Modeling count data with overdispersion

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

18 May 2020

Goals for today

- Understand the importance and source of overdispersion in Poisson models
- Understand how to assess overdispersion in count data
- Understand the options for modeling overdispersed binomial data
- Understand the pros & cons of the modeling options

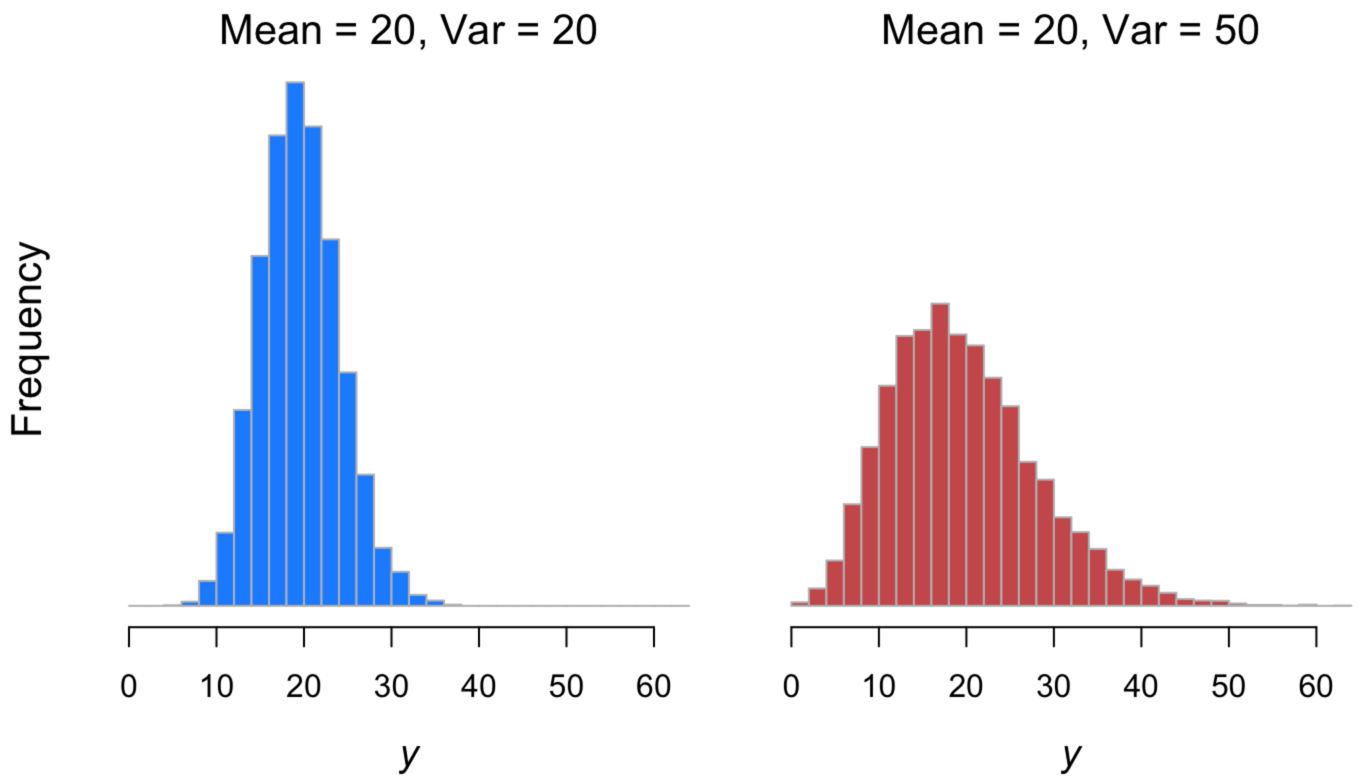
Overdispersion in counts

We saw that logistic regression models based upon the binomial distribution can exhibit overdispersion if the deviance is larger than expected

Poisson regression models are prone to the same because there is only one parameter specifying both the mean and the variance

$$y_i \sim \text{Poisson}(\lambda)$$

Overdispersion in counts



Bycatch of green sea turtles

Bycatch of sea turtles in trawl fisheries has been a conservation concern for a long time

To reduce bycatch, some trawls have been outfitted with turtle excluder devices (TEDs)

