Negative-Binomial Regression

Grinnell College

November 13, 2026

Same Mean, More Variation

A Poisson regression assumes

$$Y_i \mid \mathbf{x}_i \sim \text{Poisson}(\mu_i), \qquad \log \mu_i = \beta_0 + \beta_1 x_i,$$

with $\text{Var}(Y_i \mid \mathbf{x}_i) = \mu_i$.

What if we observe more variation than the model predicts?

Can we keep the mean model but allow counts to vary more?

Overdispersion is the phenomenon in which the variability in our GLM is greater than expected, given the random component

Negative Binomial

The **negative binomial** distribution is a discrete distribution, typically used to model the number of failures (y) in a sequence of independent Bernoulli trials before observing a fixed number of successes (r)

$$f(Y \mid r, p) = \binom{y + r - 1}{y} p^r (1 - p)^y, \qquad x = 0, 1, 2, \ldots$$

Like the Poisson, the response is a count defined on the non-negative integers

One Origin: Waiting for Successes

Suppose independent trials each succeed with probability p. Let Y count **failures before the rth success**.

For $r = 2$, the sequence

$$F,\ S,\ F,\ F,\ S$$

has $Y = 3$ failures when the second success occurs.

Y has a negative-binomial distribution:

$$E(Y) = \frac{r(1-p)}{p}, \qquad \text{Var}(Y) = \frac{r(1-p)}{p^2} = \left(\frac{r(1-p)}{p}\right)\frac{1}{p}$$

Connection to the Geometric Distribution

When $r = 1$, Y counts failures before the first success:

$$P(Y = y) = (1 - p)^y p, \qquad y = 0, 1, 2 \ldots$$

This is the geometric distribution

For integer r, add r independent geometric counts to obtain a negative-binomial count.

For regression, we need a more flexible story: θ need not be an integer, and the observations need not arise from waiting for successes.

Unobserved Differences Between Subjects

Imagine two female crabs of the same color.

- They need not have identical underlying attractiveness to satellites.
- A Poisson mean based only on color cannot represent every difference.

A model can let each crab have its own unobserved Poisson mean Λ_i:

$$Y_i \mid \Lambda_i \sim \text{Poisson}(\Lambda_i)$$

where Λ_i itself is a random variable. Variation in these means adds variation to the observed counts.

A Gamma Distribution for the Poisson Mean

The gamma distribution describes positive quantities. Choose

$$\Lambda_i \sim \text{Gamma}\left(\text{shape} = \theta, \text{rate} = \frac{\theta}{\mu_i}\right).$$

This choice gives

$$E(\Lambda_i) = \mu_i, \qquad \text{Var}(\Lambda_i) = \frac{\mu_i^2}{\theta}.$$

μ_i is the mean predicted by the observed covariates. Λ_i varies around it because subjects differ in unobserved ways.

Poisson–Gamma: The Mean

First average over Poisson counts at a given Λ_i; then average over possible values of Λ_i.

$$\begin{aligned} E(Y_i) &= E\{E(Y_i \mid \Lambda_i)\} \\ &= E(\Lambda_i) \\ &= \mu_i. \end{aligned}$$

Allowing the underlying Poisson mean to vary does not change our specified overall mean.

Poisson–Gamma: The Variance

The law of total variance separates two sources of variation:

$$\begin{aligned}
\text{Var}(Y_i) = & \underbrace{E\{\text{Var}(Y_i \mid \Lambda_i)\}}_{\text{variation within a given Poisson mean}} \\
& + \underbrace{\text{Var}\{E(Y_i \mid \Lambda_i)\}}_{\text{variation between Poisson means}} \\
& = E(\Lambda_i) + \text{Var}(\Lambda_i) \\
& = \mu_i + \frac{\mu_i^2}{\theta}.
\end{aligned}$$

The full mixture distribution is negative binomial.

