Lecture 8. Too many zeros.

We've covered models for count data, and models for binomial data. Now we'll cover a kind of model that mixes these two kinds of distribution. When analyzing count data, there is a situation that is similar to overdispersion but with important differences: too many zeros, which is often called *zero inflation*. The rationale for zero inflation is a little different than the rationale for overdispersion.

<u>The rationale for overdispersion is something like</u>:

The Poisson distribution assumes that the variance equals the mean. But variation in the mean count due to predictors that we haven't measured/modeled will show up in the data as too much variance. So we need to account for too much variance with a negative binomial distribution, or a quasi-poisson correction, or some other method.

<u>The rational for zero-inflation is:</u>

We are counting something, e.g. fish at different sites. There are multiple reasons to get a count of zero. (1) Maybe the conditions for the fish are poor, and so even though fish are in the vicinity you will see some zeros by chance. (2) Maybe the conditions for the fish are OK, but there are no fish there because they never even showed up, e.g. due to recruitment limitation or dispersal limitation. (3) Or maybe the fish were actually there but we missed them during the survey, because we are flawed creatures.

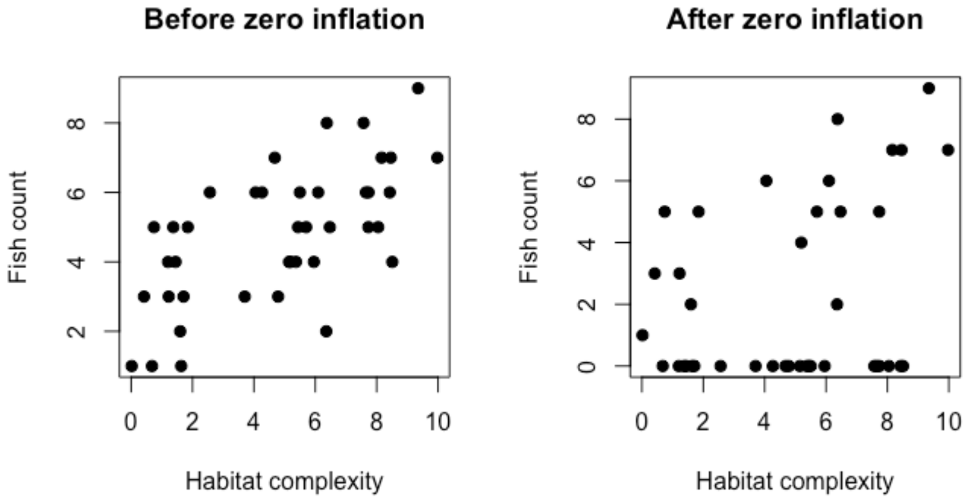
Now let's imagine that our model has predictors quantifying the different sites with variables that might predict whether they're favorable for this fish species. In this case option #1 represents the kind of zeros that our model will predict. But options #2 and #3 are going to represent 'extra' zeros that our model will not predict, because we aren't modeling the fact that our detection is imperfect (#3), or that the fish may be recruitment- or dispersal-limited (#2).

Why do we care about the extra zeros? Two main reasons. First, zero inflation is a kind of overdispersion, and therefore causes the same problems for confidence intervals and p-values. Second, if the extra zeros are caused by different processes than the rest of the data, they are going to obscure the effects we are trying to estimate. And in some cases the processes that cause the extra zeros are interesting in themselves and should be modeled. Let's look at a simulated example.

A simulated zero inflation example

Let's imagine that a reef fish tends to be more abundant at sites that offer more habitat complexity, in the sense of more nooks and crannies in which to hide from predators. I simulated some count data for what this might look like, which is the plot on the left:

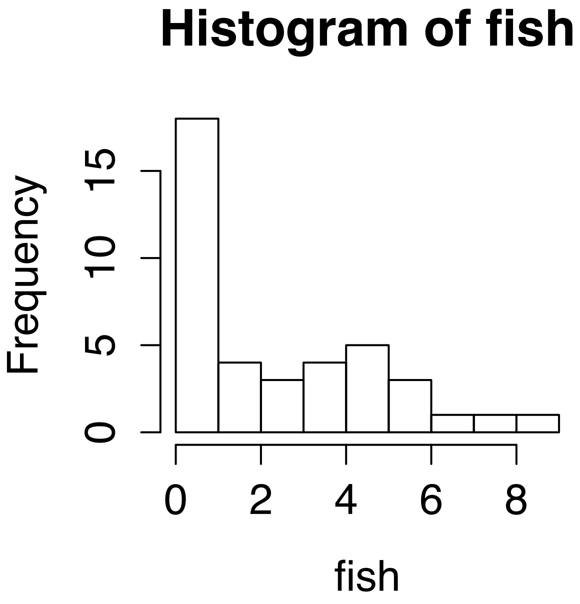
```
#simulated example of zero inflation problem
n = 40
habit = runif(n,0,10)
fish = rpois(n, exp(1 + 0.1*habit))

par(mfrow = c(1,2))
ymax = max(fish)
plot(fish ~ habit, ylab = 'Fish count', xlab = 'Habitat complexity',
pch = 19, main = 'Simulated effect of habitat complexity')
```