Turtle excluder device

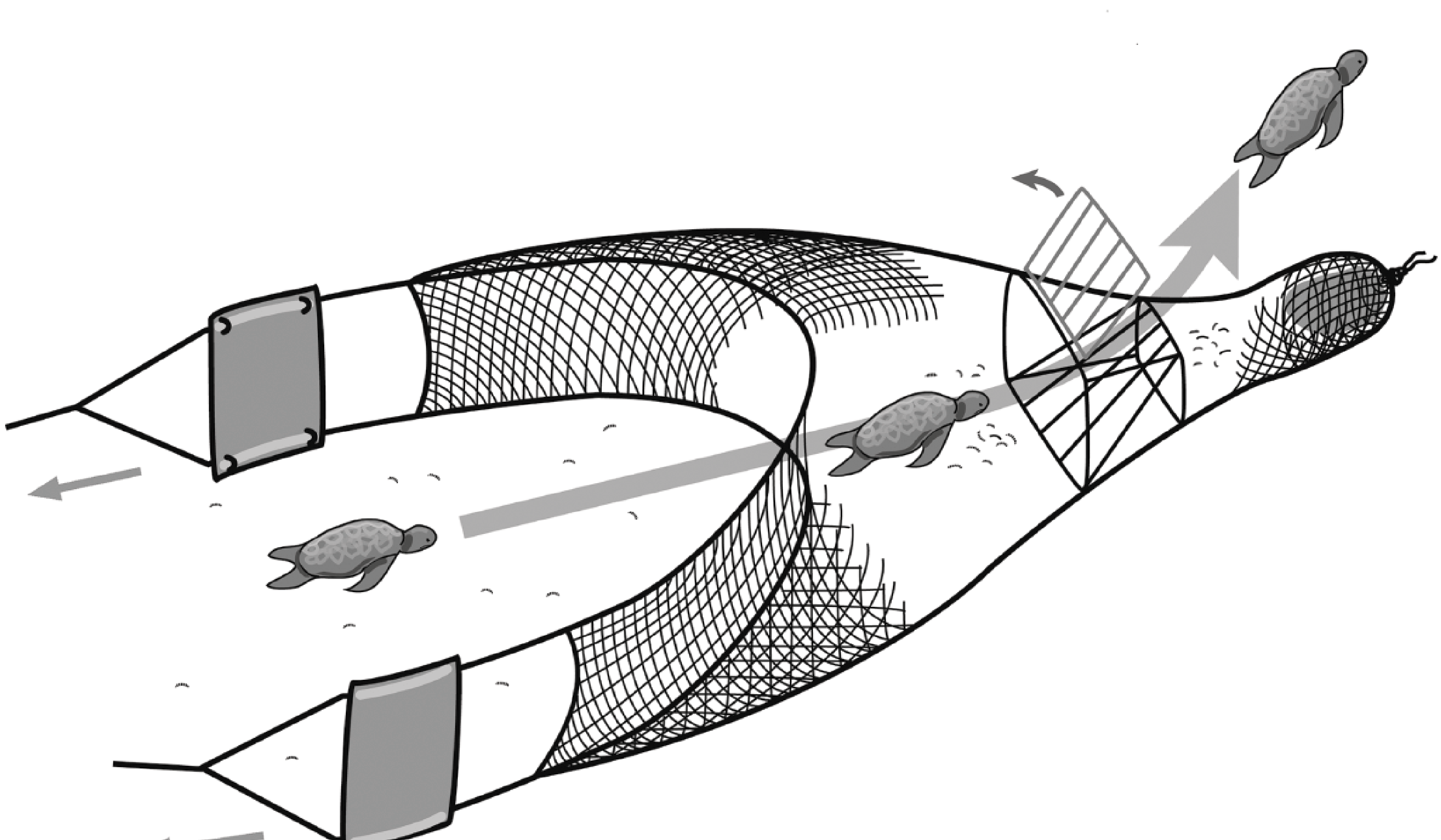


Image from Paul Probert (2017)

