldsymbol{\beta})\end{aligned}$$

Zero-Inflated Count Model Quantities of Interest

QoI 1 *The probability of a structural zero*

$$\Pr(y \text{ is a structural zero}|\mathbf{x}_c, \mathbf{z}_c) \;=\; \mathbb{E}(\psi|\mathbf{z}_c) \;=\; [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1}$$

QoI 2 *The probability of either a structural or incidental zero*

$$\begin{aligned}\Pr(y=0|\mathbf{x}_c, \mathbf{z}_c) &= \mathbb{E}(\psi|\mathbf{z}_c) - \mathbb{E}(1-\psi|\mathbf{z}_c)\Pr(y=0|\psi=0,\lambda) \\ &= [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1} + [1+\exp(\mathbf{z}_c\boldsymbol{\gamma})]^{-1}\exp(-\mathbf{x}_c\boldsymbol{\beta})\end{aligned}$$

QoI 3 *The expected count unconditional on the presence of structural zeros*

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}_c, \mathbf{z}_c) &= \mathbb{E}(\psi|\mathbf{z}_c)\times 0 \quad + \quad \mathbb{E}(1-\psi|\mathbf{z}_c)\mathbb{E}(\lambda|\mathbf{x}_c) \\ &= \quad [1+\exp(\mathbf{z}_c\boldsymbol{\gamma})]^{-1}\exp(\mathbf{x}_c\boldsymbol{\beta})\end{aligned}$$

Zero-Inflated Count Model Quantities of Interest

QoI 1 *The probability of a structural zero*

$$\Pr(y \text{ is a structural zero}|\mathbf{x}_c, \mathbf{z}_c) \;=\; \mathbb{E}(\psi|\mathbf{z}_c) \;=\; [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1}$$

QoI 2 *The probability of either a structural or incidental zero*

$$\begin{aligned}\Pr(y=0|\mathbf{x}_c, \mathbf{z}_c) &= \mathbb{E}(\psi|\mathbf{z}_c) - \mathbb{E}(1-\psi|\mathbf{z}_c)\Pr(y=0|\psi=0,\lambda) \\ &= [1+\exp(-\mathbf{z}_c\boldsymbol{\gamma})]^{-1} + [1+\exp(\mathbf{z}_c\boldsymbol{\gamma})]^{-1}\exp(-\mathbf{x}_c\boldsymbol{\beta})\end{aligned}$$

QoI 3 *The expected count unconditional on the presence of structural zeros*

$$\begin{aligned}\mathbb{E}(y|\mathbf{x}_c, \mathbf{z}_c) &= \mathbb{E}(\psi|\mathbf{z}_c)\times 0 \quad + \quad \mathbb{E}(1-\psi|\mathbf{z}_c)\mathbb{E}(\lambda|\mathbf{x}_c) \\ &= \quad [1+\exp(\mathbf{z}_c\boldsymbol{\gamma})]^{-1}\exp(\mathbf{x}_c\boldsymbol{\beta})\end{aligned}$$

QoI 4 *The expected count for a case assumed not to be a structural zero*

$$\mathbb{E}(y|\psi=0, \mathbf{x}_c, \mathbf{z}_c) \;=\; \mathbb{E}(\lambda|\mathbf{x}_c) \;=\; \exp(\mathbf{x}_c\boldsymbol{\beta})$$

Zero-Inflated Negative Binomial

What if we have excess zeros *and* expect contagion among the counts?

Zero-Inflated Negative Binomial

What if we have excess zeros *and* expect contagion among the counts?

We can construct a Zero-Inflated Negative Binomial the same way we made the ZIP

ZINB parameters have same interpretation as ZIP parameters

The algebra of the likelihood is just a bit more complicated. . .

$$\log \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha} | \mathbf{y}, \mathbf{Z}, \mathbf{X}) =$$

$$\sum_{y_i=0} \log \left\{ \frac{1}{1+\exp(-\mathbf{z}_i\boldsymbol{\gamma})} + \frac{1}{1+\exp(\mathbf{z}_i\boldsymbol{\gamma})} \left(\frac{1}{1+\alpha \mathbf{x}_i \boldsymbol{\beta}} \right)^{\frac{1}{\alpha}} \right\}$$

$$+ \sum_{y_i>0} \left\{ \log \left(\frac{1}{1+\exp(\mathbf{z}_i\boldsymbol{\gamma})} \right) - \log \Gamma \left(\frac{1}{\alpha} + y_i \right) - \log \Gamma \left(\frac{1}{\alpha} \right) \right.$$

$$\left. + \frac{1}{\alpha} \log \left(\frac{1}{1+\alpha \mathbf{x_i} \boldsymbol{\beta}} \right) + y_i \log \left(1 - \frac{1}{1+\alpha \mathbf{x}_i \boldsymbol{\beta}} \right) \right\}$$

Checking Goodness of Fit

We can use the usual assortment of fit test: AIC, BIC, RMSE, MAE, etc.

But we have some special concerns to check. . .

Is there overdispersion after accounting for excess zeros?

We can test the ZINB against the ZIP
using a t-test of α against a null of zero
again noting a one-sided test is appropriate

Usually, the t-statistic is so large a formal test is superfluous

Are there excess zeros?

To test the ZIP against the Poisson, or the ZINB against the NB,
we'll need a *non-nested test*

A popular option in this case is the Vuong test,
a likelihood test for non-nested models

Zeros models: practical considerations

The `R` package `pscl` has many helpful functions for estimating count models

`zeroinfl` estimates ZIP and ZINB

`hurdle` estimates hurdle Poisson and hurdle NB

`vuong` conducts non-nested Vuong tests

Writing the ZINB in a paper – very (overly?) complete description

The outcome y_i is modeled using the Zero-inflated Negative Binomial (ZINB), a two component mixture model combining a Negative Binomial event count with an additional point mass at zero:

$$\begin{aligned} y_i &\sim \text{ZINB}(\psi_i, \lambda_i, \alpha) \\ \text{logit}(\psi_i) &= \mathbf{z}_i\boldsymbol{\gamma} \\ \log(\lambda_i) &= \mathbf{x}_i\boldsymbol{\beta} + \log(t_i) \end{aligned}$$

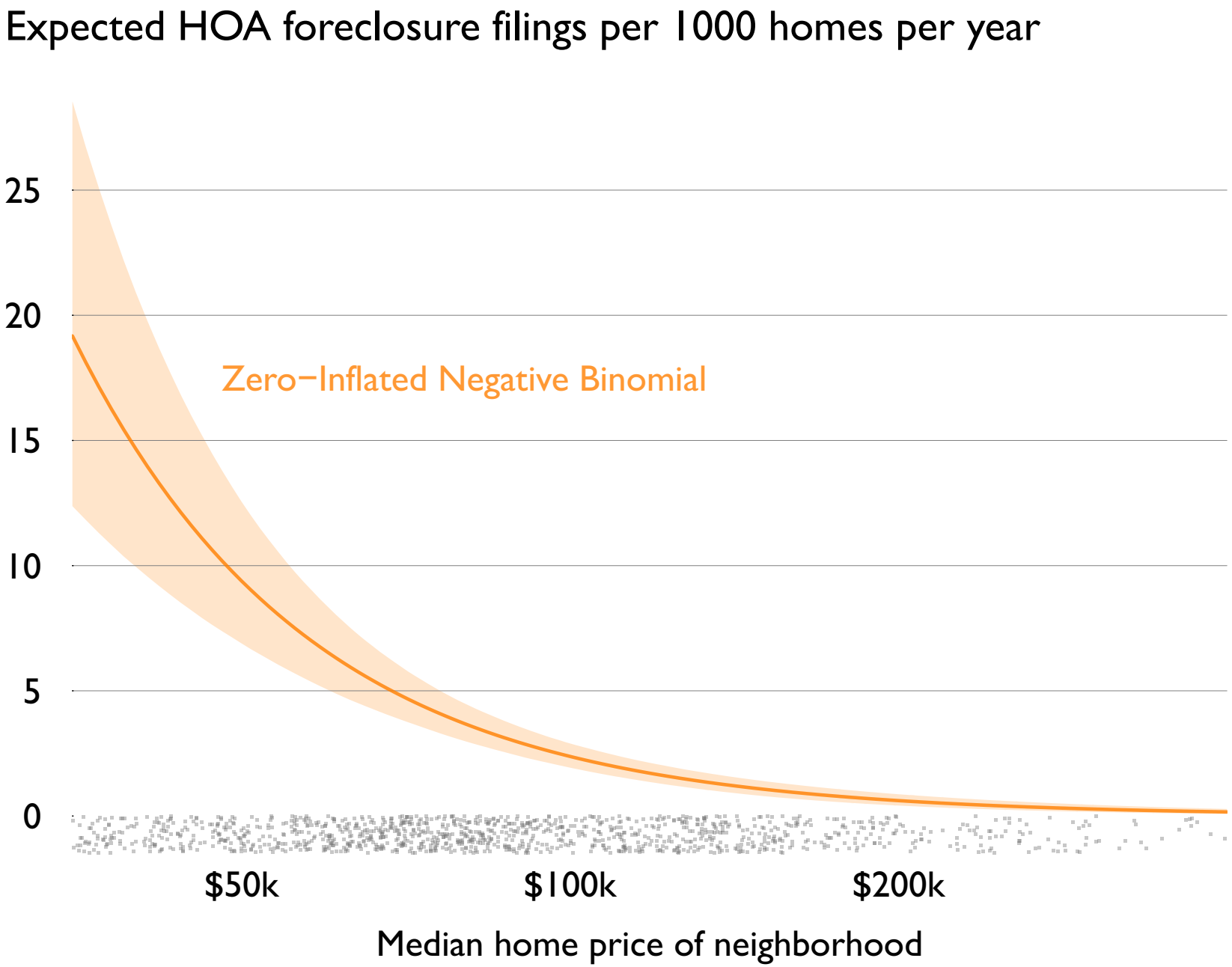
y_i is a structural zero with probability ψ_i. Otherwise, it is a (potentially zero) count with expected value λ_i and overdispersion α. The covariate vectors $\mathbf{x}_i$ and $\mathbf{z}_i$ are potentially overlapping. The offset $\log(t_i)$ adjusts for the size of the at-risk group.