Then I assumed that at half of the sites, the fish simply aren't there due to other factors such as limited supply of recruits from the planktonic stage.

```
fish = fish*sample(c(0,1), size = n, replace = TRUE, prob = c(0.5,
0.5))
plot(fish ~ habit, ylab = 'Fish count', xlab = 'Habitat complexity',
pch = 19, ylim = c(0, ymax), main = 'After zero inflation')
```

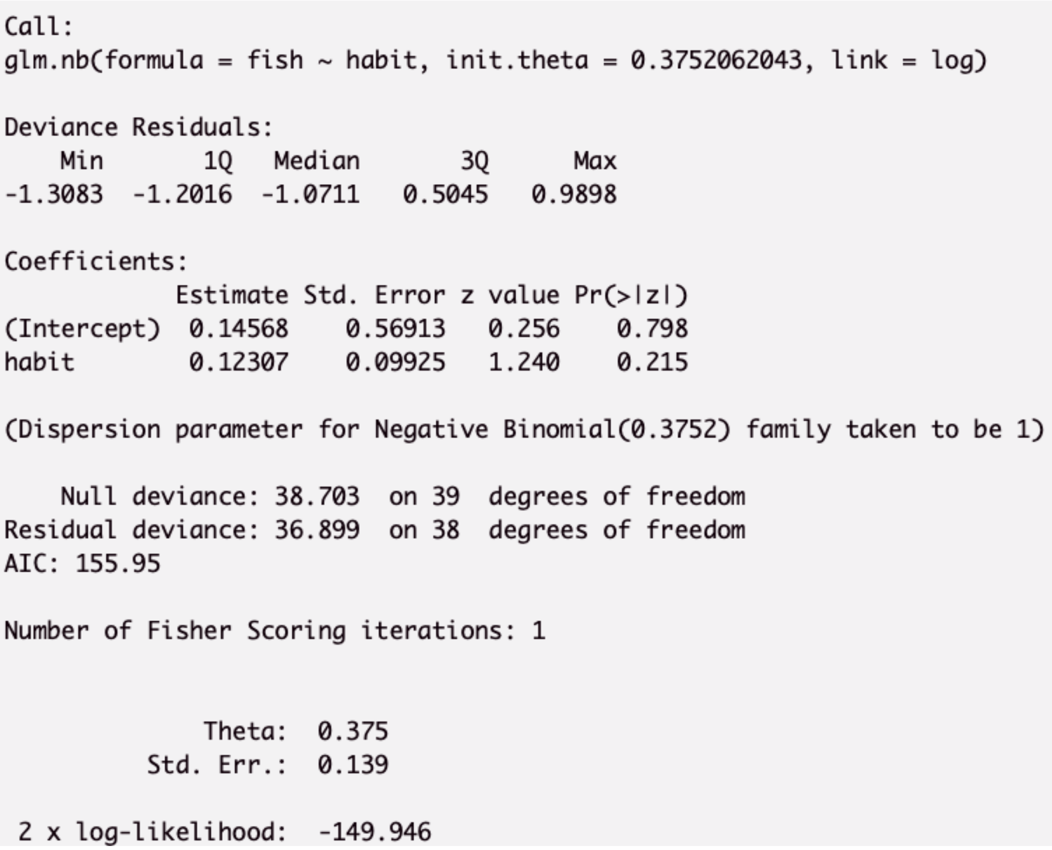
Here I've randomly turned half of the data into zeros, and that is the plot on the right. We can see that the overall pattern is still there in the positive data, but there are a bunch of zeros scattered along the x-axis. If we look at a histogram the excess zeros are clear:



This seems like a clear example of zero-inflation, because this distribution even has two peaks, i.e. it is *bimodal*, with a peak at zero and a peak around 4. However, an exploratory plot is not really sufficient to know whether zero inflation is important or not, which is a point I'll return to later.

