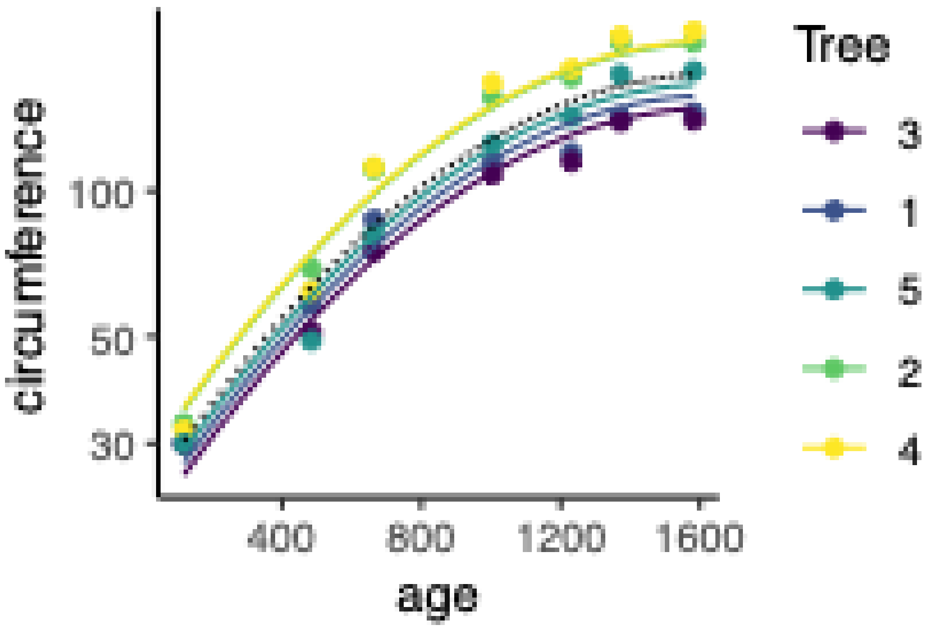
# Generalized Linear Modelling 2 – fitting

Starting point:

- Clarity about what is your response variable and what are explanatory variables
- Clarity about how you expect your response variable to be distributed
- Scientific questions about the relationship between response and explanatory variables
- Data in a long-form dataframe or tibble