	5: count	5: zeros	6: count	6: zeros
log median valuation	-1.59	-0.30	-1.97	-1.03
	(0.04)	(0.15)	(0.20)	(0.31)
Post-1975 neighborhood	0.74	-2.36	0.46	-2.54
	(0.04)	(0.18)	(0.17)	(0.23)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00	-1.16	1.00	-1.21
	(—)	(0.10)	(—)	(0.14)
Constant	12.00	14.95	16.48	23.37
	(0.49)	(1.88)	(2.29)	(3.91)
"log(theta)"			-0.27	
			(0.18)	
Model	Zero-inflated Poisson		Zero-inflated Negative Binomial	
N	1417		1417	
AIC	7122		3422	
Vuong test vs. no ZI	$p <$0.0001		$p <$0.0001	
In-sample mean absolute error (MAE)	5.52		5.74	
5-fold cross-validated MAE	5.16		5.26	

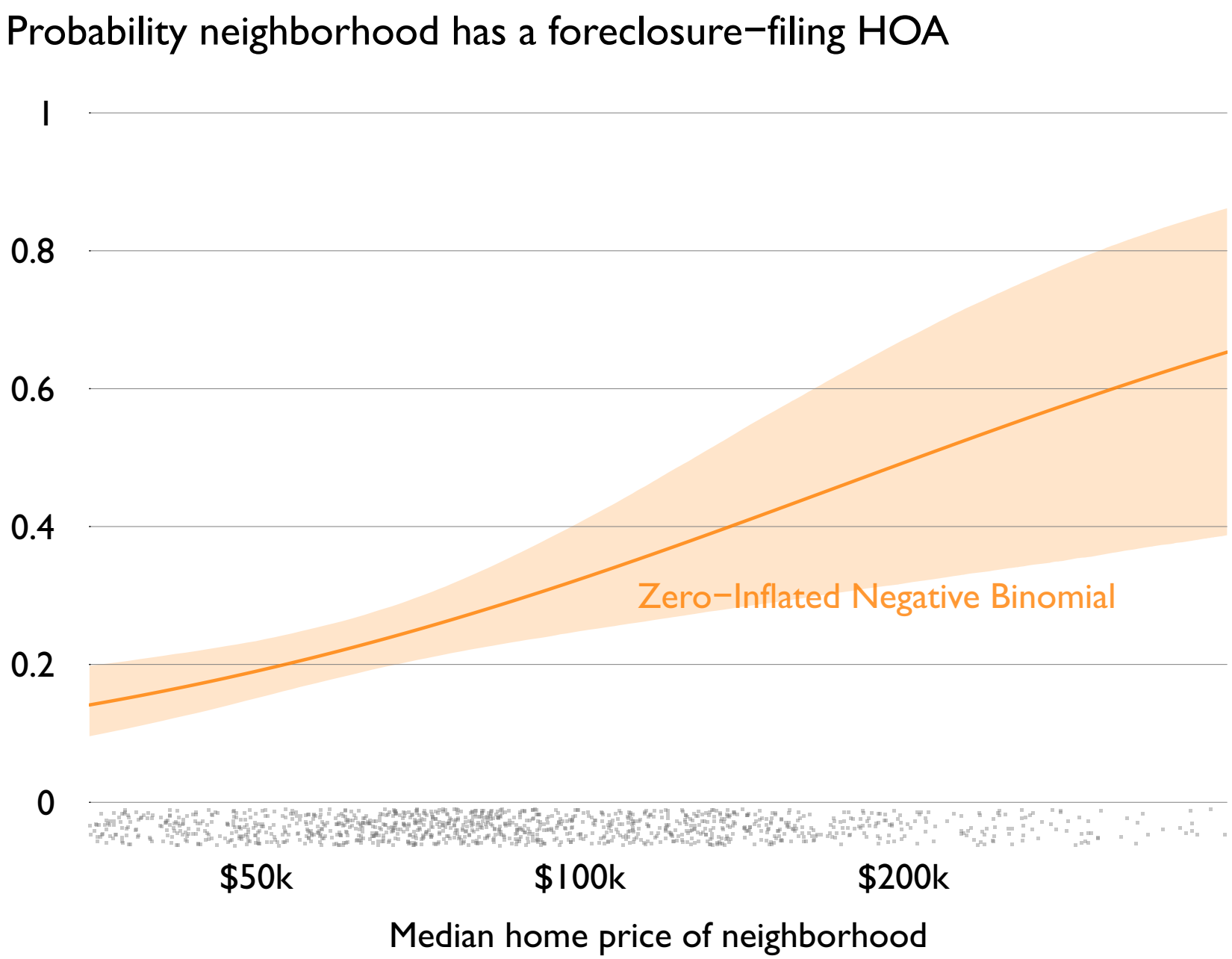
We now have a hard-to-interpret "log(theta)" dispersion parameter

	5: count	5: zeros	6: count	6: zeros
log median valuation	-1.59	-0.30	-1.97	-1.03
	(0.04)	(0.15)	(0.20)	(0.31)
Post-1975 neighborhood	0.74	-2.36	0.46	-2.54
	(0.04)	(0.18)	(0.17)	(0.23)
log $N_{\text{homes}} \times N_{\text{years}}$	1.00	-1.16	1.00	-1.21
	(—)	(0.10)	(—)	(0.14)
Constant	12.00	14.95	16.48	23.37
	(0.49)	(1.88)	(2.29)	(3.91)
Dispersion (α)			1.32	
			(0.24)	
Model	Zero-inflated Poisson		Zero-inflated Negative Binomial	
N	1417		1417	
AIC	7122		3422	
Vuong test vs. no ZI	$p <$0.0001		$p <$0.0001	
In-sample mean absolute error (MAE)	5.52		5.74	
5-fold cross-validated MAE	5.16		5.26	

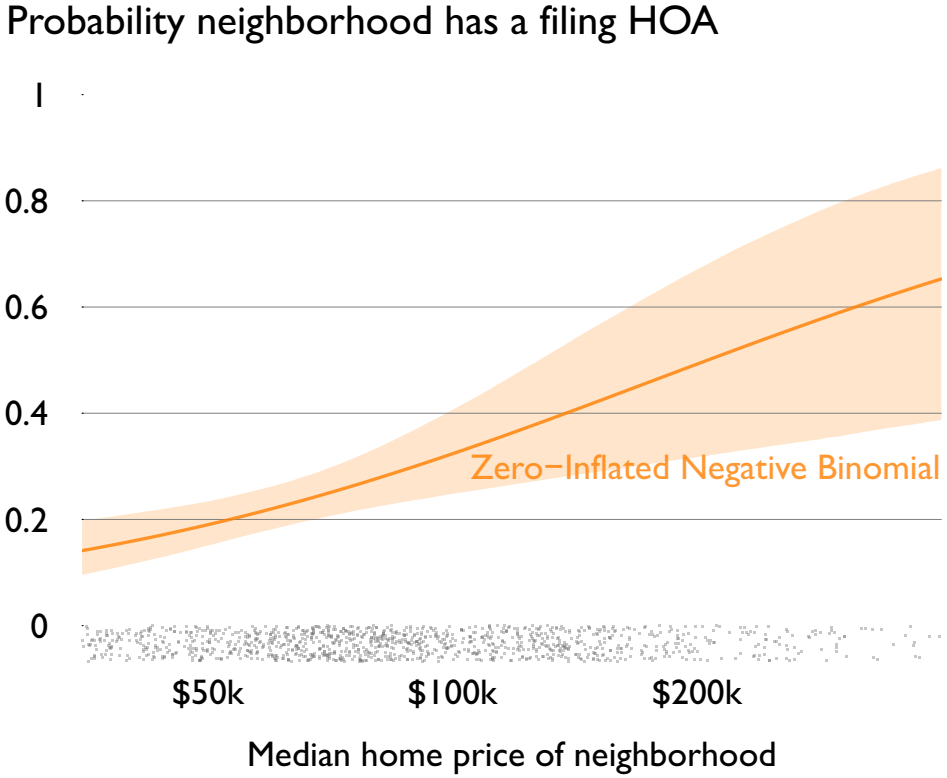
Simulation reveals $\hat{\alpha}$ and $\text{se}(\hat{\alpha})$ *What else have we learned?*

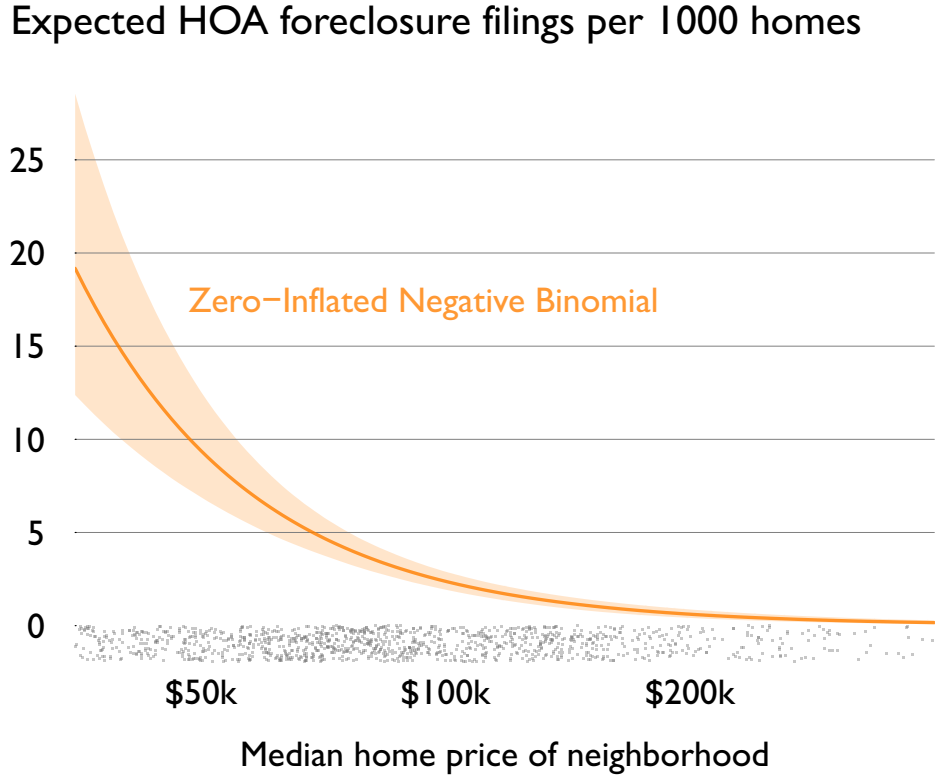


We simulate the expected number of foreclosure filings
given the assumption the HOA is able & willing to file (QoI 4)



The model also estimates the probability of a structural zero (QoI 1), which we use to simulate the probability an HOA is a "foreclosure filing type"