Now let's fit the simulated zero-inflated data. If we fit the relationship between fish abundance and habitat complexity using a Poisson GLM, we will find that the data is overdispersed. So let's skip that step here and fit a negative binomial GLM that better accounts for overdispersion:

```
#negative binomial model
mod = glm.nb(fish ~ habit)
summary(mod)
```



```
Call:
glm.nb(formula = fish ~ habit, init.theta = 0.3752062043, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3083  -1.2016  -1.0711   0.5045   0.9898

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14568    0.56913   0.256    0.798
habit        0.12307    0.09925   1.240    0.215

(Dispersion parameter for Negative Binomial(0.3752) family taken to be 1)

    Null deviance: 38.703  on 39  degrees of freedom
Residual deviance: 36.899  on 38  degrees of freedom
AIC: 155.95

Number of Fisher Scoring iterations: 1


              Theta:  0.375
          Std. Err.:  0.139

 2 x log-likelihood:  -149.946
```

The parameter *theta* for the negative binomial distribution is 0.375, which indicates pretty strong overdispersion (remember that a lower theta means a greater overdispersion). The estimate for habitat complexity is close to zero, with a large standard error, and a likelihood ratio test on this model finds that the effect of habitat complexity has p = 0.17. So even though the negative binomial distribution is fitting and accounting for overdispersion, the zero-inflation has basically erased the signal of habitat complexity from the data.

The zero-inflated model

Now let's compare with a model that better accounts for the extra zeros. This will be a *zero-inflated poisson* model. The details of the model are somewhat challenging, but the logic is fairly straightforward. The model has two parts:

1) Binomial part. There is some binomial probability π_i that count *i* is an 'extra' zero, e.g. because of recruitment limitation or some similar cause. These zeros are often referred to as 'false zeros', because they are not generated from the Poisson distribution of the counts. In the fish example we are imagining that the zeros have a cause (recruitment limitation), but that cause is distinct from the fact that the Poisson distribution will sometimes gives you a zero when mean abundance is low. So these are extra zeros beyond what the Poisson will produce.

To model the probability π_i, we will use the standard binomial GLM that we've already covered, i.e. a logistic regression / logit link function. We might want to include predictors to test whether the observing an extra zero is predictable. But in this case we just want to model how many extra zeros there are, and so we'll just include one parameter for the mean probability of seeing an extra zero:

$$\text{logit}(\pi_i) = \beta_{00}$$

2) Poisson part. If we don't observe an extra zero, then count *i* is drawn from a Poisson distribution. This will occur with probability $(1 - \pi_i)$. The mean of the Poisson is determined by whatever predictors we want, in standard GLM format. In this case the predictor is habitat complexity, i.e.

$$\log(\mu_i) = \beta_{10} + \beta_{11} * \text{habit}_\text{i}$$

Importantly, *this draw from the Poisson distribution could still be zero*. But this zero is derived from the Poisson counting process, not from the binomial trial that represents zero inflation.

This kind of model is called a *mixture model*, because two probability distributions are combined to produce the observed data. In particular, the zeros have two sources. There could be a zero due to recruitment limitation, and these zeros are binomially distributed with probability π_i and *n* = 1. In addition, even when there isn't recruitment limitation we could observe a zero just because mean abundance is low.

I won't write out the full formula that defines this model, because it is a bit tricky, but think of it as specifying a binomial GLM and a Poisson GLM at the same time, and the maximum likelihood estimate for the model will estimate what proportion of the zeros come from the 'false' binomial part.

Let's look at what the syntax for a zero-inflated model looks like:

```r
library(glmmTMB)
mod.z = glmmTMB(fish ~ habit, ziformula = ~ 1, family = "poisson")
```

I'm using the package glmmTMB, which handles zero-inflated models well, in addition to GLMMs, which we will cover later. The syntax is similar to lm() and glm(), with a few differences. The first formula is for the count model, i.e. the non-zero counts and non-inflated zeros. The second formula, 'ziformula', is where we put predictors for the extra zeros. In these case I have not put any predictors, which means extra zeros occur with some constant probability, to be estimated from the data. And I set family = ″poisson″ for the count data part of the mixture distribution. The negative binomial is also an option, which we'll look at later.

Here’s what summary() returns:

```
##  Family: poisson  ( log )
## Formula:          fish ~ habit
## Zero inflation:        ~1
##
##      AIC      BIC   logLik deviance df.resid
##    130.5    135.6    -62.3    124.5       37
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.01317    0.24767   4.091  4.3e-05 ***
## habit        0.10604    0.03657   2.900  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1695     0.3228   0.525      0.6
```

So we have two sets of coefficients, one for the count model (poisson with log link), and one for the extra zeros (binomial with logit link). The count model is called ‘conditional’, because these coefficients are used to calculated the expected value of the response variable, conditional on the value(s) of the predictor(s). Let’s think about the coefficient for the binomial model. There is just one, the Intercept, which is quantifying the probability the count *i* is an extra zero. The coefficient estimate is ~0.17. Recall that with the logit link, the coefficients can be interpreted as the *log-odds*. So this is saying that the mean log-odds of observing an extra zero is 0.17. Which is equivalent to saying that the *odds* of observing an extra zero is exp(0)=1.2. So the odds are roughly 1:1 that we will observe an extra zero vs. observing an actual count drawn from the Poisson. These are indeed the odds that I used to simulate the data (I randomly turned half the data to zero), so that’s reassuring.