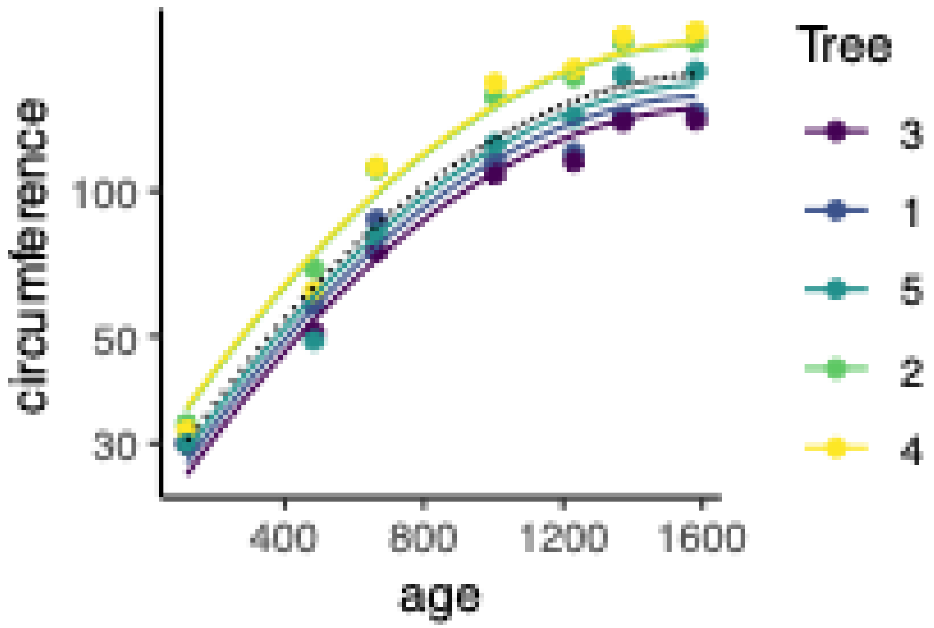
# Types of response variable

- Have considered different kinds of explanatory variable
  - Numeric, categorical, fixed, random
- Have silently assumed things about the response variable
  - Numeric, continuous
  - Normally distributed (have at least checked that one)
- What if I'm counting pigeons in my garden?
  - Counts: Numeric, categorical – integers only
  - Don't know how many pigeons there could be
  - Never negative
  - Is an average number (e.g. 2)



# Types of response variable

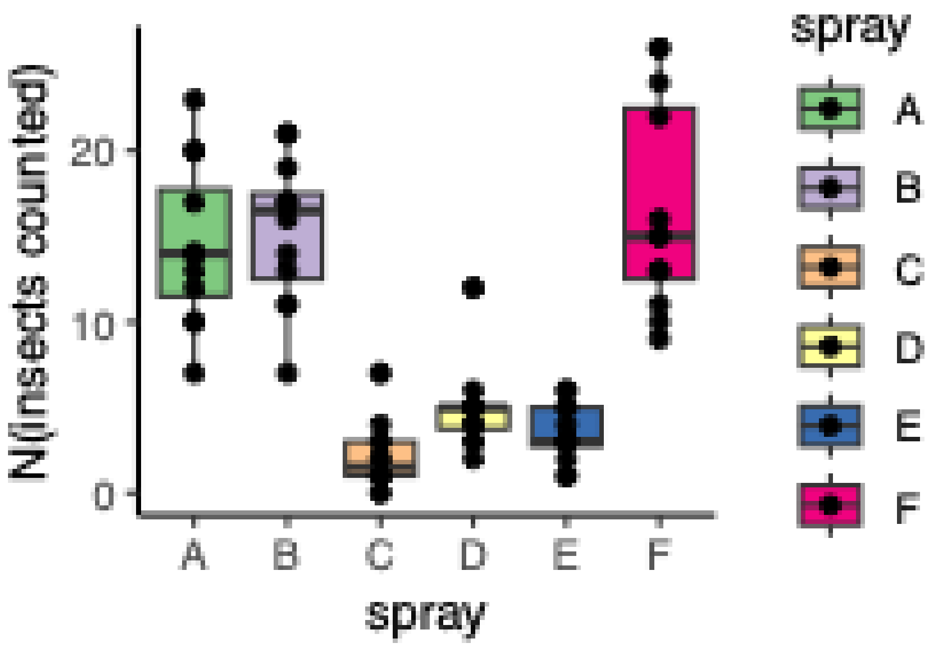
- What if I’m counting pigeons in my garden?
  - Counts: Numeric, categorical – integers only
  - Don’t know how many pigeons there could be
  - Never negative
  - Is an average number (e.g. 2)