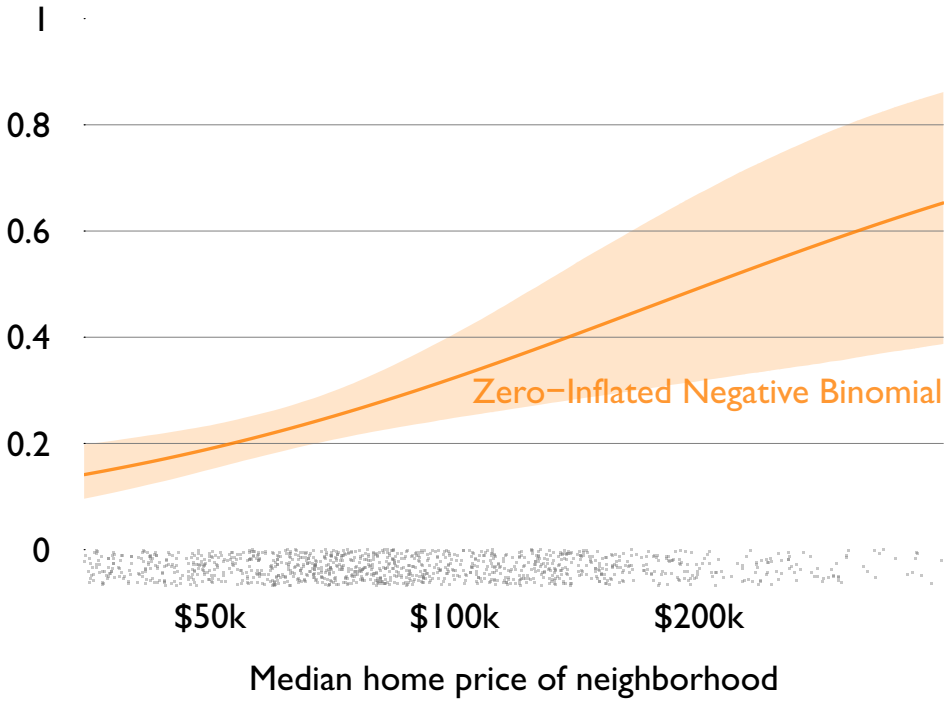


Let's put these two quantities of interest side-by-side

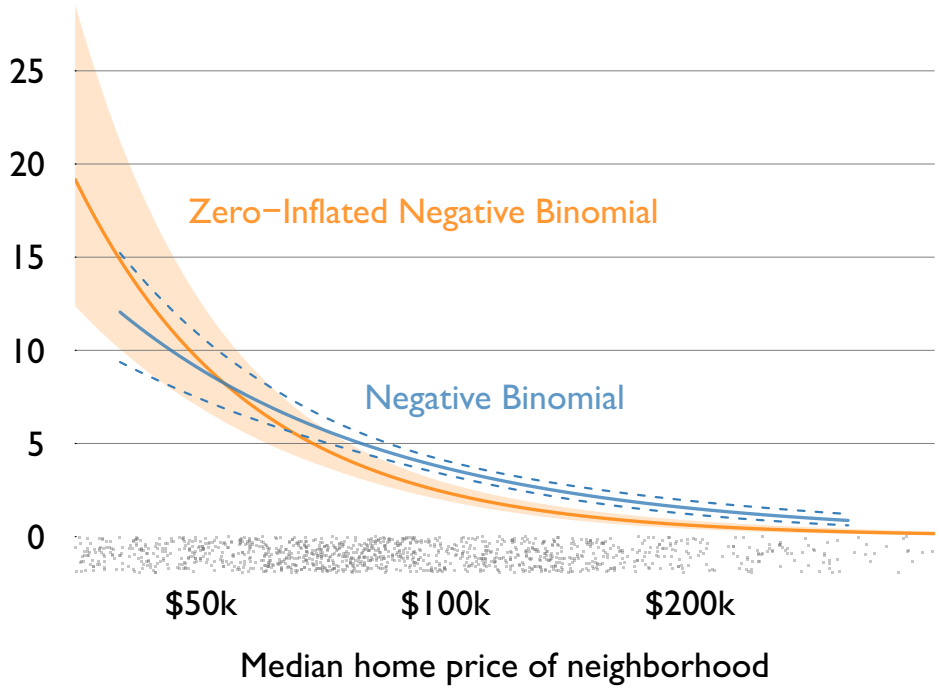
Neighborhoods with more expensive homes are more likely to have filing HOAs:
often these neighborhoods are persuaded powerful HOAs will protect home values

But among HOAs that file,
those in poorer neighborhoods are far more likely to file foreclosures against residents

Probability neighborhood has a filing HOA



Expected HOA foreclosure filings per 1000 homes

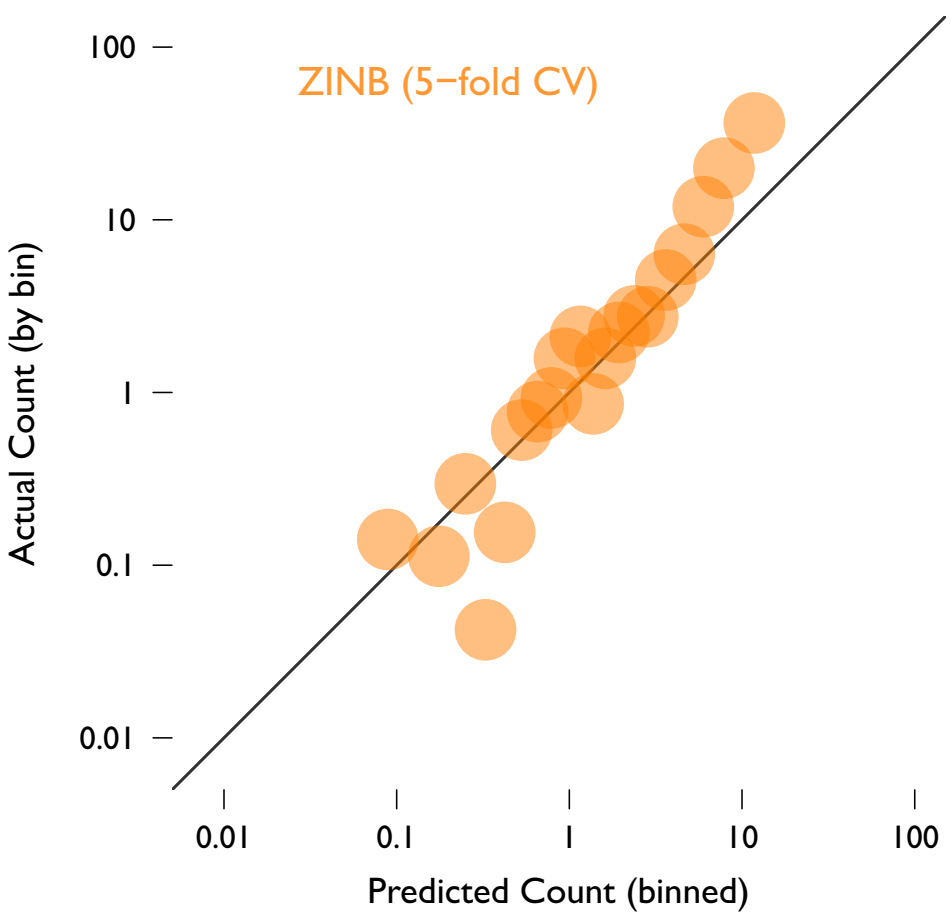


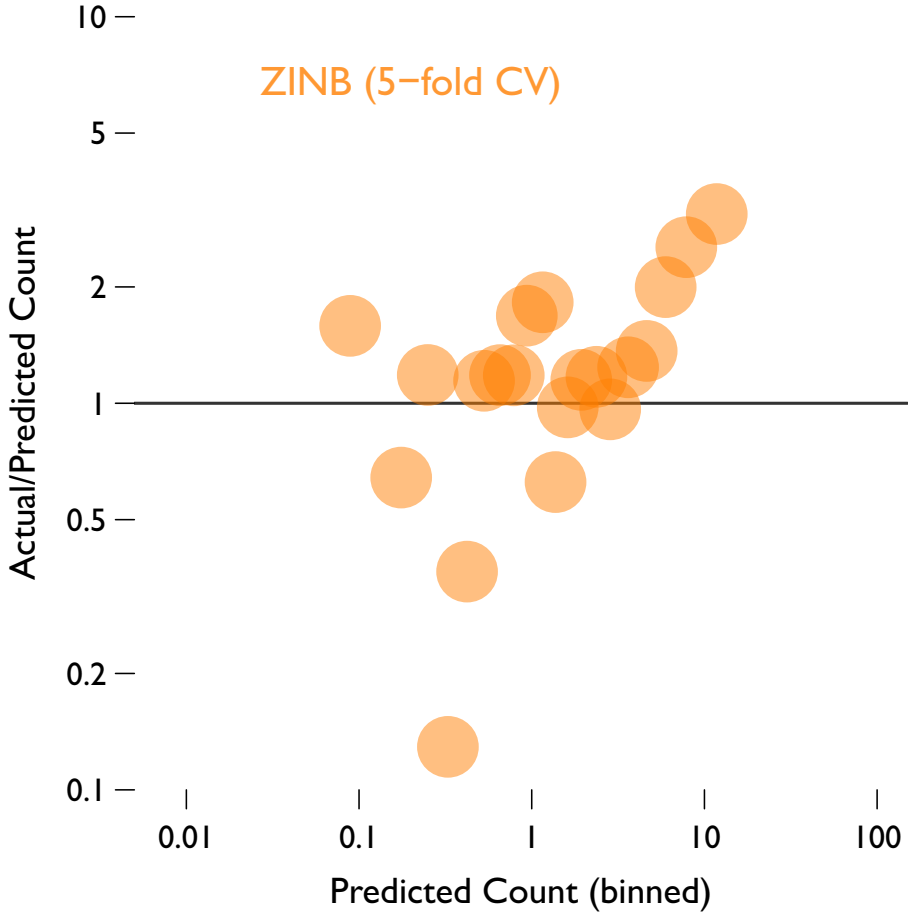
Results are stronger than the Negative Binomial with zeros deleted

This difference in substantive findings is due to

1. the use of different samples

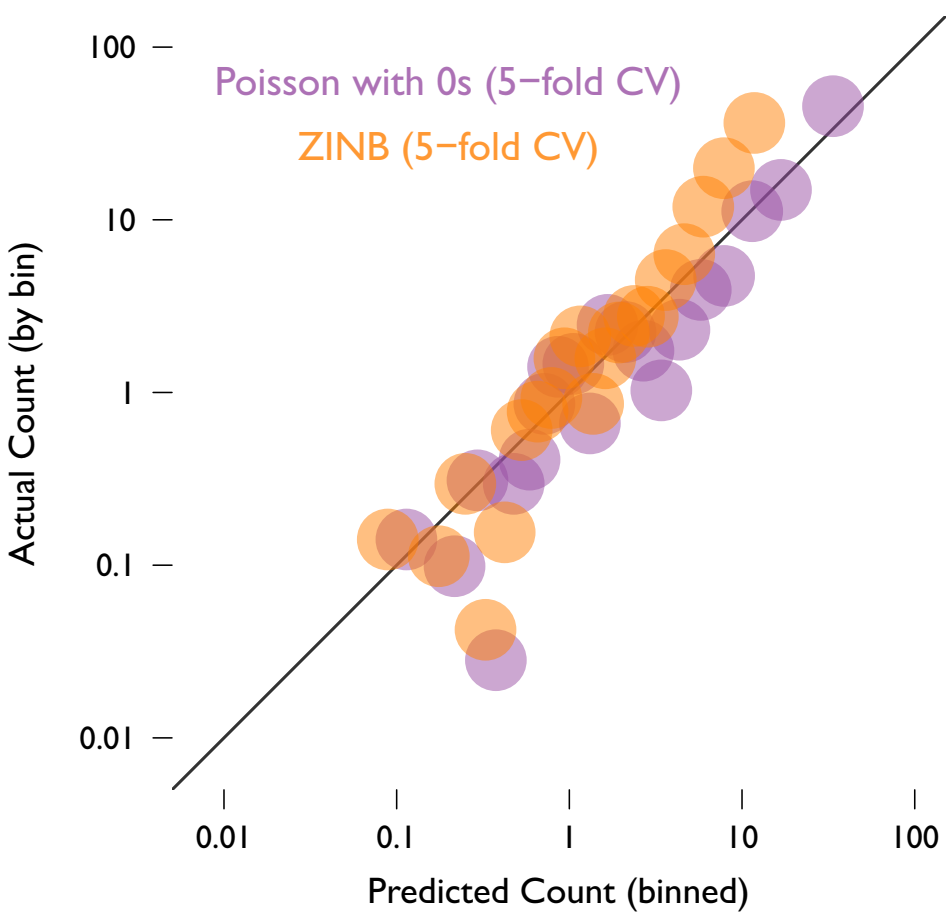
2. the addition of a model of excess zeros competing to explain the source of overdispersion