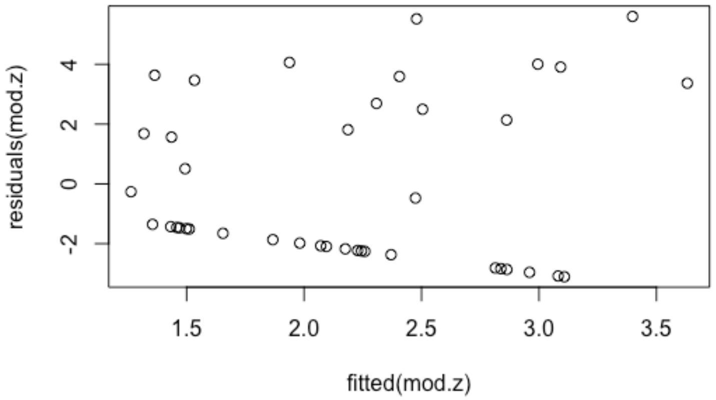
The count model coefficients say that the count increases with habitat complexity, and the z-test is significant. I’d prefer to do a likelihood ratio test, because it is more accurate than the z-test. To do this in this case, we can fit a second model lacking the habitat predictor and compare them with anova():

```
mod.z0 = glmmTMB(fish ~ 1, ziformula = ~ 1, family = "poisson")
anova(mod.z, mod.z0)
```

```
## Data: NULL
## Models:
## mod.z0: fish ~ 1, zi=~1, disp=~1
## mod.z: fish ~ habit, zi=~1, disp=~1
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod.z0  2 137.56 140.93 -66.778   133.56
## mod.z   3 130.52 135.59 -62.262   124.52 9.0327      1   0.002652 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

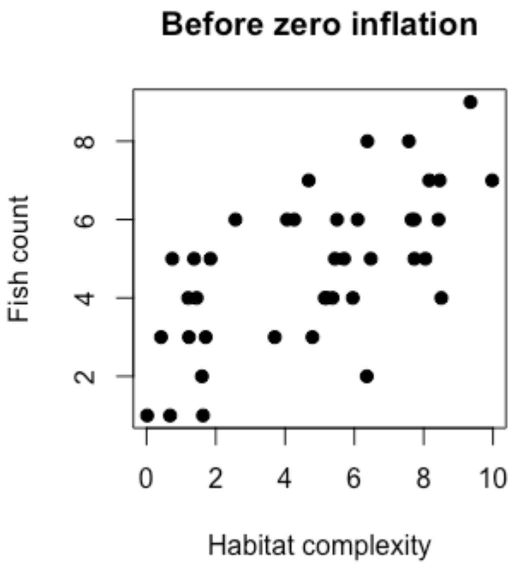
We can also look at residual plots. As for simpler GLMs, we want to look at residuals vs. fitted values, and residuals vs. predictors. In this case there is just one predictor, so we'll just look at residuals vs. fitted values.

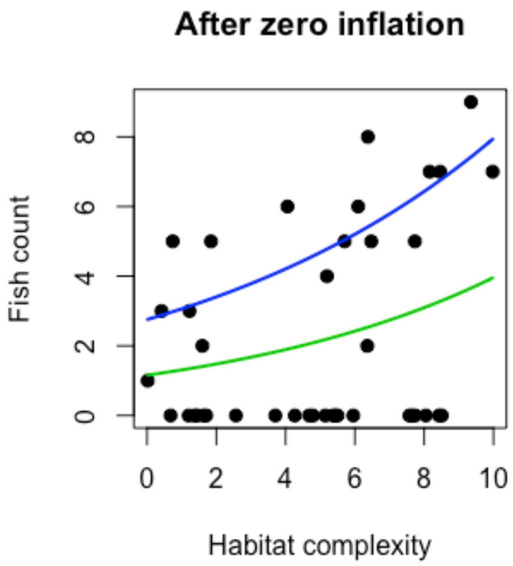
```
plot(residuals(modz) ~ fitted(modz))
```



No obvious patterns, which is good.

Let's review what we've learned so far from this example. Because that data is simulated, we know there's a real effect of habitat complexity, and also a bunch of extra zeros that are not related to habitat complexity. This causes the data to be overdispersed, so one option is to fit a negative binomial GLM. But this model finds no relationship between fish abundance and habitat complexity, so it appears that the zeros have swamped the signal we're testing for. In contrast, the zero-inflated model correctly accounts for the fact that there are a bunch of extra zeros, and is therefore able to distinguish the pattern in the actual count data from the extra zeros. We can plot the two models to see this difference:



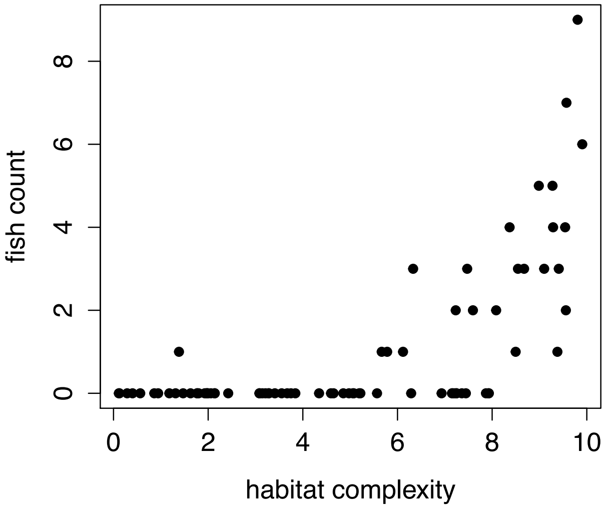


The green line shows the negative binomial fit, and the blue line shows the fit for the count model portion of the zero-inflated model. It seems almost like magic that the zero-inflated model can do this, but it's all based on the fact that the Poisson distribution is expected to generate a certain number of zeros, depending on the mean of the distribution, and so the extra zeros can be inferred relative to that expectation.

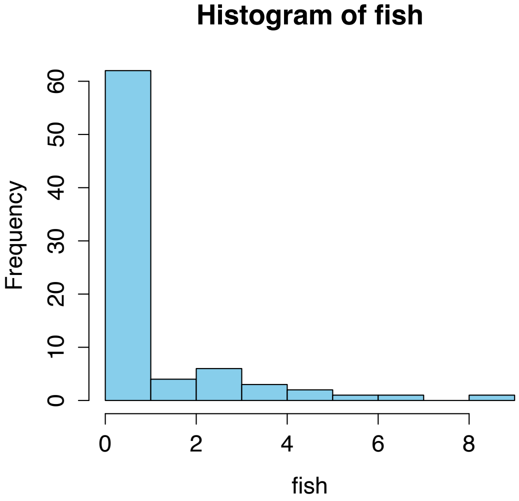
Lots of zeros doesn't always mean zero inflation

How do we know when to use a zero-inflated model? This is a fairly subtle question for which there is no simple algorithm or test. For example, in the Zuur et al. chaper on zero-inflated models, they say "make a frequency plot of the data [i.e. a histogram] and you will know whether there is zero inflation". In contrast, David Warton has a paper titled "Many zeros does not mean zero inflation…". Let's look at a simulated example where data might appear to be zero-inflated, but is not. I've taken the same code for the fish-habitat complexity example and made the relationship steeper, but haven't added any extra zeros:

```r
n = 80
habit = runif(n,0,10)
fish = rpois(n, exp(-5 + 0.7*habit))

plot(fish ~ habit, xlab = 'habitat complexity', ylab = 'fish count',
pch = 19)
```

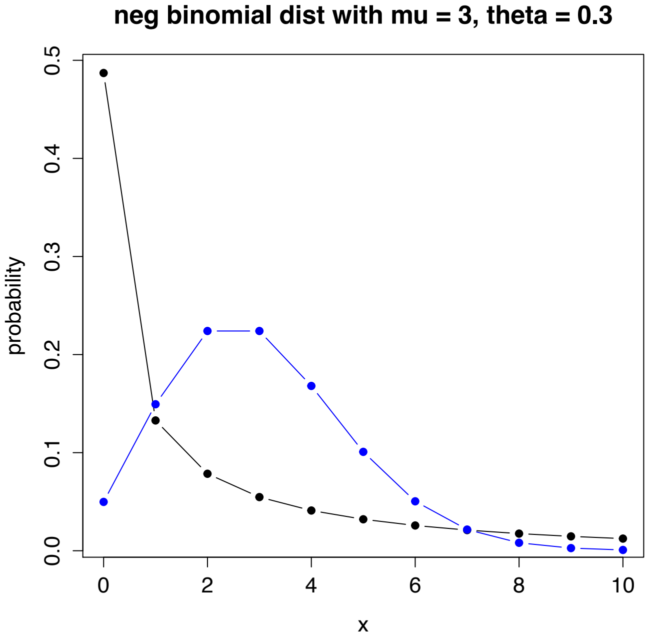


This data has a lot of zeros, but they are caused by a strong relationship with a predictor.



In this case if we just looked at a histogram of the raw data it might look like zero-inflation, but we know this data is Poisson-distributed once you account for the predictor, because that's how I created it. Likewise, this data isn't even overdispersed.

Another way data can look zero-inflated is if the data is just highly overdispersed. For example, a negative binomial distribution with a low mean and a low value of theta will have a lot of zeros:



Here I'm comparing a negative binomial with mean=3 and theta=0.3 to a Poisson with mean=3. The negative binomial has many more zeros, but not because there is some second process causing extra zeros *per se*. Rather, some underlying heterogeneity is causing overdispersion relative to the Poisson, and this changes the whole shape of the distribution, including more zeros and the loss of the peak around 2-3.

Something like a protocol for count data, overdispersion, and zero inflation

At this point we have a lot of options for analyzing count data. Poisson is the default, and quasi-Poisson and negative binomial can account for overdispersion. Zero-inflation is a special kind of overdispersion where the extra zeros may exhibit different patterns compared to the rest of the data. And in a zero-inflated model, the count data itself can be modeled with a Poisson or a negative binomial distribution. How can we sort out all these possibilities for analyzing count data? I think the best strategy is to combine some careful thinking with comparison of multiple models. For some kinds of data you may think that zero-inflation is likely, even before you look at the data. Here are some possibilities:

- Observer error, e.g. using bird count data from amateur birders who may mistake a rare species for a common species.
- Other kind of imperfect detection will lead to similar zero-inflation issues
- A species that is patchily distributed, such that they are absent from many areas that seem otherwise habitable (e.g. the fish recruitment limitation example)

For these kinds of data it is definitely worth considering zero inflation, and it is plausible that the extra zeros will be driven by different processes than the actual counts. This is probably the most useful feature of zero-inflated models: the extra zeros can be driven by different processes than the count data, and separating these components can help reveal the processes driving the counts.

So the situations listed above are candidates for zero inflation, but it is often unclear whether a given dataset really exhibits this issue or not. In this case, a process of model selection is probably the best route.

Here's a protocol that seems defensible (though be aware that there is a lot of different advice in the literature):