# Types of response variable

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models

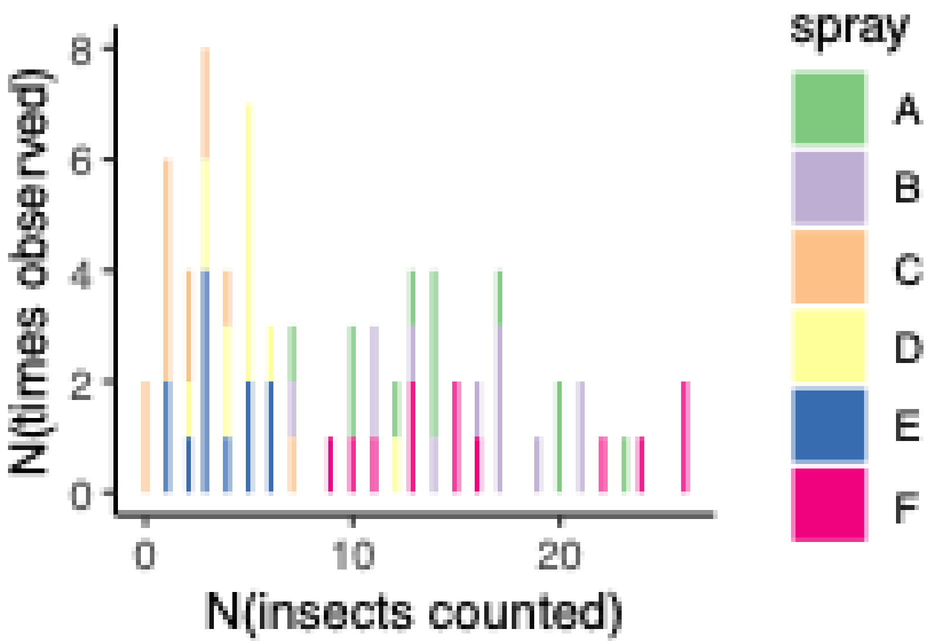



Beall, G. **The Transformation of Data from Entomological Field Experiments so that the Analysis of Variance Becomes Applicable**. *Biometrika* 32 (1942). 10.2307/2332128

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models




- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

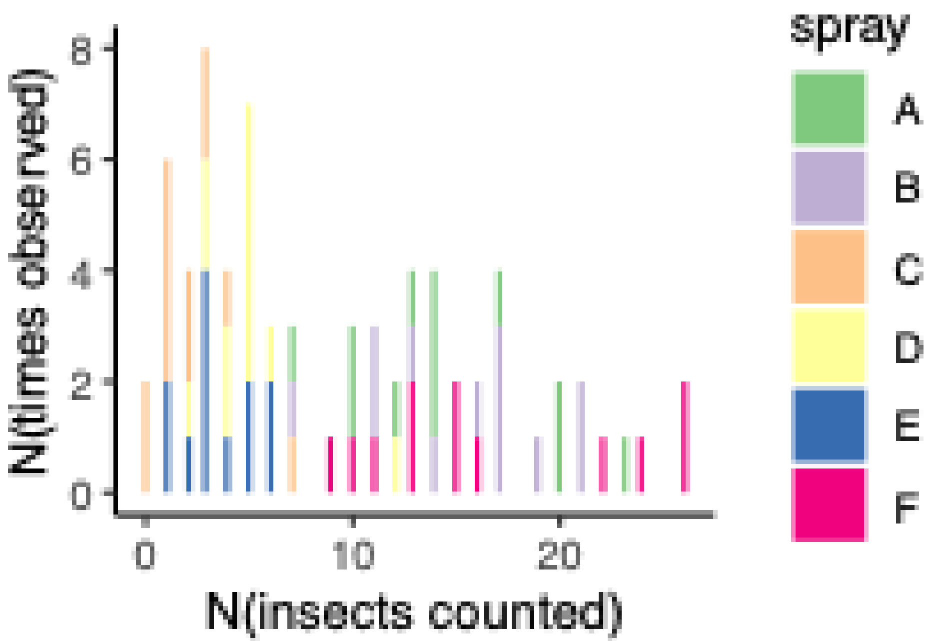
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used?
  - How much does insect count change with spray used?
  - What's the expected count for spray A?
  - How sure are we of the answers?

```
> model <- glm(count ~ spray, family = "poisson", data = InsectSprays)

> summary(model)
```