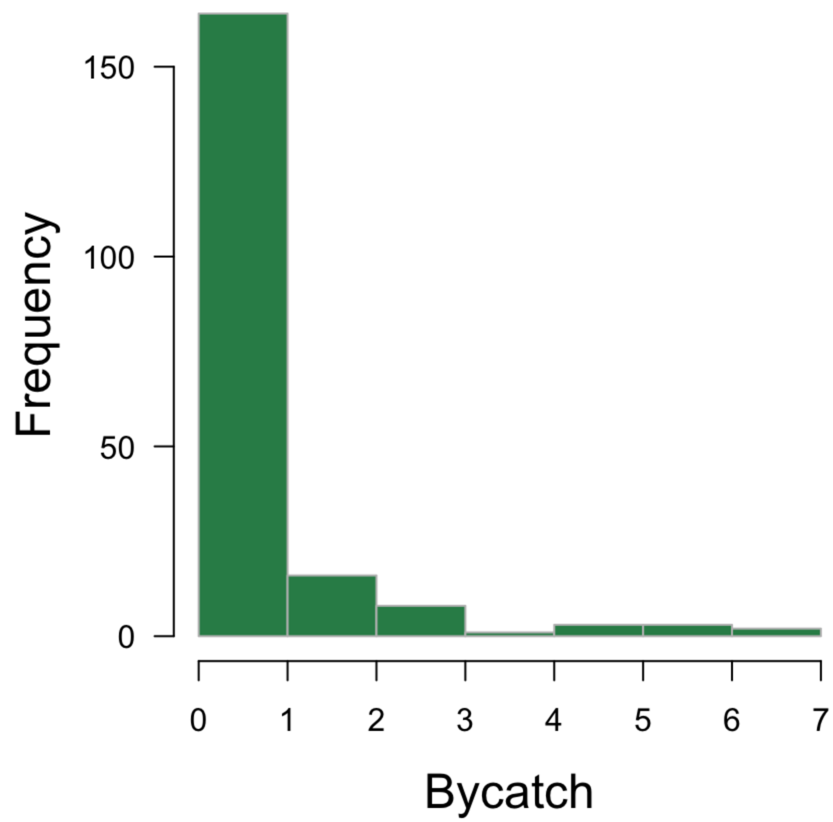
Bycatch of green sea turtles

Let's examine some data on the effectiveness of TEDs in a shrimp fishery

~50% of the fleet was outfitted with TEDS

The number of turtles caught per 1000 trawl hours was recorded along with water temperature

Bycatch of green sea turtles



Our model for bycatch

Bycatch of turtles y_i as a function of TED presence/absence T_i and water temperature W_i

$$
\begin{aligned}
\text{data distribution:} &\quad y_i \sim \text{Poisson}(\lambda_i) \\
\text{link function:} &\quad \log(\lambda_i) = \eta_i \\
\text{linear predictor:} &\quad \eta_i = \alpha + \beta_1 T_i + \beta_2 W_i
\end{aligned}
$$

Bycatch of green sea turtles

```
## Poisson regression
ted_mod <- glm(bycatch ~ TED + temp, data = turtles,
               family = poisson(link = "log"))
## model summary
faraway::sumary(ted_mod)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -1.496656   0.632031 -2.3680   0.01788
## TED         -0.757065   0.176934 -4.2788 1.879e-05
## temp         0.073133   0.030238  2.4185   0.01558
##
## n = 197 p = 3
## Deviance = 345.29336 Null Deviance = 369.77029 (Difference = 24.47693)
```

Pearson's χ^2 statistic

Recall that we can use Pearson's χ^2 statistic as a goodness-of-fit measure

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(n-k)}$$

where O_i is the observed count and E_i is the expected count

Pearson's χ^2 statistic

For $y_i \sim \text{Poisson}(\lambda_i)$

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$\Downarrow$$

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i}$$

Goodness of fit

H_0: Our model is correctly specified

```
## Pearson's X^2 statistic
X2 <- sum((bycatch - fitted(ted_mod))^2 / fitted(ted_mod))
## likelihood ratio test
pchisq(X2, df = nn - length(coef(ted_mod)),
       lower.tail = FALSE)
```

```
## [1] 6.535142e-24
```

The p-value is small so we reject H_0

Variance of Poisson

Recall that the variance for a Poisson is

$$\mathrm{Var}(y) = \mathrm{Mean}(y) = \lambda$$

General variance for count data

We can consider the possibility that the variance scales linearly with the mean

$$\mathrm{Var}(y) = c\lambda$$

If c = 1 then $y \sim \mathrm{Poisson}(\lambda)$

If c > 1 the data are *overdispersed*

Overdispersion

We can estimate c as

$$\hat{c} = \frac{X^2}{n - k}$$

```
## overdispersion parameter
(c_hat <- X2 / (nn - length(coef(ted_mod))))
```

```
## [1] 2.381611
```