1. Fit your model with a Poisson distribution. See if overdispersion is important, using the pearson residuals formula, or by fitting a quasipoisson or negative binomial model.

- If there is no overdispersion, then strong zero-inflation is unlikely, and you can just use the Poisson. (Although if you have a strong *a priori* reason to expect zero-inflation, you still might want to fit that model, because if the counts are low on average then extra zeros will not cause very strong overdispersion.)
- If there is overdispersion, go to #2

2. Because there is overdispersion, at minimum you will need to use a negative binomial or quasipoisson approach. You will also need to decide how much you are concerned about zero inflation. The quasipoisson and negative binomial both 'account' for overdispersion, in slightly different ways, and so they should yield results that do not have artificially small p-values. But if you think the extra zeros may be caused by different processes than the rest of the count data, then the extra zeros may be obscuring the patterns you want to quantify.

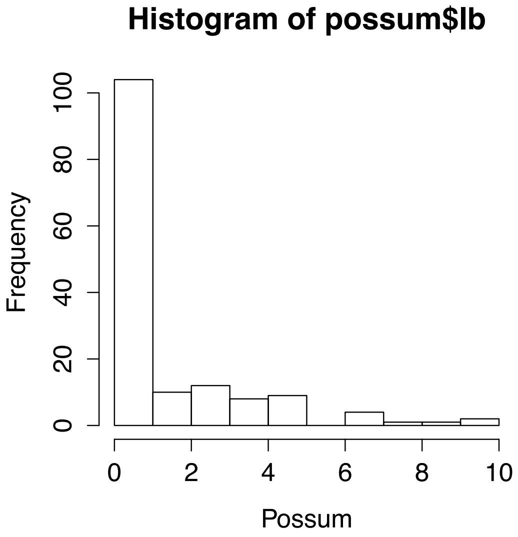
3. If you want to test for zero-inflation, fit a zero-inflated Poisson model. It is possible that the count data is still overdispersed, even after accounting for extra zeros. You can calculate a dispersion parameter using the pearson residuals from the zero-inflated Poisson. If there is overdispersion, fit a zero-inflated negative binomial model.

4. How can we test whether the zero-inflated Poisson or zero-inflated negative binomial is a better model than a model with no zero-inflation? We can't test this with a likelihood ratio test, because the zero-inflated models are mixture models with two probability distributions, and therefore a negative binomial model isn't just a special case of a zero-inflated model. Instead, we can compare the models with AIC.

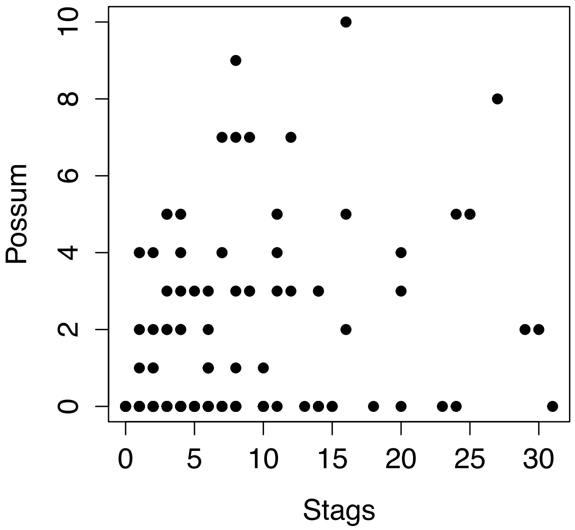
In future lectures we will learn about AIC (Akaike Information Criterion) in depth. For now you need to know a few things: AIC is calculated using the likelihood of a model; it quantifies how well a model fits the data, in a certain sense; a lower AIC indicates a better model; AIC is on a relative scale, not an absolute scale, so just look at the *difference* in AIC between models.

Zero-inflated example

I'll illustrate this protocol with an example. This data is from surveys of Leadbeater's possum, which is a rare and endangered Australian marsupial. A variety of environmental factors were measured, but let's focus on 'stags', which are trees with hollows that can act as nest sites. Data for a rare species is likely to have a lot of zeros, because the species will be absent from many sites, and because rare species are easier to miss in surveys. A histogram of the counts shows a lot of zeros:



The relationship with stags shows a potentially positive relationship, but also a lot of zeros along the x-axis that may obscure this pattern:



First I'll fit a Poisson model, and quantify the dispersion parameter:

```
#poisson model
mod.p = glmmTMB(lb ~ stags, data = possum, family = 'poisson')
#calculate dispersion parameter
sum(residuals(mod.p, type = "pearson")^2)/(nrow(possum) - length(coef(m
od.p)))

## [1] 3.247506
```

Looks like there is a lot of overdispersion. I could try to account for this with a negative binomial model, but I already suspect zero-inflation, so let's go straight to a zero-inflated Poisson model:

```
#zero-inflated poisson model
mod.pz = glmmTMB(lb ~ stags, ziformula =~ 1, data = possum, family = 'p
oisson')
summary(mod.pz)

##  Family: poisson  ( log )
## Formula:          lb ~ stags
## Zero inflation:      ~1
## Data: possum
##
##      AIC      BIC   logLik deviance df.resid
##    436.4    445.5   -215.2    430.4      148
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.007267   0.127426   7.905 2.69e-15 ***
## stags       0.025130   0.008767   2.866  0.00415 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4556     0.1746   2.608   0.0091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient for zero-inflation is 0.456, and that is the log-odds of a particular sample being an ‘extra’ zero. So the odds is exp(0.456) = 1.58. So this model thinks there are a substantial number of extra zeros. It also looks like the effect of stags is signficant, though that should be properly tested with a likelihood ratio test. It’s also possible that the counts are still overdispersed after accounting for extra zeros, which can be dealt with by using a negative binomial distribution:

```
#zero-inflated negative binomial model
mod.nbz = glmmTMB(lb ~ stags, ziformula =~ 1, data = possum, family = '
nbinom2')
summary(mod.nbz)
```

```
##  Family: nbinom2  ( log )
## Formula:          lb ~ stags
## Zero inflation:      ~1
## Data: possum
##
##      AIC      BIC   logLik deviance df.resid
##    433.7    445.8   -212.9    425.7      147
##
##
## Overdispersion parameter for nbinom2 family (): 5.91
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.89224    0.18765   4.755 1.99e-06 ***
## stags        0.03149    0.01300   2.423   0.0154 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3645     0.1974   1.846   0.0648 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ‘overdispersion parameter’ is telling us that the negative binomial has substantially more variance than the Poisson. Otherwise the coefficient estimates look similar but not identical to the zero-inflated Poisson.

I'd like to know whether the zero-inflated model(s) are really a better fit to the data than a model like the negative binomial that just accounts for overdispersion but not zero-inflation. To make this comparison I'll fit a negative binomial model:

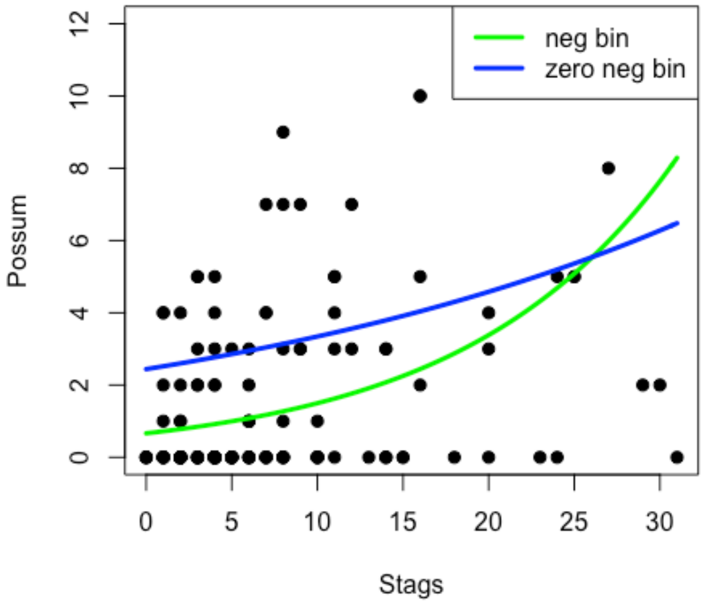
```
#negative binomial model
mod.nb = glmmTMB(lb ~ stags, data = possum, family = 'nbinom2')
```

And then I'll compare the various models with AIC:

```
AIC(mod.p)

## [1] 604.6294

AIC(mod.nb)

## [1] 447.7667

AIC(mod.pz)

## [1] 436.4344

AIC(mod.nbz)

## [1] 433.7432
```

The negative binomial model (mod.nb) has much lower AIC than the Poisson model (mod.p), which makes sense because there is a lot of overdispersion. The zero-inflated Poisson (mod.pz) is substantially lower than the negative binomial model, indicating that accounting for zero-inflation yields a better fit than just accounting for extra variance. The zero-inflated negative binomial (mod.nbz) is slighly smaller than the zero-inflated Poisson, indicating that there is still some extra variability in the data that is better accounted for by this model. We'll talk more in the future about how to interpret these numbers.

Now that we know which model is best, how do we interpret it? We can plot the fitted relationship for the count part of the zero-inflated negative binomial, and we can also compare to the negative binomial model without zero inflation:



For both models, the abundance of possums increases when there are more stags. Interestingly, the zero-inflated model actually shows a shallower slope. This is the opposite of the simulated reef fish example I showed earlier. So that means that in this case, properly accounting for zero-inflation reveals that the relationship is weaker than it might otherwise seem. Why would that be? Perhaps the probability of seeing extra zeros is also related to the availability of stags. We can put this into the model:

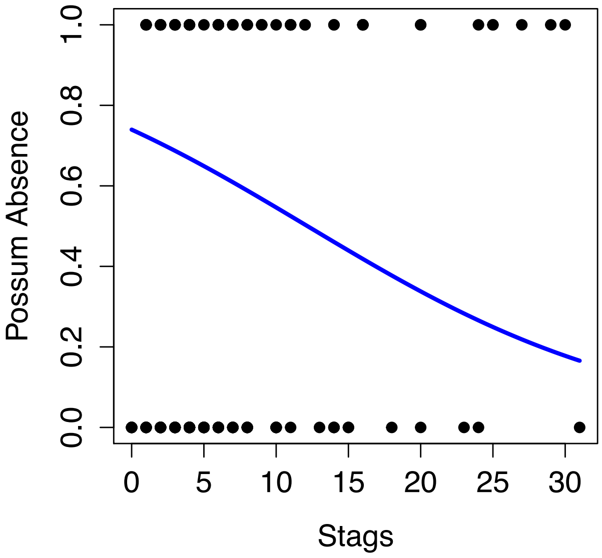
```
mod.nbz2 = glmmTMB(lb ~ stags, ziformula =~ stags, data = possum,
family = 'nbinom2')
```

Now the zero-inflated part of the model also has a predictor, and it is also ‘stags’.

```
summary(mod.nbz2)

##  Family: nbinom2  ( log )
## Formula:          lb ~ stags
## Zero inflation:      ~stags
## Data: possum
##
##      AIC      BIC   logLik deviance df.resid
##    426.1    441.2   -208.1    416.1      146
##
##
## Overdispersion parameter for nbinom2 family (): 7.02
##
## Conditional model:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.96854    0.16442   5.891 3.85e-09 ***
## stags        0.02641    0.01198   2.204   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.04436    0.29195   3.577 0.000347 ***
## stags       -0.08584    0.02995  -2.866 0.004157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks like the zero-inflation is indeed related to the availability of stags. A likelihood ratio test would show that this effect is significant (not shown). What does this coefficient mean? It is important to note that the coefficients for the zero-inflated part of the model are describing *the probability of observing an extra zero*. So the estimate of -0.0858 for 'stags' means that *the probability of observing an extra zero declines as stags increases*. We can plot this effect, and I will plot it against some raw data indicating whether possums are totally absent (zero) or not:



The raw data isn't super informative here, but the fitted relationship is fairly strong, with a much smaller chance of getting an extra zero when stags are plentiful.

OK let's take a step back and think about how one might interpret these results. We now know that the data are zero-inflated as well as mildly overdispersed, at least with stags as the only predictor. And we know that 1) possum are more abundant when there are more stags, and 2) we are less likely to see an 'extra' zero when there are more stags. To me this suggests that we can think of the distribution of possums in two parts: under what conditions are they even present? And if they are present, how abundant are they? And in this case it looks like trees with hollows makes it more likely that stags are present, and also makes possums more abundant when they are present. In this case the predictor we looked at has a similar effect on both presence/absence as well as abundance, but in other situations this may not be true, and using a zero-inflated model allows us to better parse these effects.