The Negative-Binomial Mean and Variance

We will use parameters $\mu > 0$ and $\theta > 0$:

$$E(Y) = \mu, \qquad \boxed{\text{Var}(Y) = \mu + \frac{\mu^2}{\theta}.}$$

- The response is still a count: $0, 1, 2, \ldots$.
- Smaller θ means more extra variation at a fixed mean.
- As $\theta \to \infty$, the distribution approaches Poisson.

Agresti uses $D = 1/\theta$, so the variance is $\mu + D\mu^2$.

Two Parameterizations, One Distribution

For the waiting-time version, set $r = \theta$ and

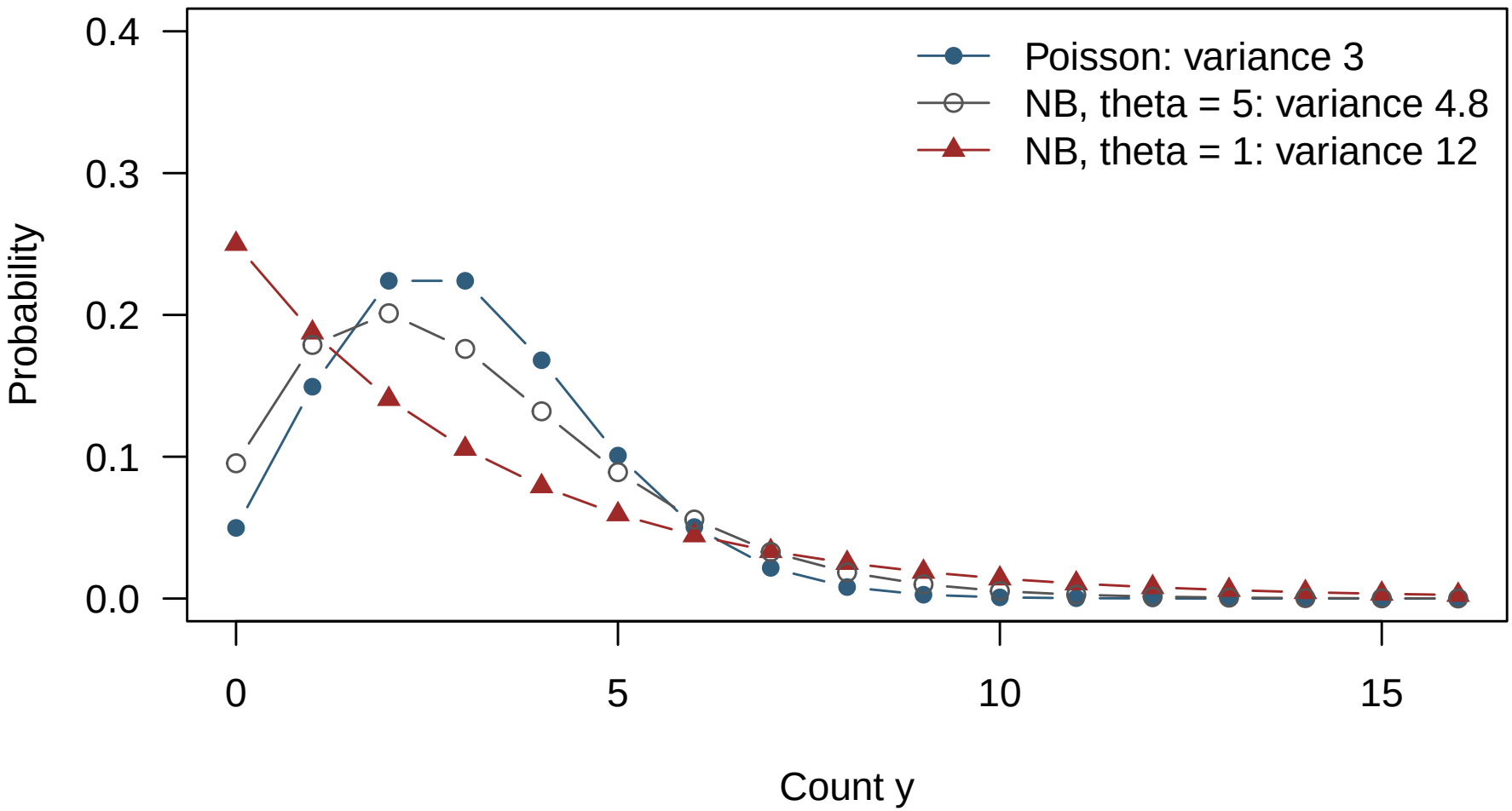
$$p = \frac{\theta}{\theta + \mu}, \qquad 1 - p = \frac{\mu}{\theta + \mu}.$$