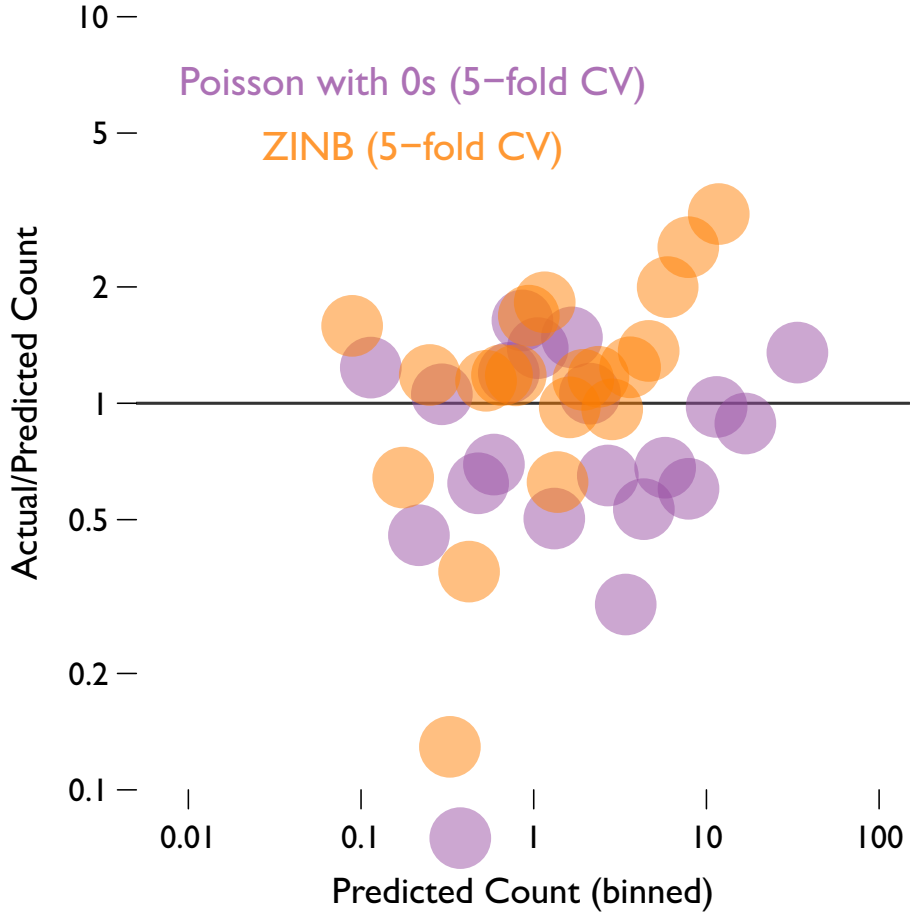




Model fit using Actual versus Predicted plots:

Decent, but not perfect



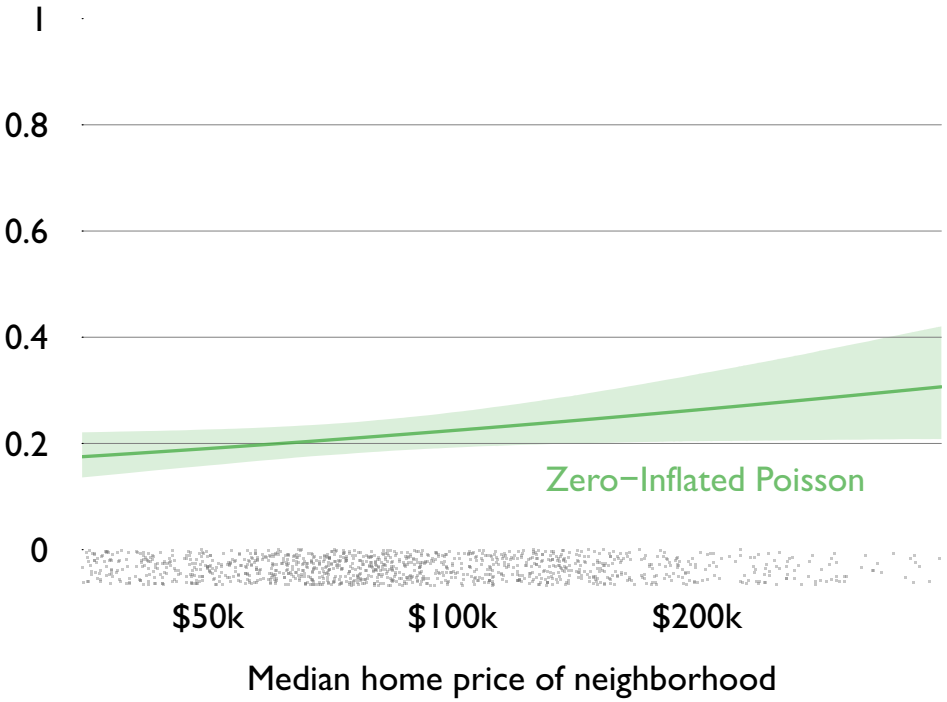


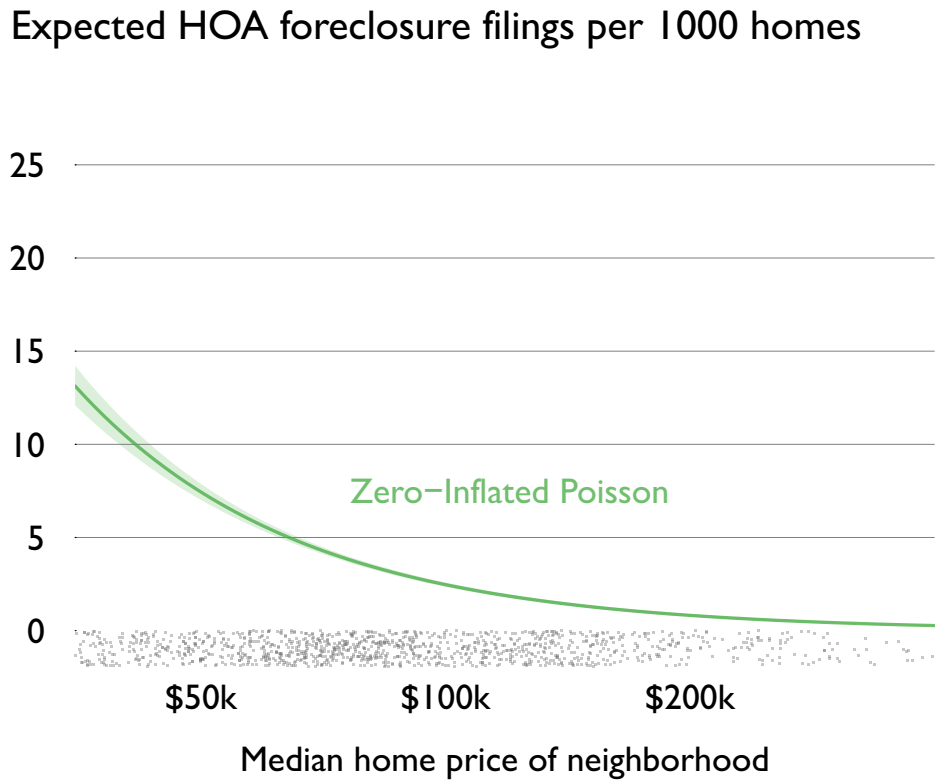
Compared to the original Poisson model, does better at low count neighborhoods

Reasonable, given the better modeling of zeros

But underpredicts the high count cases

Suggests there may be omitted variables predicting high counts

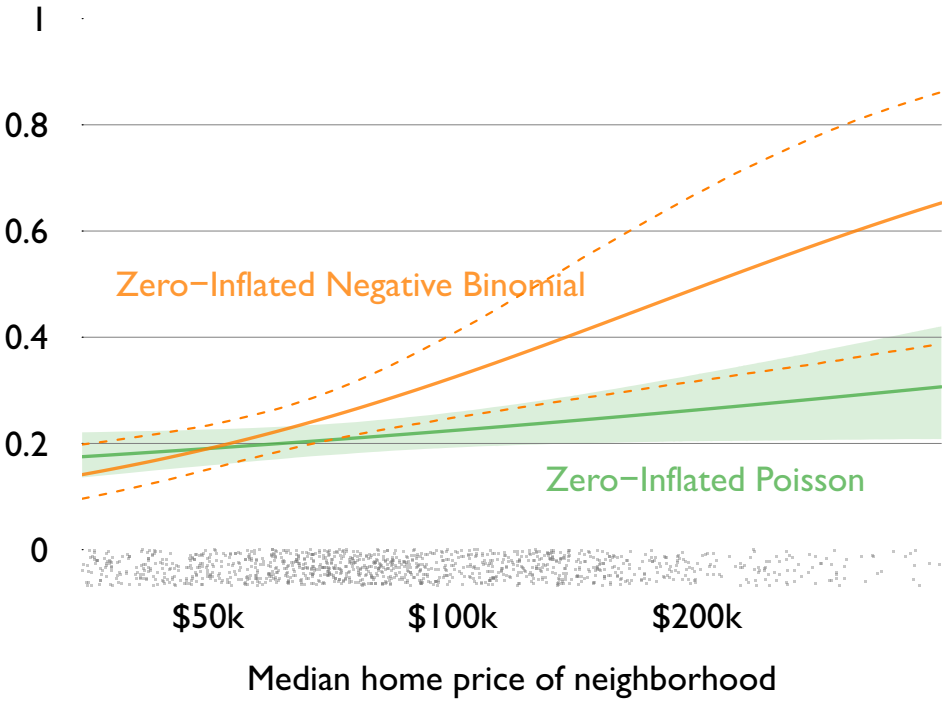




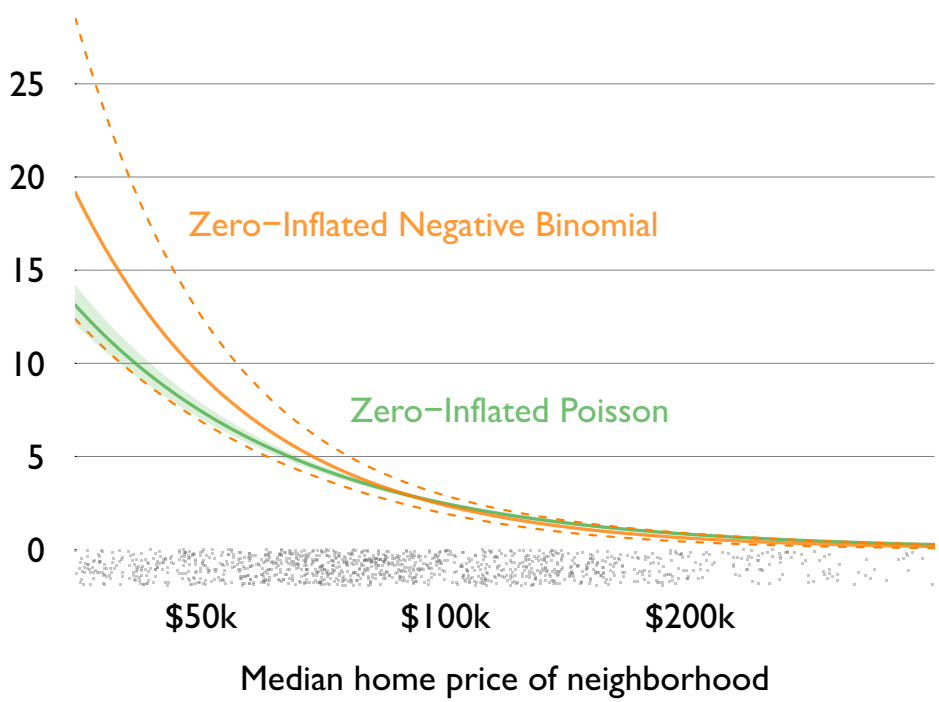
The ZIP results can be interpreted in similar fashion to the ZINB