Effects on parameter estimates

Recall that $\hat{\boldsymbol{\beta}}$ is *not* affected by overdispersion

but the variance of $\hat{\boldsymbol{\beta}}$ *is* affected, such that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{c}\left(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X}\right)^{-1}$$

$$\hat{\mathbf{W}} = \begin{bmatrix} \hat{\lambda}_1 & 0 & \dots & 0 \\ 0 & \hat{\lambda}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\lambda}_n \end{bmatrix}$$

Bycatch of green sea turtles

```
## model summary
faraway::sumary(ted_mod, dispersion = c_hat)
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.496656   0.975381 -1.5344 0.124923
## TED         -0.757065   0.273053 -2.7726 0.005561
## temp         0.073133   0.046665  1.5672 0.117074
##
## Dispersion parameter = 2.38161
## n = 197 p = 3
## Deviance = 345.29336 Null Deviance = 369.77029 (Difference = 24.47693)
```

Bycatch of green sea turtles

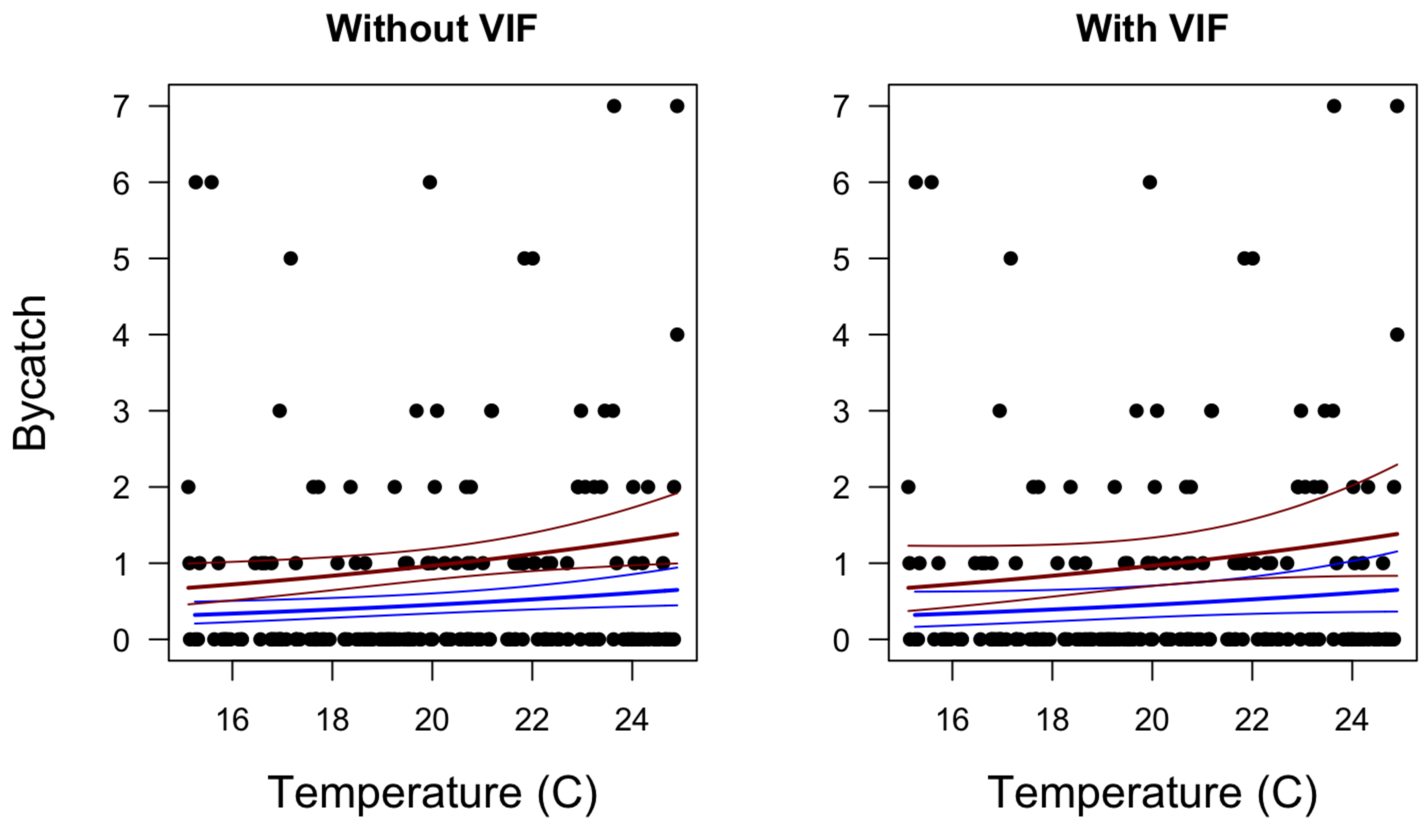
```r
## regular Poisson
signif(summary(ted_mod)$coefficients, 3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5000     0.6320   -2.37 1.79e-02
## TED          -0.7570     0.1770   -4.28 1.88e-05
## temp          0.0731     0.0302    2.42 1.56e-02
```