Then

$$\frac{\theta(1-p)}{p} = \mu, \qquad \frac{\theta(1-p)}{p^2} = \mu + \frac{\mu^2}{\theta}.$$

The gamma-mixture formulation allows any positive θ.

Hold the Mean Fixed



All three distributions have mean 3. The more dispersed distribution places more probability at zero and in the far right tail.

Why This Model Appears in Many Areas

- **Ecology:** counts vary because individuals and habitats differ.
- **Sequencing:** read counts vary across biological samples; negative-binomial models underlie methods such as DESeq2.
- **Rates:** crashes or other events may vary more than a Poisson model allows, even after accounting for exposure.

The heterogeneity within groups is one useful explanation for overdispersion.

Negative-Binomial Regression

For independent counts conditional on predictors,

$$Y_i \sim \text{NB}(\mu_i, \theta), \qquad \log \mu_i = \beta_0 + \beta_1 x_i.$$

- Slopes still describe multiplicative changes in the mean.
- One common θ is estimated for the observations.
- The random component is now a negative binomial, where the variance can be greater than the mean

Offsets with overdispersion also work the same way:

$$\log \mu_i = \log t_i + \beta_0 + \beta_1 x_i.$$

Horseshoe Crabs

We have 173 female crabs for whom we are trying to model the number of male satellites

```
> summarize(crabs, mean(sat), var(sat))
  mean(sat) var(sat)
1    2.9191    9.912
```

Clearly, we see the variance in satellites is greater than the mean, but suitability for Poisson regressions assumes equality *after* accounting for covariates

Horseshoe Crabs

If we were interested in using color as our predictor, better to check the mean/variance relation within each group

```
group_by(crabs, factor(color)) %>%
  summarize(mean = mean(sat), var = var(sat))

  factor(color)  mean   var
1 1                4.08  9.72
2 2                3.29 10.3
3 3                2.23  6.74
4 4                2.05 13.1
```

Even after accounting for color, there is still more variability than would be suggested by a Poisson random variable

Poisson Regression

```
glm(formula = sat ~ factor(color), family = poisson, data = crab)

Coefficients:
                Estimate Std. Error z value             Pr(>|z|)
(Intercept)        1.407      0.143    9.85 < 0.0000000000000002 ***
factor(color)2    -0.215      0.154   -1.40              0.16249
factor(color)3    -0.606      0.175   -3.46              0.00053 ***
factor(color)4    -0.691      0.206   -3.35              0.00081 ***


    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 609.14  on 169  degrees of freedom
AIC: 972.4
```

Residual deviance suggests that we perhaps don't have an adequate model fit, but it says nothing about the suitability of the Poisson assumption

Negative Binomial Regression

The `MASS` package contains the function `glm.nb`

```
glm.nb(formula = sat ~ factor(color), data = crab, init.theta = 0.8018786143,
    link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.407      0.353    3.99 0.000066 ***
factor(color)2  -0.215      0.375   -0.57     0.57
factor(color)3  -0.606      0.404   -1.50     0.13
factor(color)4  -0.691      0.451   -1.53     0.13


    Null deviance: 199.23  on 172  degrees of freedom
Residual deviance: 194.00  on 169  degrees of freedom
AIC: 772.3

Number of Fisher Scoring iterations: 1


              Theta:  0.802
          Std. Err.:  0.136

 2 x log-likelihood:  -762.296
```