Note that the excess zeros function has reasonable CIs,
but the expected counts looks dangerously overconfident, as in most Poisson models

Probability neighborhood has a filing HOA



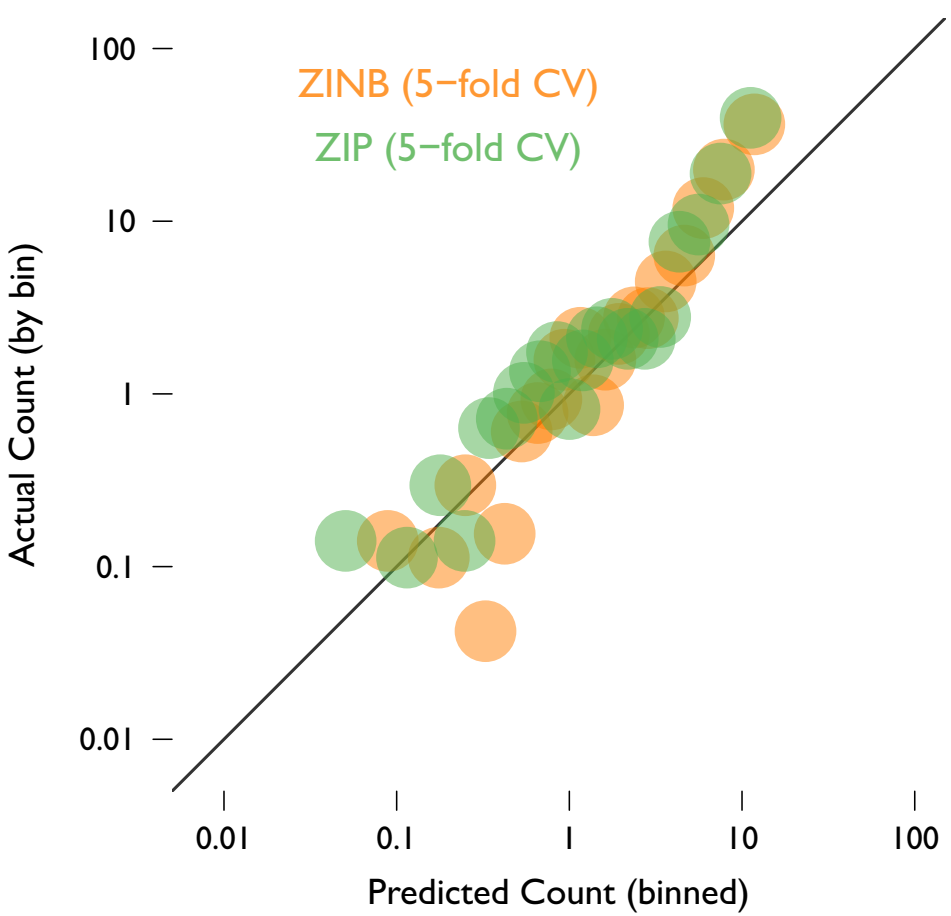
Expected HOA foreclosure filings per 1000 homes

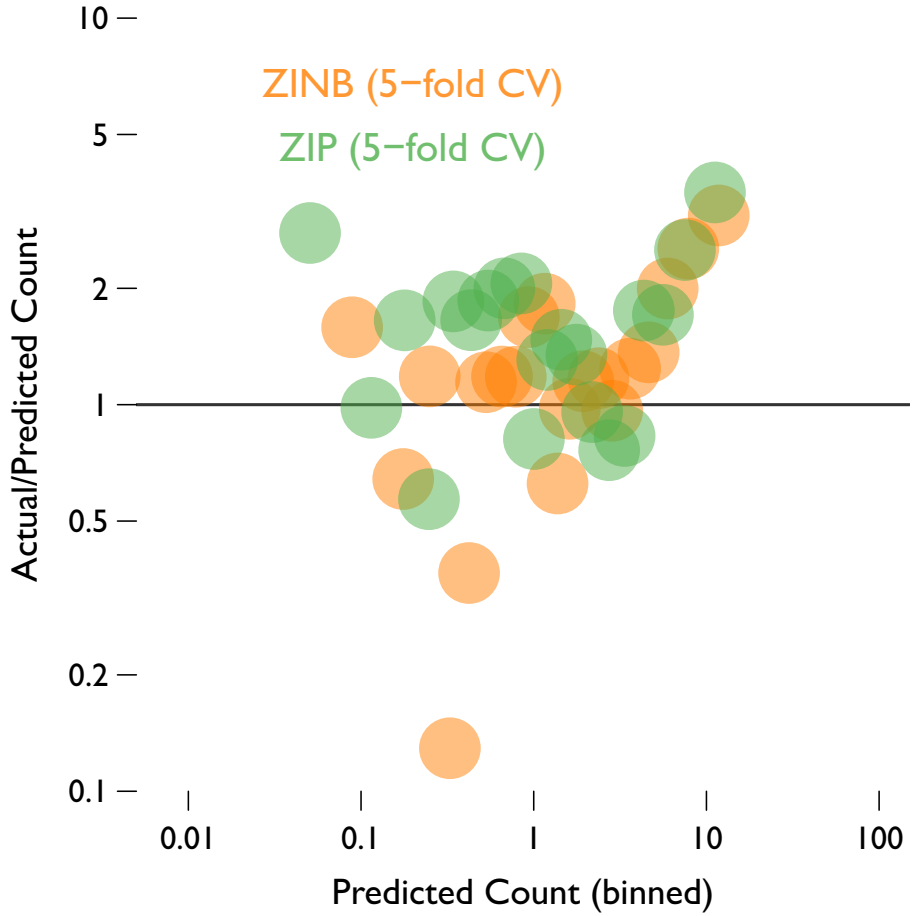


The ZIP results are substantively weaker than the ZINB

This difference in substantive findings is due to

differences in distributional assumptions;
specifically, whether there is Gamma-distributed overdispersion