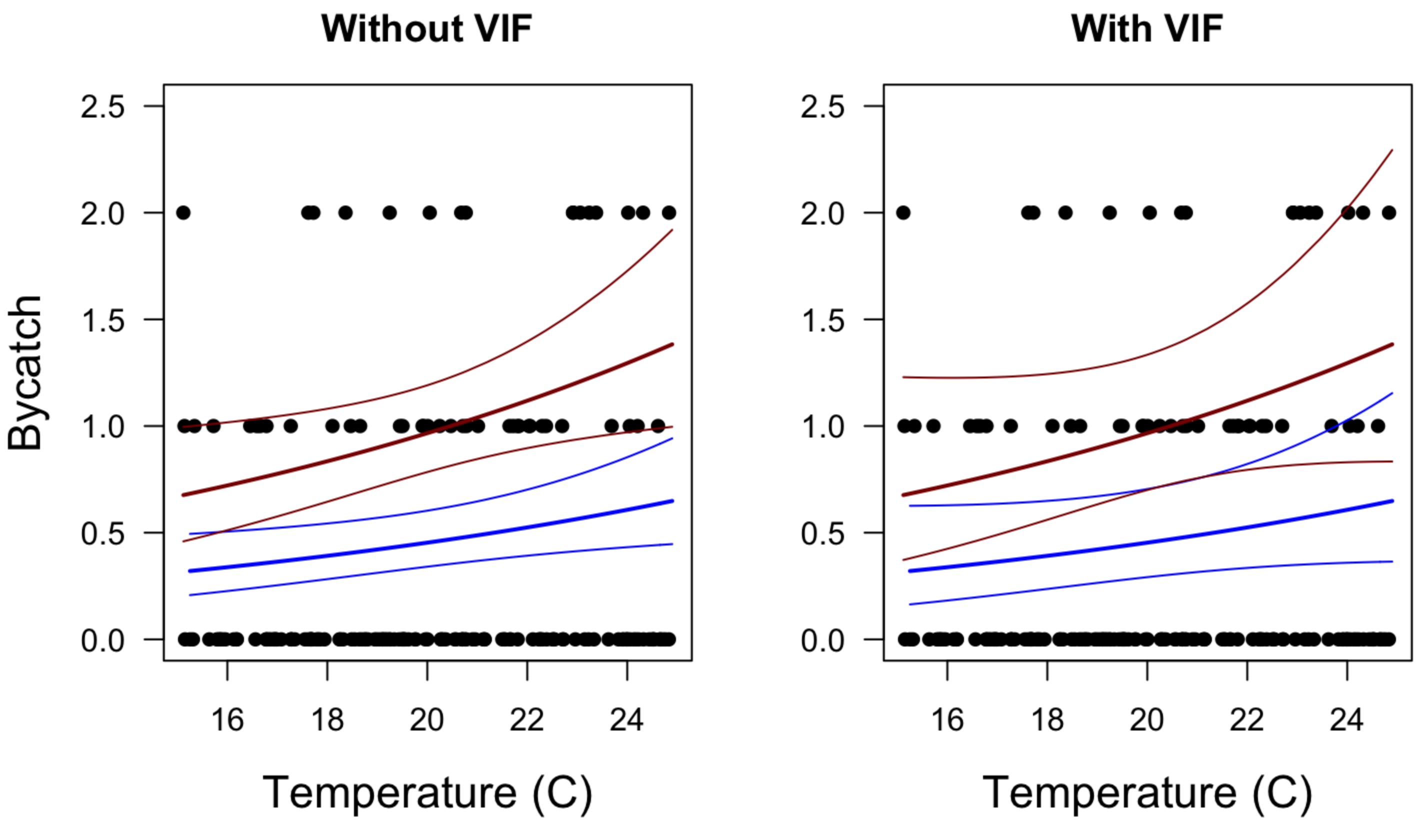
```r
## overdispersed Poisson
signif(summary(ted_mod, dispersion = c_hat)$coefficients, 3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5000     0.9750   -1.53  0.12500
## TED          -0.7570     0.2730   -2.77  0.00556
## temp          0.0731     0.0467    1.57  0.11700
```