Reported here is the dispersion term, $\theta = \frac{1}{D}$, where smaller values for θ indicated *greater* disperson

Comparing Coefficients

In this case of a single categorical covariate, the fitted estimates are identical. This is not typically true, but the they are generally fairly close

```
> coef(summary(fit_pois)) %>% round(5)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.40691    0.14286  9.8484  0.00000
factor(color)2 -0.21459    0.15363 -1.3967  0.16249
factor(color)3 -0.60614    0.17496 -3.4644  0.00053
factor(color)4 -0.69129    0.20647 -3.3481  0.00081

> coef(summary(fit_nb)) %>% round(5)
               Estimate Std. Error  z value Pr(>|z|)
(Intercept)     1.40691    0.35261  3.99005  0.00007
factor(color)2 -0.21459    0.37504 -0.57218  0.56720
factor(color)3 -0.60614    0.40358 -1.50189  0.13312
factor(color)4 -0.69129    0.45082 -1.53342  0.12517
```

What does change is the size of the standard error, where the negative binomial permits extra variation

Comparing Coefficients

Here, we consider differences in models that only consider width

```
> coef(summary(fit_pois_w)) %>% round(5)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224 -6.0946        0
width        0.16405    0.01997  8.2165        0

> coef(summary(fit_nb_w)) %>% round(5)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.05251    1.17143 -3.4594  0.00054
width        0.19207    0.04406  4.3595  0.00001
```

Now in addition to an inflated standard error, the coefficient values have adjusted somewhat while remaining similar