# Generalised linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models



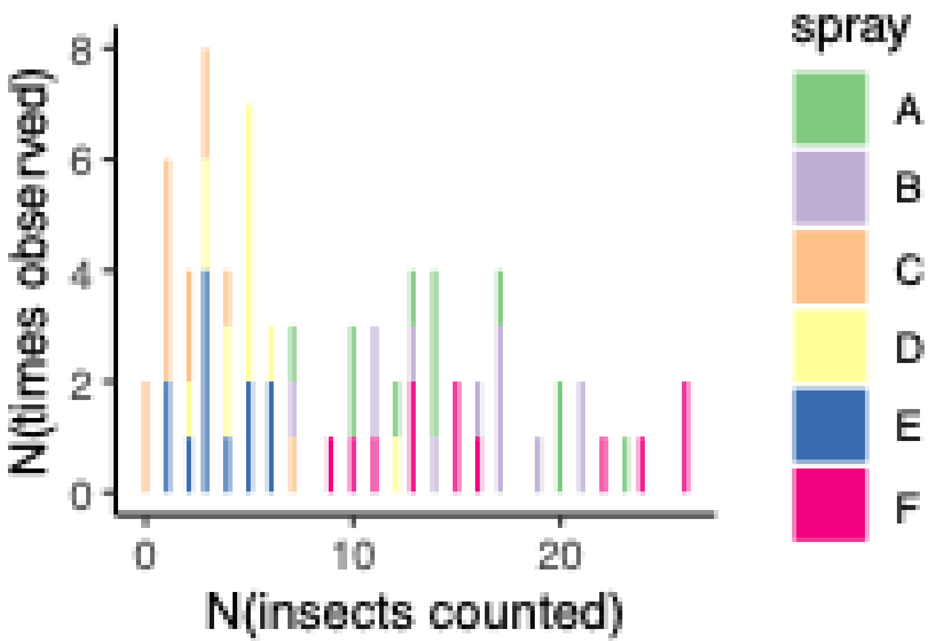

Check residual deviance

```
> model <- glm(count ~ spray, family = poisson, data = InsectSprays)
> summary(model)

Call:
glm(formula = count ~ spray, family = poisson, data = InsectSprays)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: 376.59

Number of Fisher Scoring iterations: 5
```

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models




```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
```

Check residual deviance

Moderate **over-dispersion**

library(DHARMa)

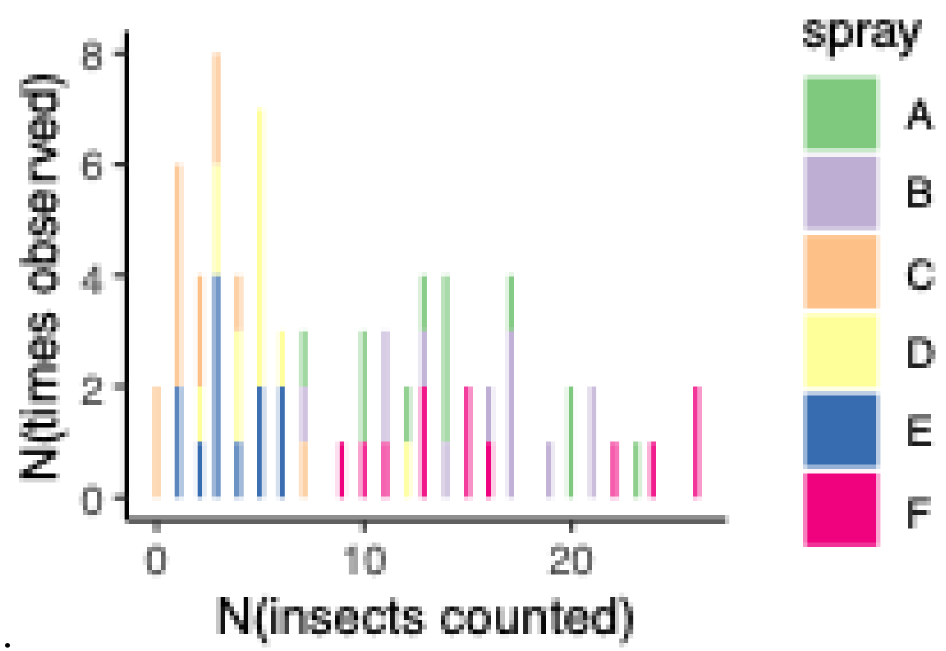
- Use to test over-dispersion
- Use for diagnostics plots

- For these purposes '**deviance**' means pretty much the same as '**variance**'
- Expect the residuals to have a Poisson distribution where mean = variance
- So expect **residual deviance** to equal the residual degrees of freedom
- **Dispersion parameter** is just the ratio residual deviance / residual df
- i.e. dispersion parameter here is 98/66 = 1.5