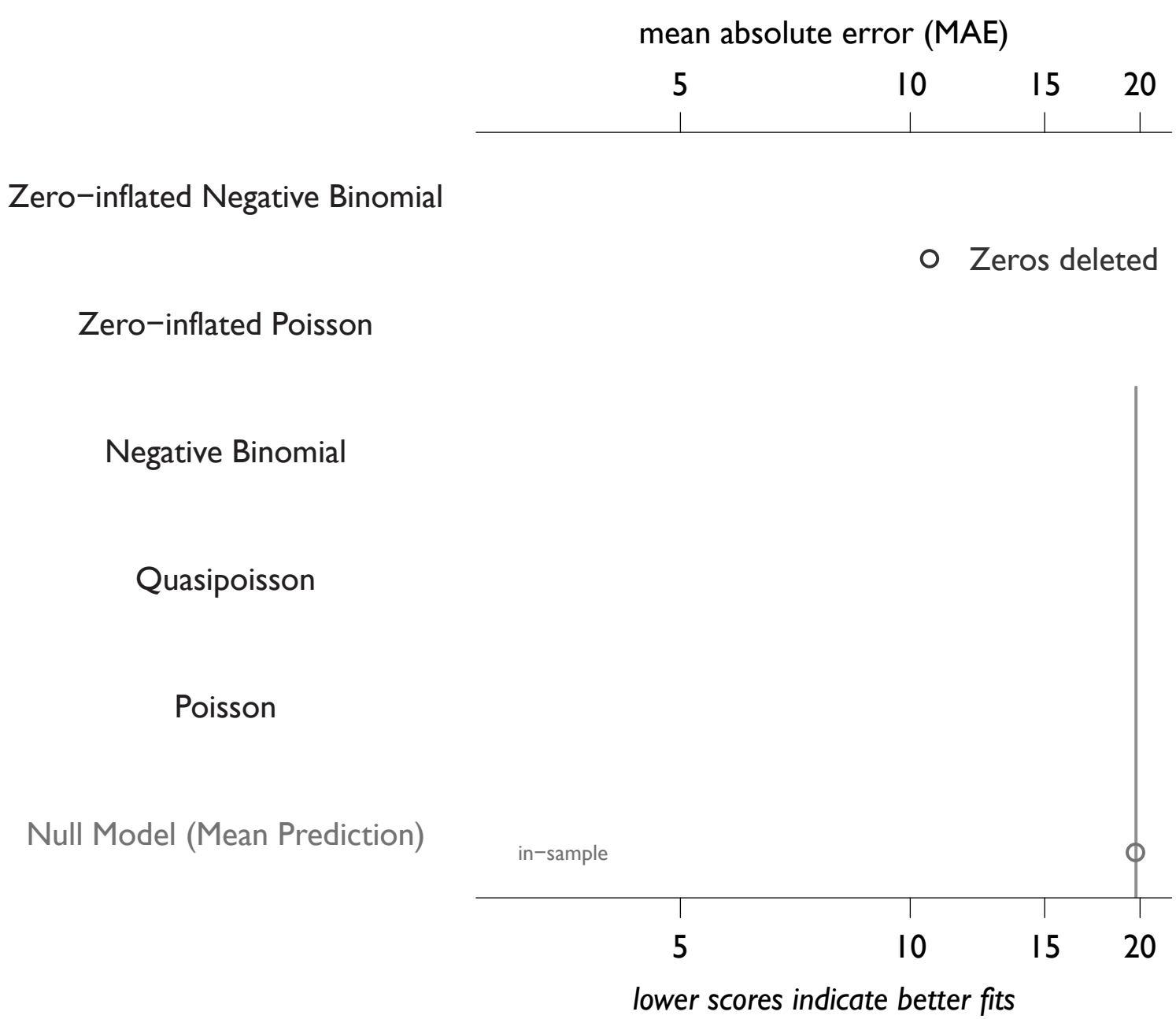




ZIP and ZINB fits look very similar, including underpredicting high counts

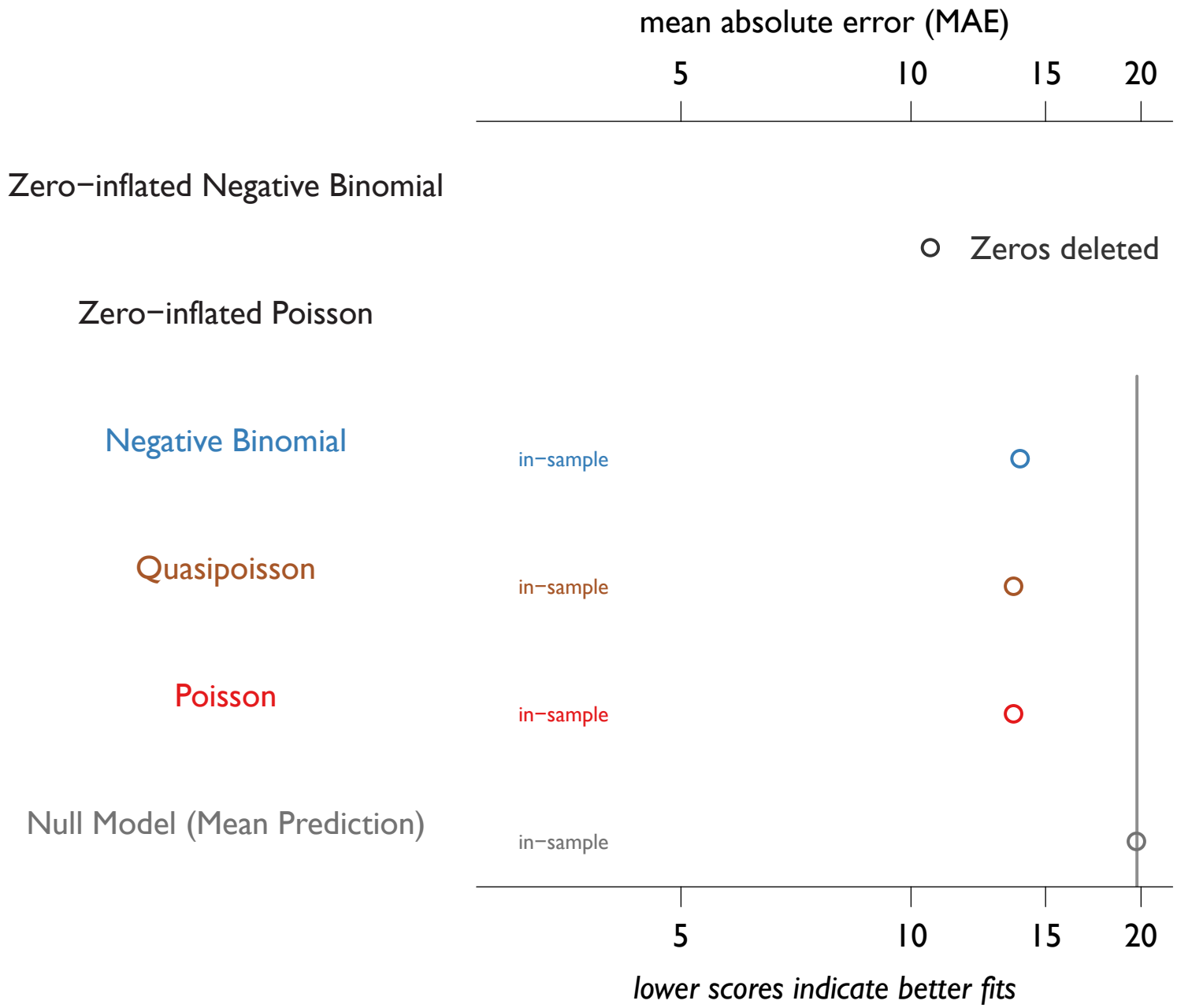
Remember: the key difference between these models lies in producing credible CIs, not in goodness of fit *per se*

That said, let's review the fit of all our models
using the single dimension of mean absolute error



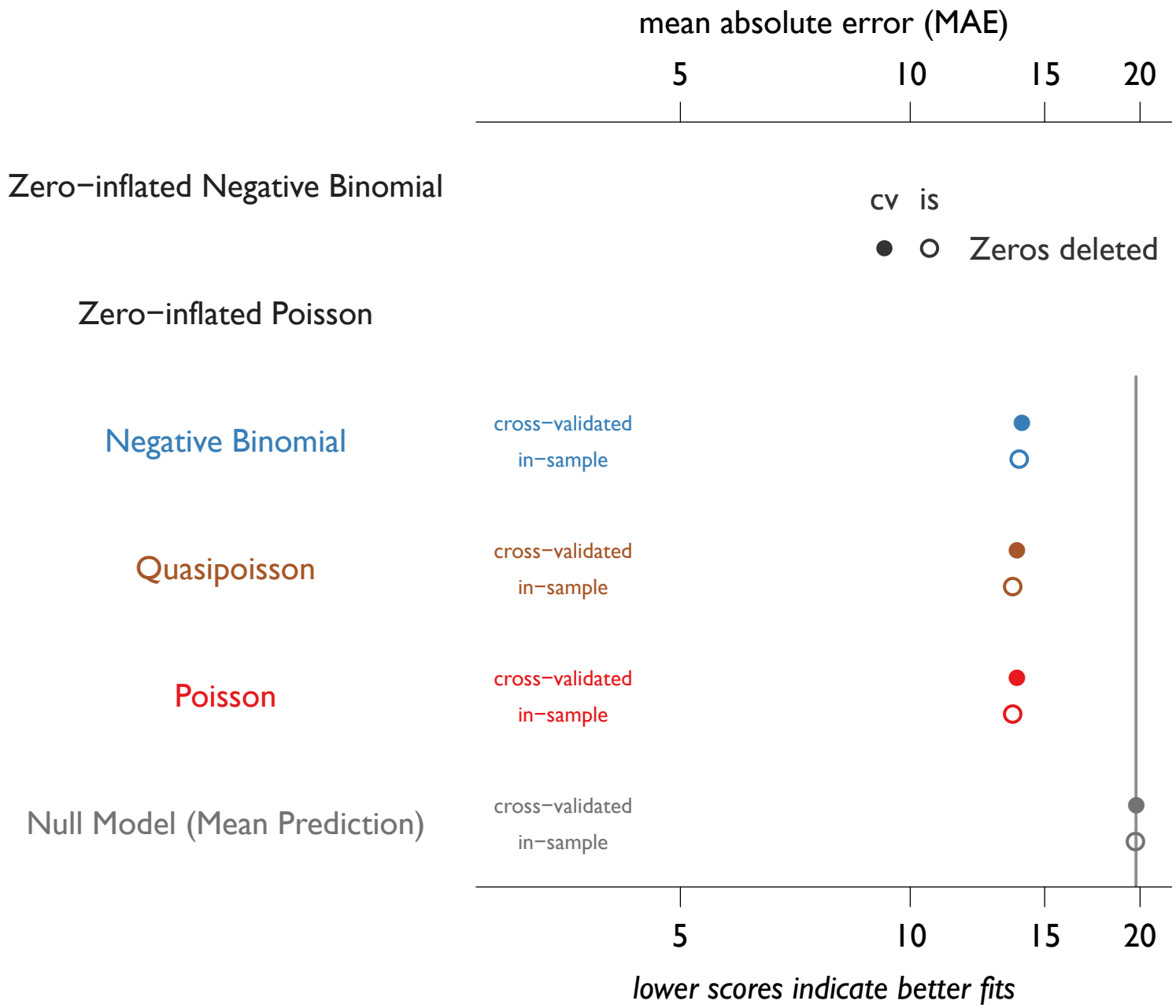
Let's start with in-sample fits for the "zeros-deleted" models

As usual, a baseline helps: how well does the mean-only model predict?



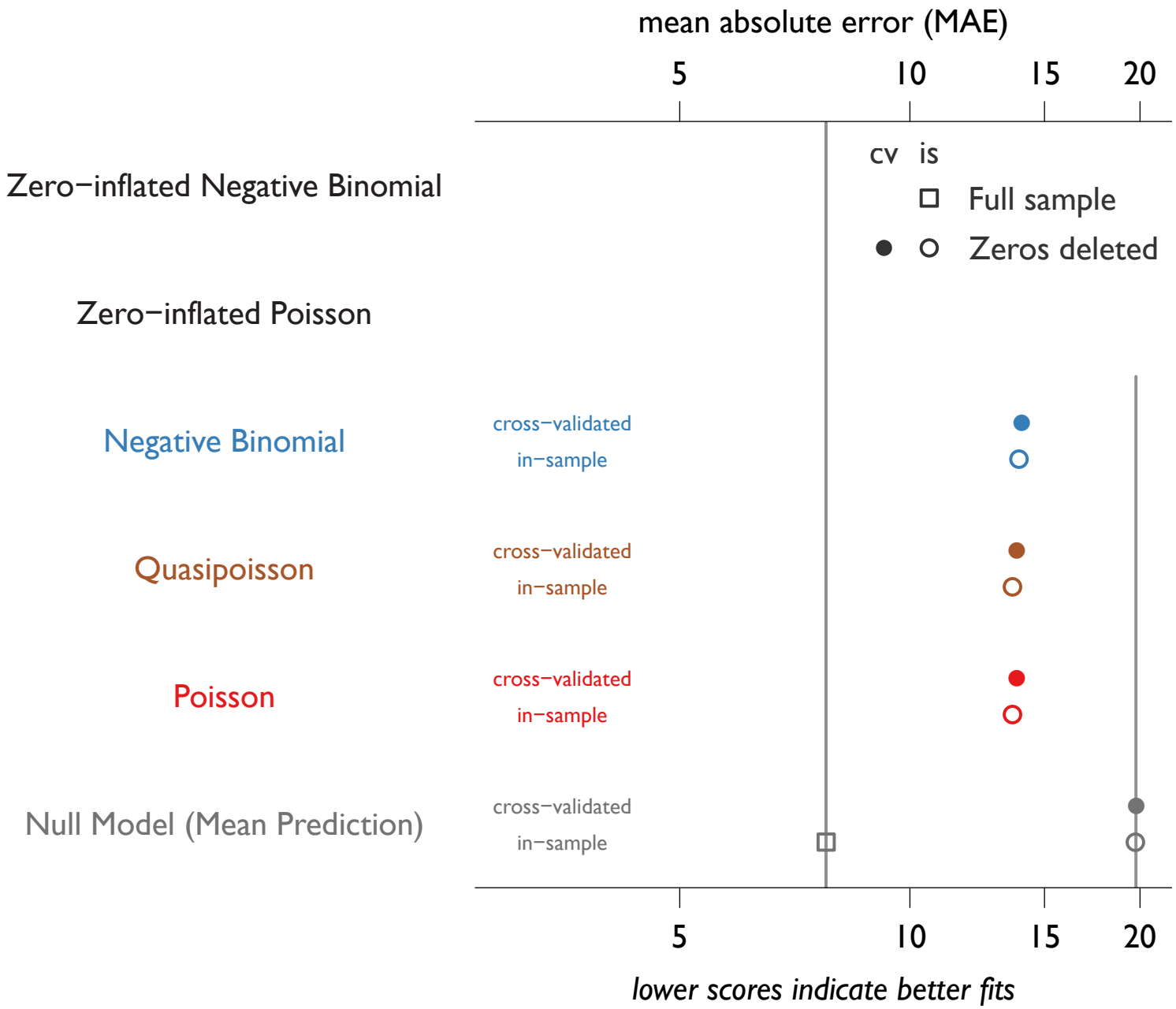
Fits for the Poisson, Quasipoisson, and Negative Binomial are similar