Width Only Model – Dispersion

Interpretation of coefficients (and the construction of CI) works identically to the Poisson model. Here, we see an increase of width is associated with a multiplicative change in the number of satellites by $\exp(0.192) = 1.21$

```
> summary(fit_nb_w)

Call:
glm.nb(formula = sat ~ width, data = crabs, init.theta = 0.90456808,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0525     1.1714   -3.46  0.00054 ***
width         0.1921     0.0441    4.36 0.000013 ***
---

    Null deviance: 213.05  on 172  degrees of freedom
Residual deviance: 195.81  on 171  degrees of freedom
AIC: 757.3

Number of Fisher Scoring iterations: 1


              Theta:  0.905
          Std. Err.:  0.161

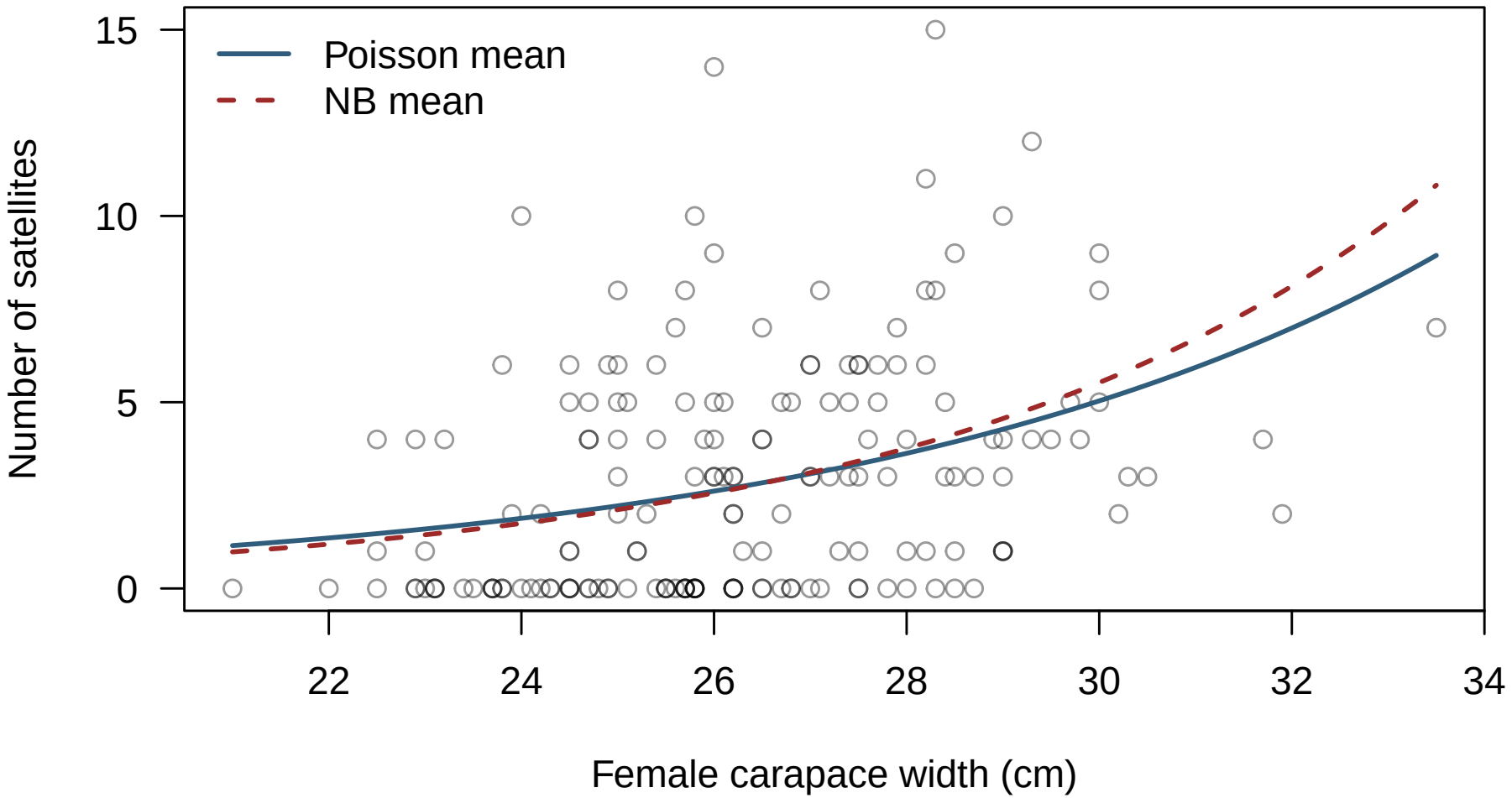
 2 x log-likelihood:  -751.291
```

Compare Fitted Values

```
> new <- data.frame(width = c(22, 26, 30))
> data.frame(new, Poisson = predict(pois, new, type="response"),
    NB = predict(fit, new, type="response"))
  width Poisson    NB
1    22   1.356 1.189
2    26   2.613 2.564
3    30   5.036 5.527
```