Effect of overdispersion



Effect of overdispersion



Quasi-Poisson models

We saw with the case of overdisperse binomial models that we could use a *quasi-likelihood* to estimate the parameters

Quasi-likelihood

Recall that for many distributions we use a *score* (U) as part of the log-likelihood, which can be thought of as

$$U = \frac{(\text{observation} - \text{expectation})}{\text{scale} \cdot \text{Var}}$$

Quasi-likelihood

For example, a normal distribution has a score of

$$U = \frac{y - \mu}{\sigma^2}$$

and a quasi-likelihood of

$$Q = -\frac{(y - \mu)^2}{2}$$

Quasi-likelihood

A Poisson has a score of

$$U = \frac{y - \mu}{\mu\sigma^2}$$

and a quasi-likelihood of

$$Q = y \log \mu - \mu$$

Quasi-Poisson for bycatch

```
## Poisson regression
ted_mod_q <- glm(bycatch ~ TED + temp, data = turtles,
                 family = quasipoisson(link = "log"))
## model summary
faraway::sumary(ted_mod_q)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.496656   0.975381 -1.5344 0.126553
## TED         -0.757065   0.273054 -2.7726 0.006103
## temp         0.073133   0.046665  1.5672 0.118704
##
## Dispersion parameter = 2.38161
## n = 197 p = 3
## Deviance = 345.29336 Null Deviance = 369.77029 (Difference = 24.47693)
```

Quasi-Poisson for bycatch

```r
## quasi-Poisson
signif(summary(ted_mod_q)$coefficients, 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5000     0.9750   -1.53   0.1270
## TED          -0.7570     0.2730   -2.77   0.0061
## temp          0.0731     0.0467    1.57   0.1190
```

```r
## overdispersed Poisson
signif(summary(ted_mod, dispersion = c_hat)$coefficients, 3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5000     0.9750   -1.53  0.12500
## TED          -0.7570     0.2730   -2.77  0.00556
## temp          0.0731     0.0467    1.57  0.11700
```

Quasi-AIC

Just as we di for binomial models, we can use a *quasi*-AIC to compare models

$$QAIC = 2k - 2\frac{\log \mathcal{L}}{\hat{c}}$$

Comparison of bycatch model

Here's a comparison of some models for bycatch

```
##                    k neg-LL   AIC deltaAIC  QAIC deltaQAIC
## B0 + TED + temp   3  255.7 517.3      0.0 220.7       0.0
## B0 + TED          2  258.6 521.2      3.9 221.2       0.5
## B0 + temp         2  265.3 534.7     17.4 226.8       6.1
## B0 only           1  267.9 537.8     20.5 227.0       6.3
```

QUESTIONS?

Negative binomial distribution

The negative binomial distribution describes the *number of failure*s in a sequence of independent Bernoulli trials *before* obtaining a predetermined number of *successes*