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models




- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
```

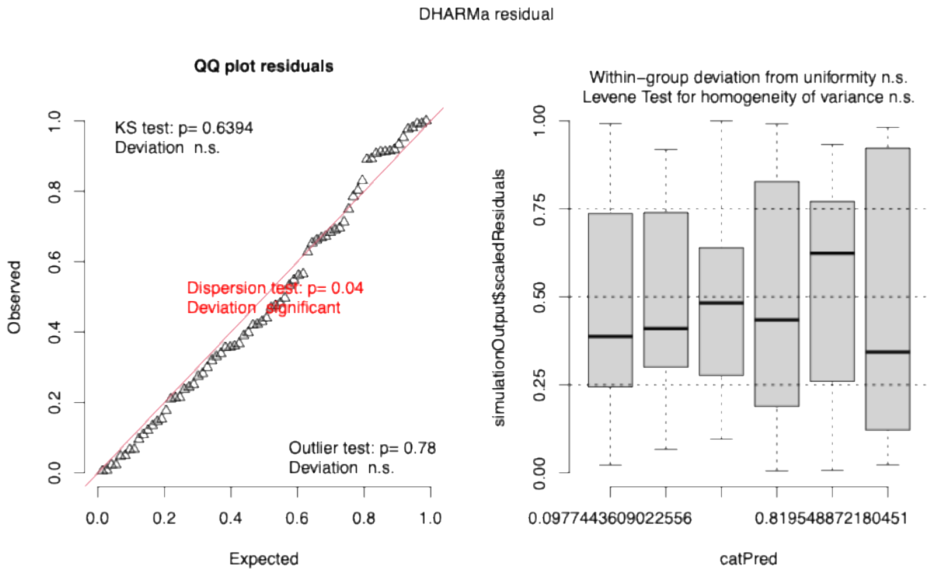
Check residual deviance

Moderate **over-dispersion**

library(DHARMa)

- Use to test over-dispersion
- Use for diagnostics plots

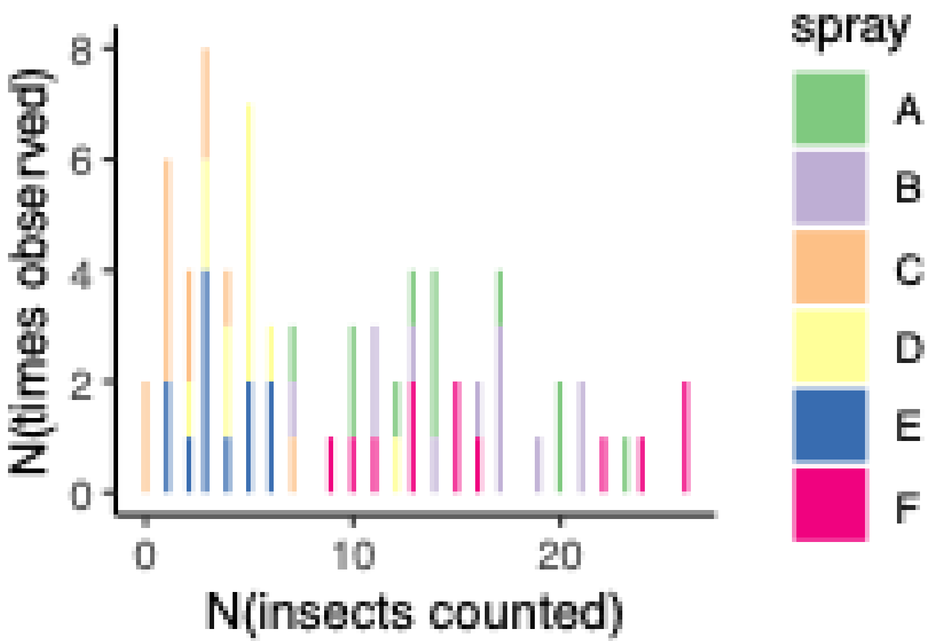
```
plot(simulateResiduals(model))
```



# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models

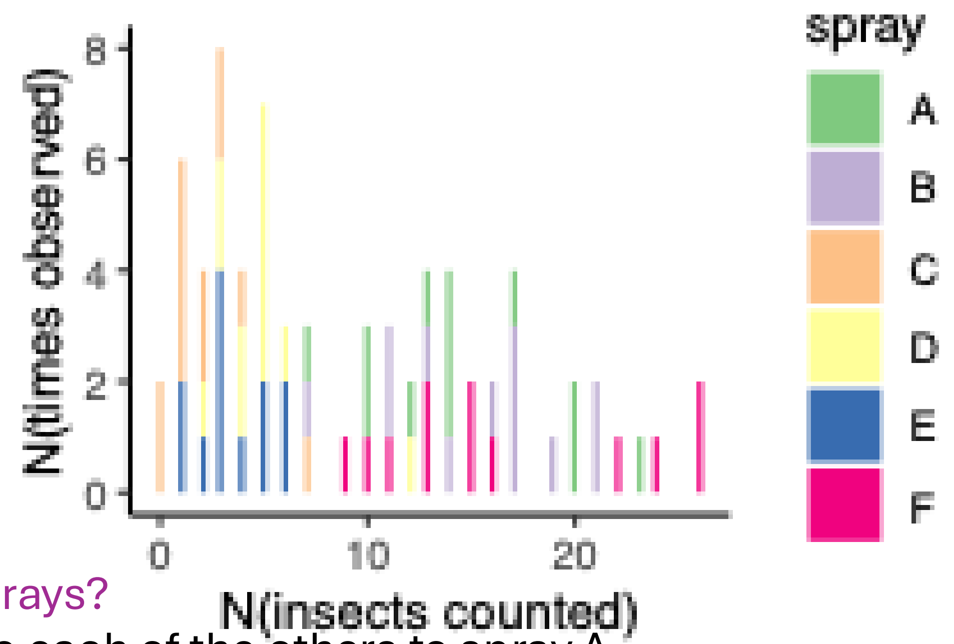