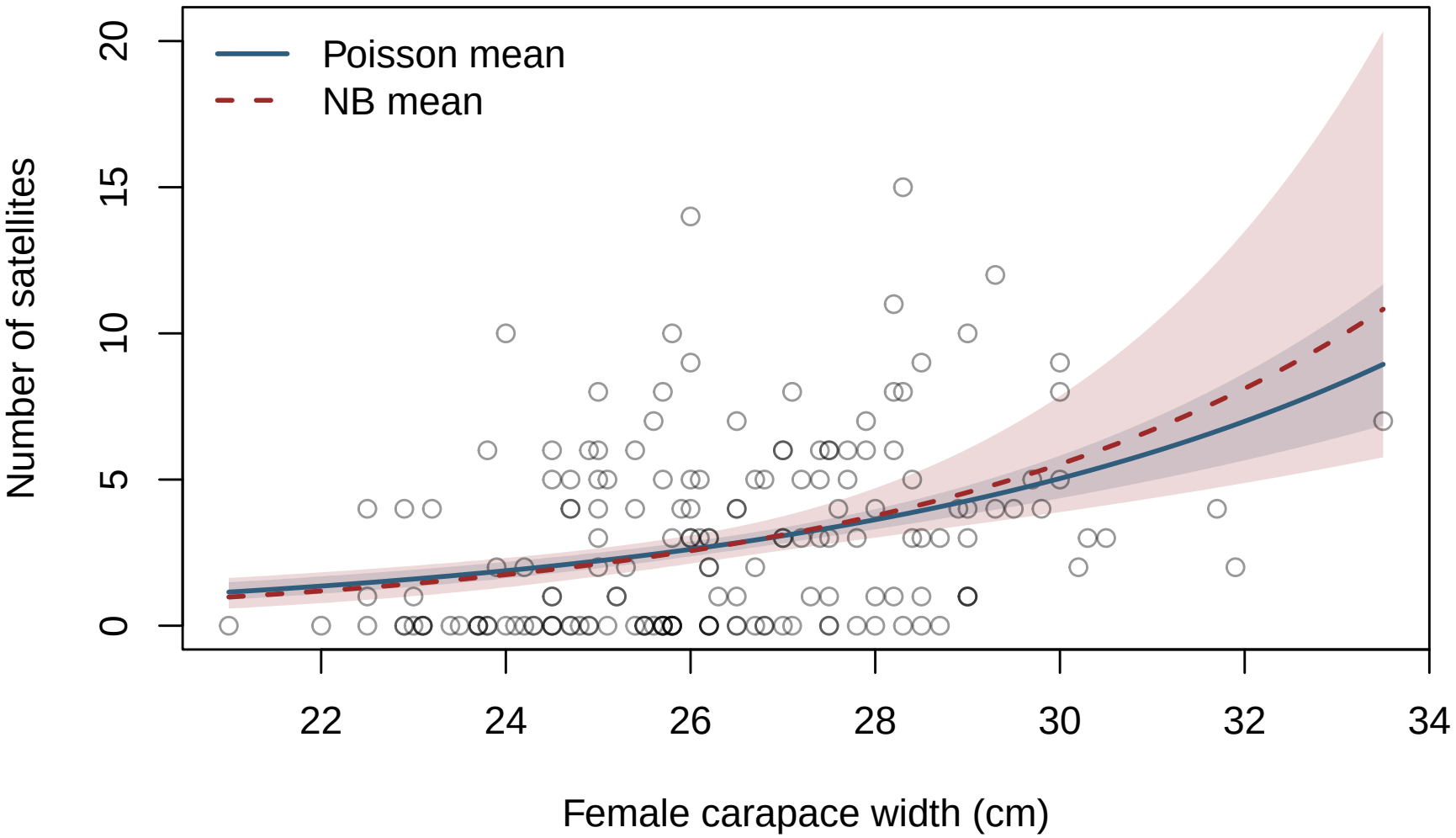
Both fits tell a similar story about mean satellite counts across widths.

Similar Mean Curves



The mean curves are fairly close. The distribution of counts around those curves is the main distinction.

Different SE



How Much Extra Variation?

```
> c(theta = fit$theta, SE = fit$SE.theta,
    D = 1 / fit$theta)
  theta      SE       D
0.90457 0.16121 1.10550
```

- $\hat{\theta} \approx 0.905$; $\hat{D} = 1/\hat{\theta} \approx 1.106$.
- `SE.theta` is an approximate standard error based on observed information.