Negative binomial distribution

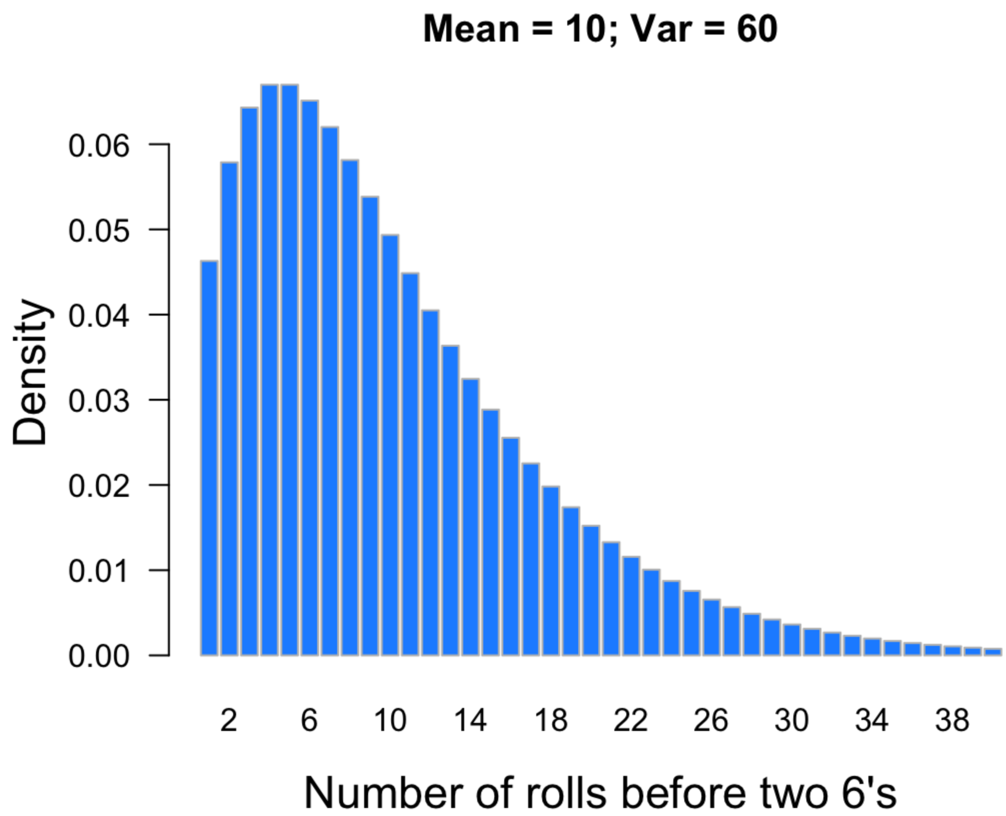
Example

How many times do we have to roll a single die before getting two 6’s?

$$y_i \sim \text{negBin}(k, p)$$

with successes k = 2 and probability p = 1/6

Negative binomial distribution



Negative binomial distribution

The probability mass function is given by

$$f(y; k, p) = \frac{(y + k - 1)!}{(k - 1)! y!} p^k (1 - p)^y$$

$$\text{mean}(y) = \frac{k(1 - p)}{p}$$

$$\text{Var}(y) = \frac{k(1 - p)}{p^2}$$

Negative binomial distribution

The negative binomial distribution can also arise as a mixture of Poisson distributions, each with a mean that follows a gamma distribution

$$y \sim \mathrm{Poisson}(\lambda)$$

$$\lambda \sim \mathrm{Gamma}\left(r, \frac{p}{1-p}\right)$$

Negative binomial distribution

In terms of the mean and variance

$$f(y;k,\mu) = \frac{(y+k-1)!}{(k-1)!y!}\left(\frac{\mu}{\mu+r}\right)^k\left(\frac{r}{\mu+r}\right)^y$$