```
> model <- glm(count ~ spray, family = poisson, data = InsectSprays)
> summary(model)

Call:
glm(formula = count ~ spray, family = poisson, data = InsectSprays)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: 376.59

Number of Fisher Scoring iterations: 5
```

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models




```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
```

Is there variation among sprays?

- Wald tests only compare each of the others to spray A
- Want one test of whether spray matters

`anova(model)`

```
> anova(model)
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     71     409.04
spray  5   310.71        66      98.33 < 2.2e-16 ***
---
```

How much variation?

- Linear model so numbers add up

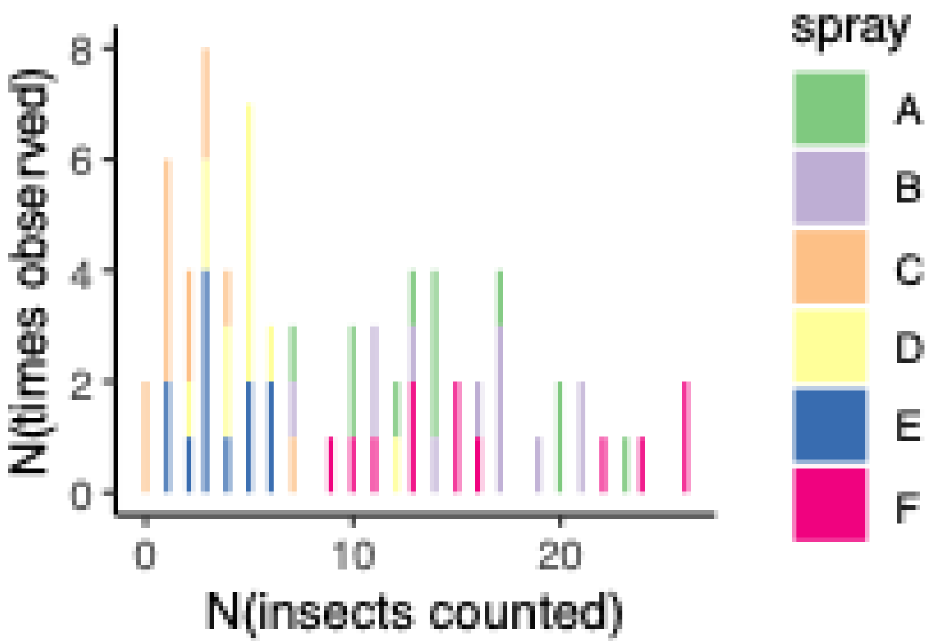
$$\hat{y}_i = \beta_0 + \beta_1 x_1$$

- So expect spray A to be 2.67, spray E to be 2.67 – 1.42 = 1.25 etc.

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)
- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

Poisson models