Put in context, this suggests for a crab Y_i with expected count μ_i given x_i, we should see

$$\text{Var}(Y_i \mid x_i) = \mu_i + (1.1)\mu_i^2$$

wald test on theta, note that we not testing theta $= 0$

Pearson Dispersion Statistic

Alternatively, the **Pearson dispersion statistic**, ϕ, compares the ratio of the sum of squared Pearson residuals (which are approximately χ^2)

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(Y_i)}}.$$

with the associated residual degrees of freedom, $n - p$

$$\phi = \frac{\sum r_i^2}{n - p}$$

Values near 1 indicate little evidence for overdispersion

Pearson Dispersion Statistic

$$\phi = \frac{\sum r_i^2}{n - p}$$

```
> c(Poisson = sum(residuals(pois, type="pearson")^2) /
    df.residual(pois), NB = sum(residuals(fit, type="pearson")^2) /
    df.residual(fit))
Poisson      NB
 3.1822  0.8465
```

The Poisson ratio is much larger than 1. The NB ratio is much closer

Compare AIC

AIC is still valid, as the likelihoods describe the joint probabilities of the data

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\theta}) + 2k.$$

```
> AIC(pois, fit)
     df   AIC
pois  2 927.2
fit   3 757.3
```

- Smaller AIC is preferred among models fit to the same response data.
- NB estimates three parameters: intercept, slope, and θ.
- The improvement in log likelihood outweighs the penalty here.

AIC is a comparison criterion, not a goodness-of-fit test or a p-value.

Why Not Subtract the Residual Deviances?

However, we can't compare residual deviances to determine improvement of fit as the saturated models are different

$$G^2_{\text{Pois}} = 2\big(\ell_{\text{sat,Pois}} - \ell_{\text{Pois}}\big),$$
$$G^2_{\text{NB}} = 2\big(\ell_{\text{sat,NB}} - \ell_{\text{NB}}\big).$$

Subtracting gives

$$\begin{aligned} G^2_{\text{Pois}} - G^2_{\text{NB}} &= 2\big(\ell_{\text{sat,Pois}} - \ell_{\text{Pois}}\big) - 2\big(\ell_{\text{sat,NB}} - \ell_{\text{NB}}\big) \\ &= \underbrace{2\big(\ell_{\text{NB}} - \ell_{\text{Pois}}\big)}_{\text{likelihood-ratio statistic}} + \underbrace{2\big(\ell_{\text{sat,Pois}} - \ell_{\text{sat,NB}}\big)}_{\text{saturated terms do not cancel}}. \end{aligned}$$

The Likelihood Ratio Test

Even though the saturated models are different, we could theoretically perform a likelihood ratio test of $H_0 : D = 0$

$$\underbrace{2\left(\ell_{\text{NB}} - \ell_{\text{Pois}}\right)}$$

The problem, however, is that $D = 0$ falls on the boundary of the parameter space

If the null is correct and the data truly is Poisson, then about half the time the observed variability will be less than μ

As the negative binomial can only handle positive dispersion, it will set $D = 0$, resulting in the LRT having a value of zero far more than expected in a chi-squared random variable

The Likelihood Ratio Test

Long story short: LRT can be done, but it takes some maneuvering

Too many zeros

What if some female crabs are unlikely to have any satellites, for reasons beyond what we have observed in our model covariates

In other words, is it possible that there is a group of crabs that will always have zero satellites, regardless of their covariate values? How would we determine this?

In general, a negative binomial will predict more zeros than a Poisson, but is it enough?

Oops all zero?

How many zeros do our fitted count models predict?

	Crabs with zero satellites
Observed	62
Expected under Poisson	14.0
Expected under NB	50.5

We see that the NB model, while more than Poisson, still falls short. Where these zeros by chance, or is there something we are missing?

Next question: What would a model with two ways to get a zero look like?

Several Ways to Assess a Model

- **Fitted values:** what mean pattern does the model describe?
- **Dispersion and residuals:** does the assumed variation look plausible?
- **AIC:** does the added flexibility improve the fit enough to justify its cost?
- **LRT:** Gotta be careful, but it can be done

For these crabs, NB retains the mean trend while describing much more of the observed variation. No single diagnostic answers every question.