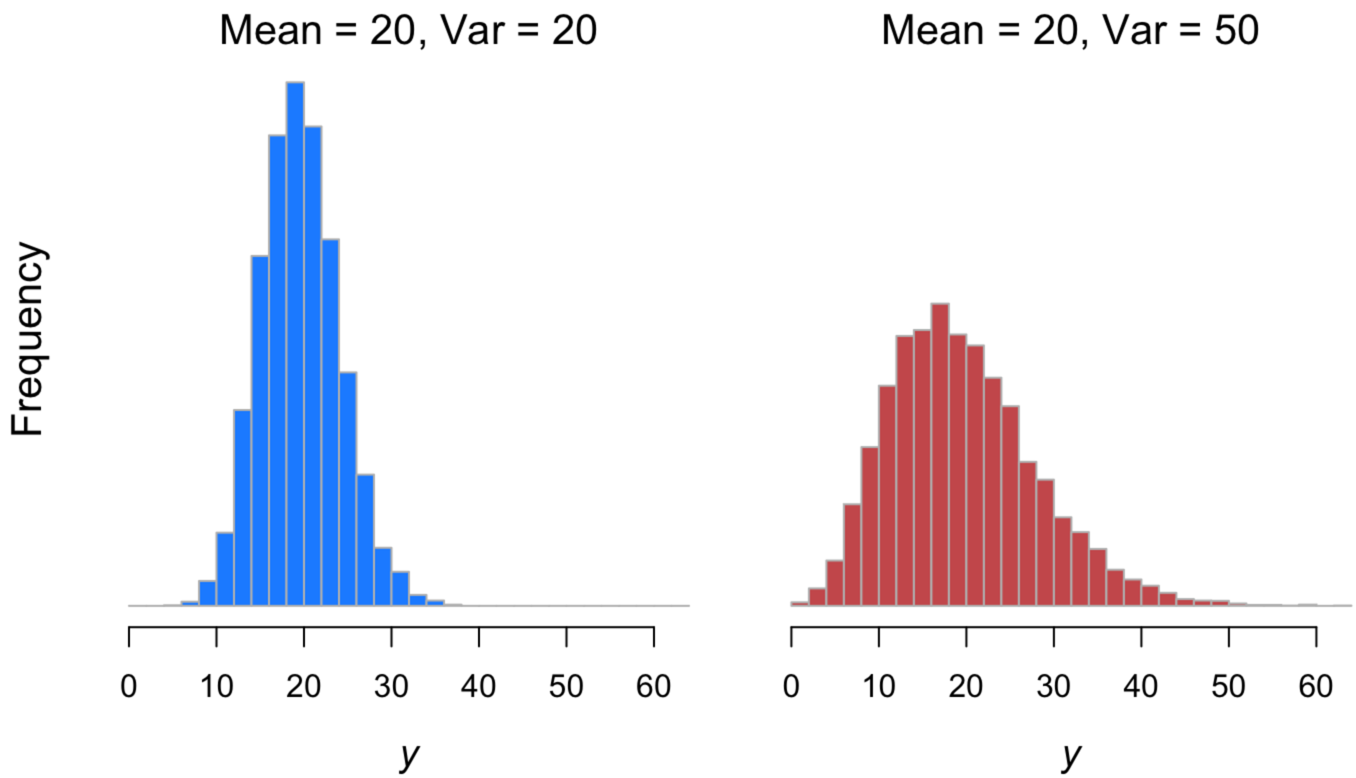
$$\text{mean}(y) = \mu$$

$$\text{Var}(y) = \mu + \frac{\mu^2}{r}$$

Negative binomial distribution

The extra parameter r allows more variance than the Poisson, which allows us greater flexibility in fitting the data

Negative binomial distribution



Poisson as limiting case

Note that

$$
\begin{gathered}
\operatorname{Var}(y) = \mu + \frac{\mu^2}{r} \\
\Downarrow \\
\lim_{r \to \infty} \operatorname{Var}(y) = \mu
\end{gathered}
$$

As r gets large, the negative binomial converges to the Poisson

Our model for bycatch

Bycatch of turtles y_i as a function of TED presence/absence T_i and water temperature W_i

$$\begin{aligned} \text{data distribution:} \quad & y_i \sim \text{negBin}(k, \mu_i) \\ \text{link function:} \quad & \log(\mu_i) = \eta_i \\ \text{linear predictor:} \quad & \eta_i = \alpha + \beta_1 T_i + \beta_2 W_i \end{aligned}$$

Bycatch of green sea turtles

Let's model our bycatch data with a negative binomial using `glm.nb()` from the **MASS** package

```
## load MASS
library(MASS)
## neg binomial regression
ted_mod_nb <- glm.nb(bycatch ~ TED + temp, data = turtles,
                     link = "log")
```

Bycatch of green sea turtles

```
## model summary
signif(summary(ted_mod_nb)$coefficients, 3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5200     0.9990   -1.52  0.12900
## TED          -0.7850     0.2720   -2.88  0.00393
## temp          0.0748     0.0486    1.54  0.12400
```

Bycatch of green sea turtles

```
## overdispersed Poisson
signif(summary(ted_mod, dispersion = c_hat)$coefficients, 3)
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5000     0.9750   -1.53  0.12500
## TED          -0.7570     0.2730   -2.77  0.00556
## temp          0.0731     0.0467    1.57  0.11700
```

```
## negative binomial
signif(summary(ted_mod_nb)$coefficients, 3)
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5200     0.9990   -1.52  0.12900
## TED          -0.7850     0.2720   -2.88  0.00393
## temp          0.0748     0.0486    1.54  0.12400
```

Summary

There are several ways to model overdispersed count data, each with its own pros and cons

Model	Pros	Cons
Poisson	Easy	Underestimates variance
Poisson with VIF	Easy; estimate of variance	Ad hoc
quasi-Poisson	Easy; estimate of variance	No distribution for inference
negative-binomial	Easy; estimate of variance	None