All improve notably on the null model in-sample



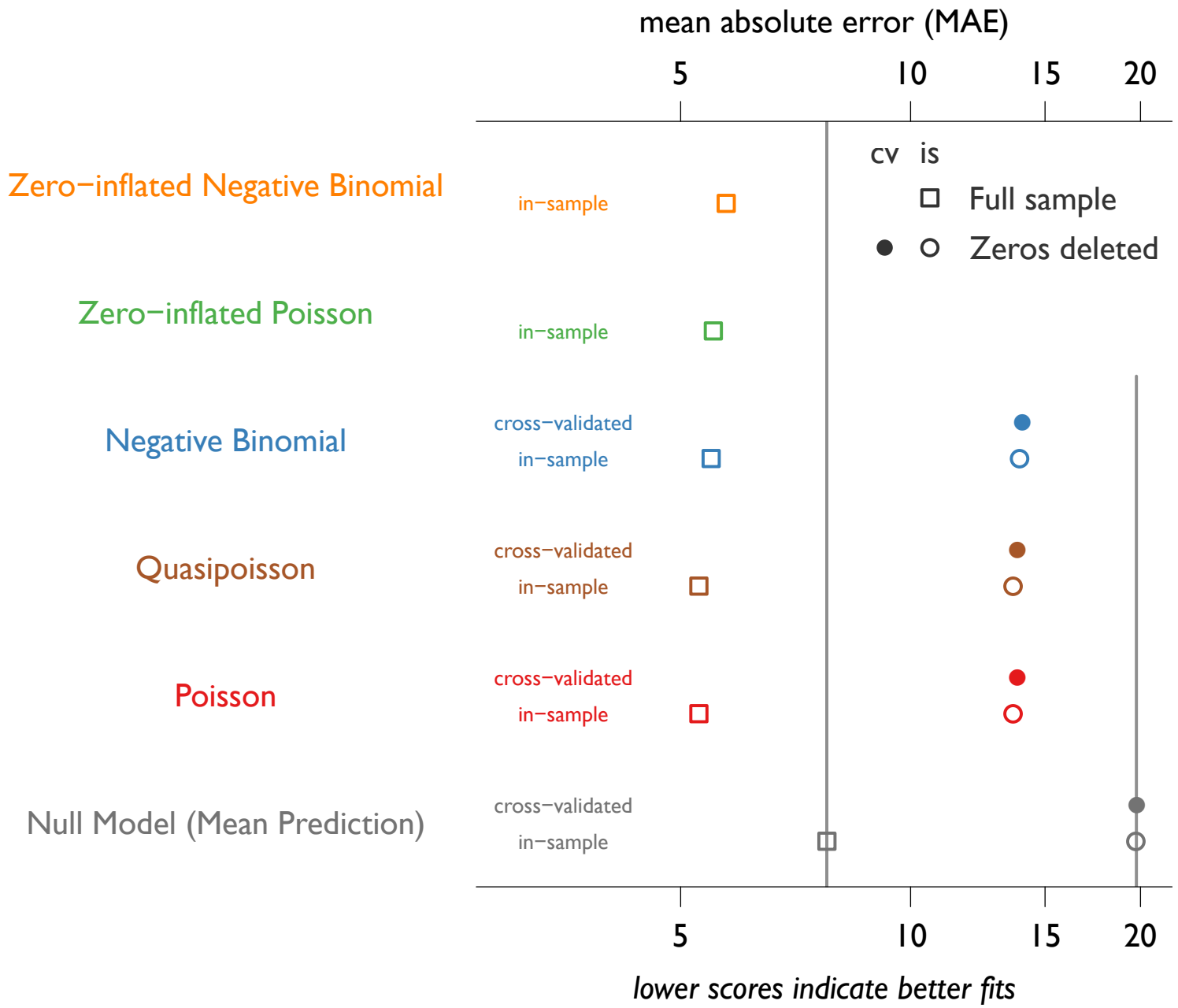
Cross-validation diminishes this only slightly

Null fits are different in-sample & in cv, but not noticeably here



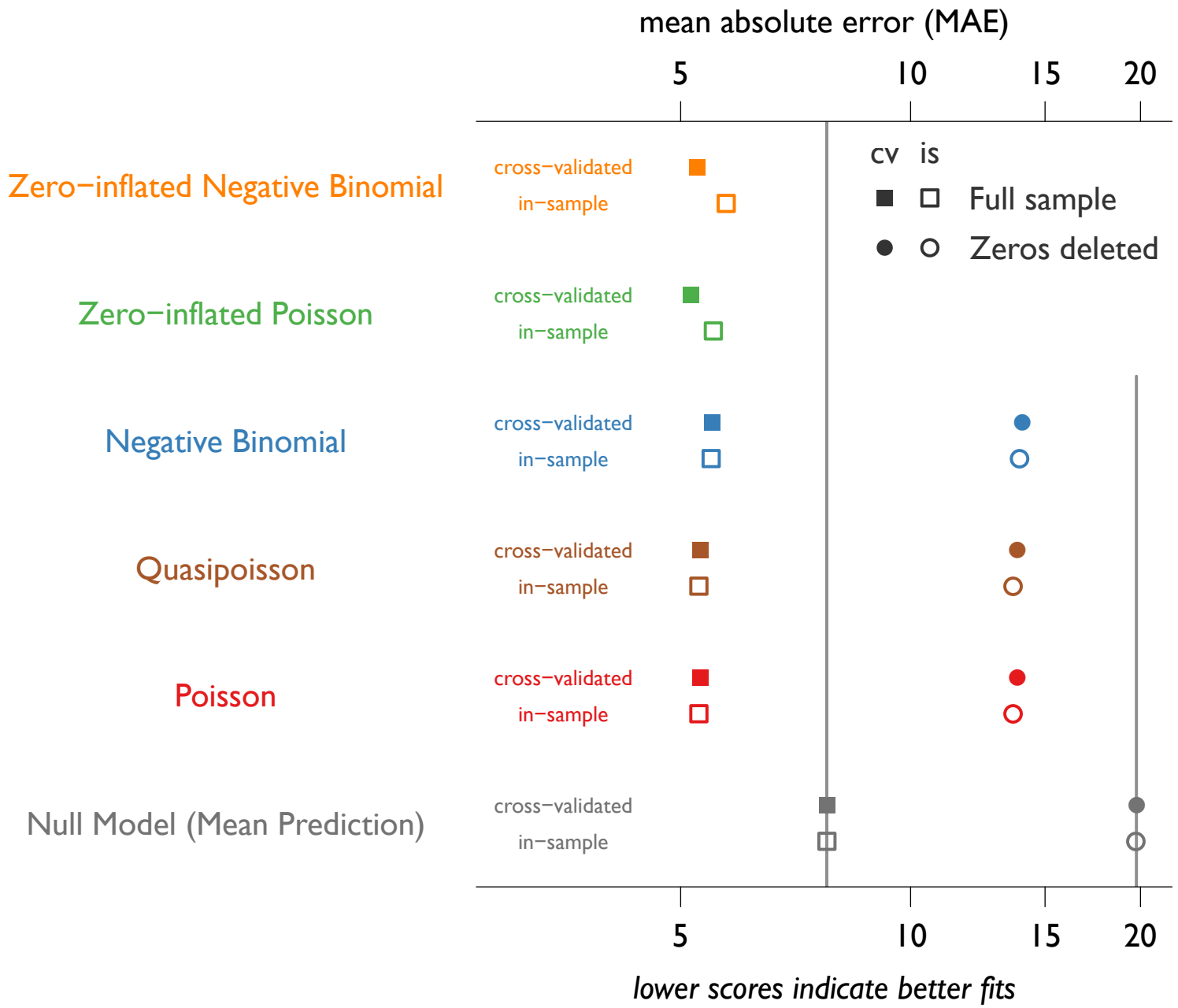
Let's look at the full-data models

As we effectively have a new dataset, we have a new baseline fit



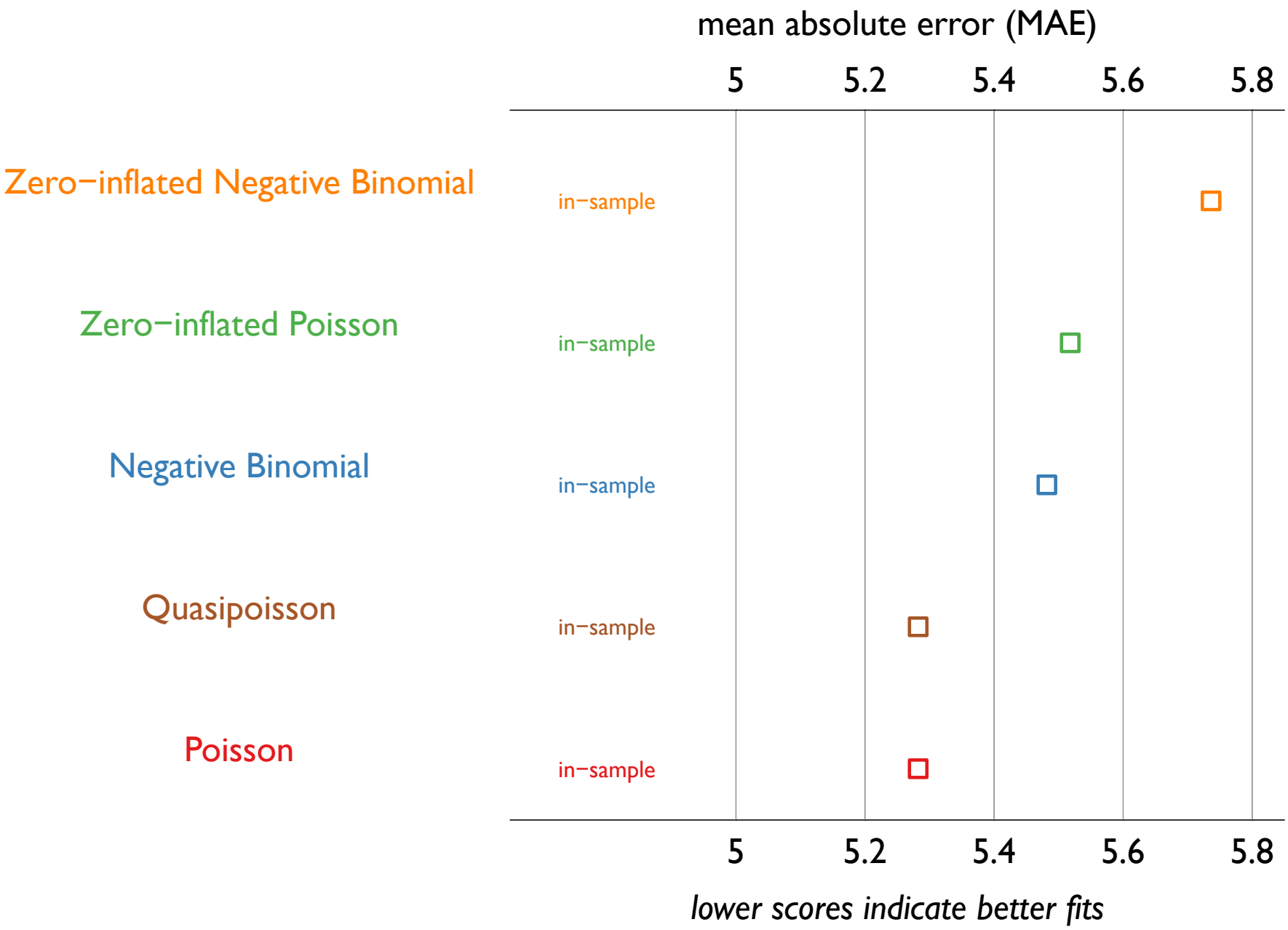
In-sample, the Poisson and the Quasipoisson fit slightly better *A surprise?*

We might still discard the Poisson itself as having untrustworthy CIs



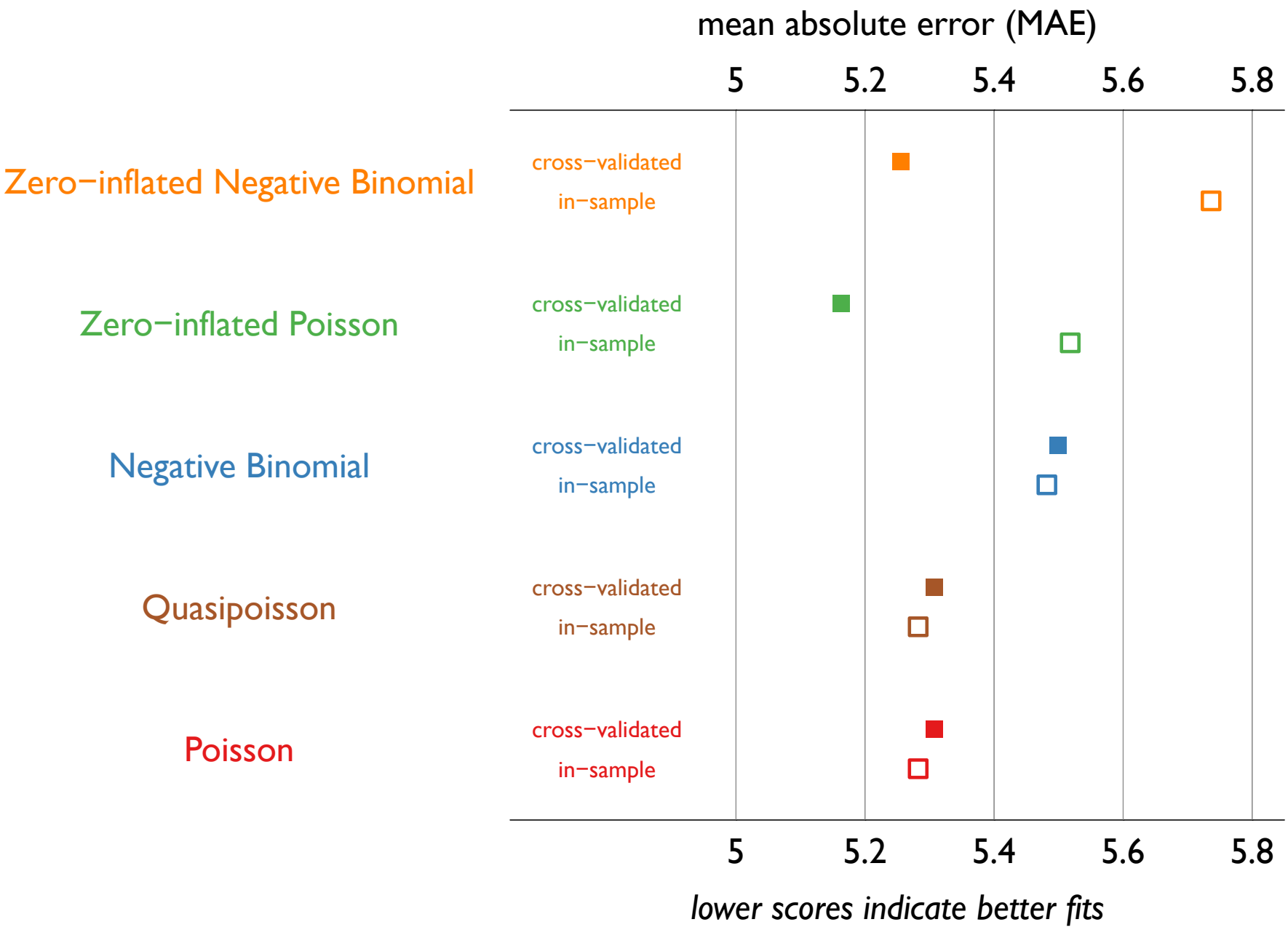
Cross-validation reveals the in-sample results are a bit misleading

Now, the zero-inflated models do best



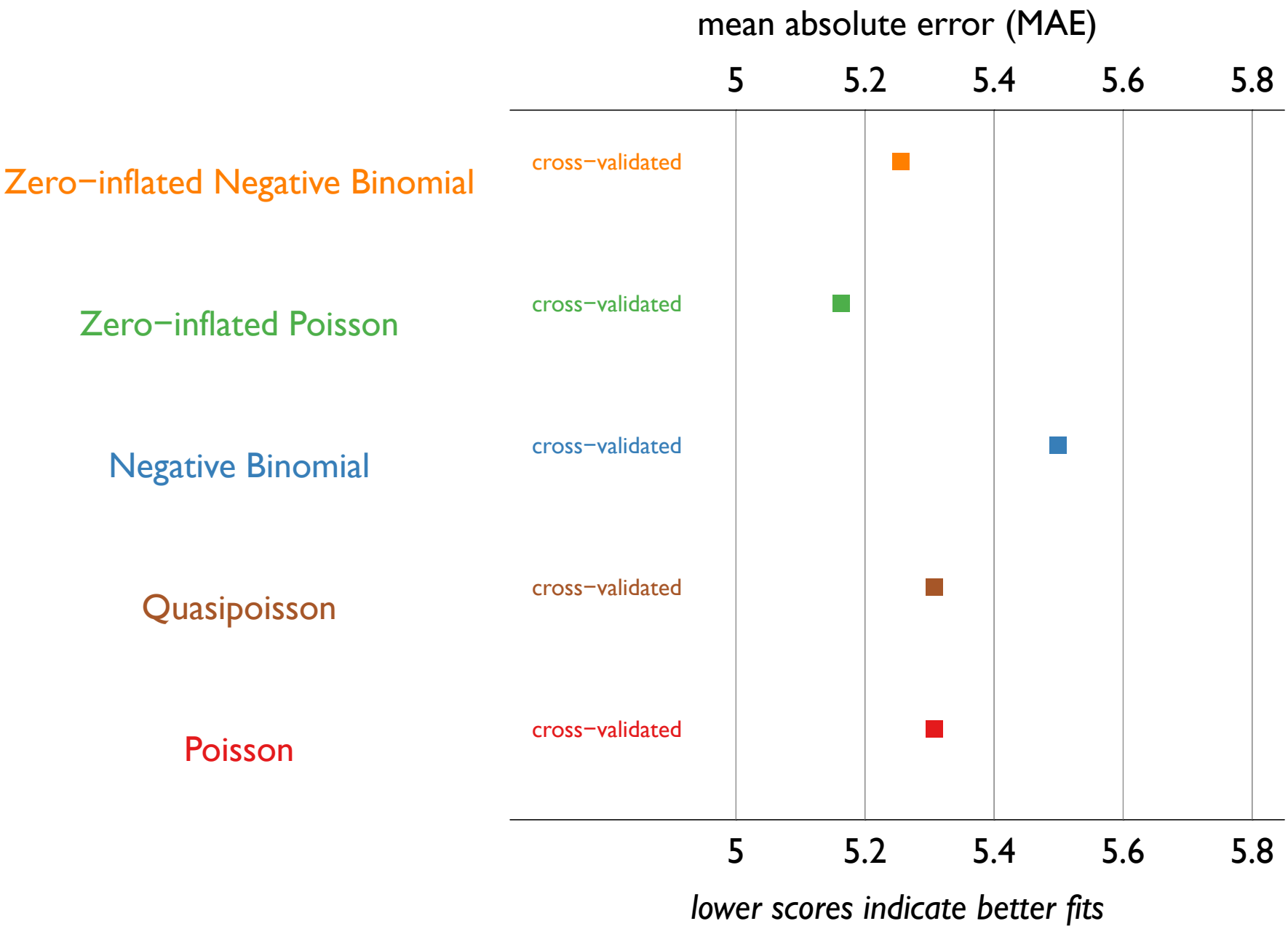
Let's zoom in to see how much difference model choice makes for MAE

We start with the in-sample fits, which have a counter-intuitive ordering



The more reliable 5-fold cross-validation improves only the zero-inflated models

Other models are slightly worse out of sample



In the end, these aren't huge differences

It is reassuring the more appropriate models did best

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial

Beta-Binomial

Quasibinomial

Poisson

Negative Binomial

Quasipoisson

Zero-inflated Poisson

Zero-inflated Negative Binomial

we've discussed eight count models in detail

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

the models we've discussed imply other models

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that assume a maximum count

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that allow for unbounded counts

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that assume independent events

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that allow events to be correlated

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that avoid distributional assumptions

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that allow higher barriers to the initial event

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models that allow some cases to be structural zeroes

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

models you might actually use for social science data

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial

does it seem like something is missing?

A TAXONOMY OF COUNT REGRESSION MODELS

Binomial	Poisson
Beta-Binomial	Negative Binomial
Quasibinomial	Quasipoisson
Hurdle Binomial	Hurdle Poisson
Hurdle Beta-Binomial	Hurdle Negative Binomial
Zero-inflated Binomial	Zero-inflated Poisson
Zero-inflated Beta-Binomial	Zero-inflated Negative Binomial
Zero-inflated Quasibinomial	Zero-inflated Quasipoisson

do zero-inflated quasilikelihood models exist?

Zero-inflated Quasipoisson models?

The quasipoisson model scaled up Poisson standard errors to match overdispersion

By doing this without distributional assumptions (just a mean & variance),
the quasipoisson was more robust to misspecification than the Negative Binomial,
but less efficient than correctly specified Negative Binomial models

It is possible to create a zero-inflated quasipoisson (ZIQP),
or does the lack of a true likelihood for the quasipoisson prevent it?

Zero-inflated Quasipoisson models?

The quasipoisson model scaled up Poisson standard errors to match overdispersion

By doing this without distributional assumptions (just a mean & variance),
the quasipoisson was more robust to misspecification than the Negative Binomial,
but less efficient than correctly specified Negative Binomial models

It is possible to create a zero-inflated quasipoisson (ZIQP),
or does the lack of a true likelihood for the quasipoisson prevent it?

Staub and Winkelmann (2012, *Health Economics*) derive such a model
and perform some Monte Carlo tests

They find:

1. The ZIQP is unidentified in the typical case where $\mathbf{x} = \mathbf{z}$,
but assuming the sign of one parameter identifies the model

2. Performance is poor relative to the ZINB in "small" samples (e.g., $N < 5000$)

3. In large samples, may provide useful robustness to misspecification

Zero-inflated Quasipoisson models?

Not available in R yet – Staub & Winkelmann offer Stata code

Would be very interesting to see if it fits the HOA data better than ZINB

Or whether the ZINB and ZIQP results would converge as the HOA dataset grows, which would help validate the Gamma-Poisson assumptions of ZINB

Do you think a hurdle quasipoisson is possible?

Zero-inflated Quasipoisson models?

Not available in R yet – Staub & Winkelmann offer Stata code

Would be very interesting to see if it fits the HOA data better than ZINB

Or whether the ZINB and ZIQP results would converge as the HOA dataset grows, which would help validate the Gamma-Poisson assumptions of ZINB

Do you think a hurdle quasipoisson is possible?

Instead of matching the mean of the poisson, λ

Match the mean of the zero-truncated poisson: $\frac{\lambda}{1 - \exp(-\lambda)}$

Concluding thoughts on count models in social science

Many social science event counts are overdispersed

They often also have mixed-in structural zeros, hurdles, or other truncation

If we could identify

- which observations came from which process, and
- which omitted variables caused the overdispersion,

we'd get more millage from simple models like the binomial and Poisson

In practice, should probably turn to models like
the Beta-Binomial for bounded counts
and the Negative Binomial for unbounded counts

Quasilikelihood models are also a good check if available

In many cases, we will need to consider zero-inflation, truncation, or other quirks

MLE provides a powerful toolkit for deriving new models for these cases