```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
```

How much variation?

- Linear model so numbers add up

$$\hat{y}_i = \beta_0 + \beta_1 x_1$$

- So expect spray A to be 2.67, spray E to be 2.67 – 1.42 = 1.25 etc.
- Signs are right, but numbers aren't counts

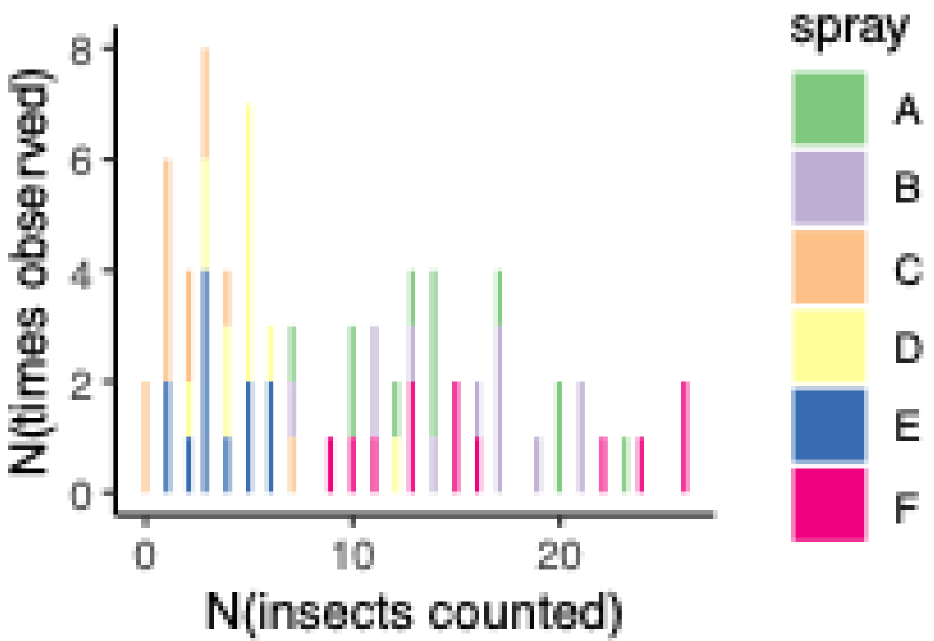
```
> anova(model)
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     71     409.04
spray  5   310.71        66      98.33 < 2.2e-16 ***
---
```

- Coefficients are on a log scale

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models




- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
```

How much variation?

- Linear model so numbers should add up

  $\hat{y}_i = \beta_0 + \beta_1 x_1$
- So expect spray A to be 2.67, spray E to be 2.67 – 1.42 = 1.25 etc.
- Signs are right, but numbers aren't counts

- To reverse the log transform need to exponentiate:

  `exp(coef(model))`

```
> exp(coef(model))
(Intercept)      sprayB      sprayC      sprayD      sprayE      sprayF
 14.5000000   1.0574713   0.1436782   0.3390805   0.2413793   1.1494253
```

- So expect 14.5 insects with spray A
- Expect spray E to give `exp(1.25)` = 3.5

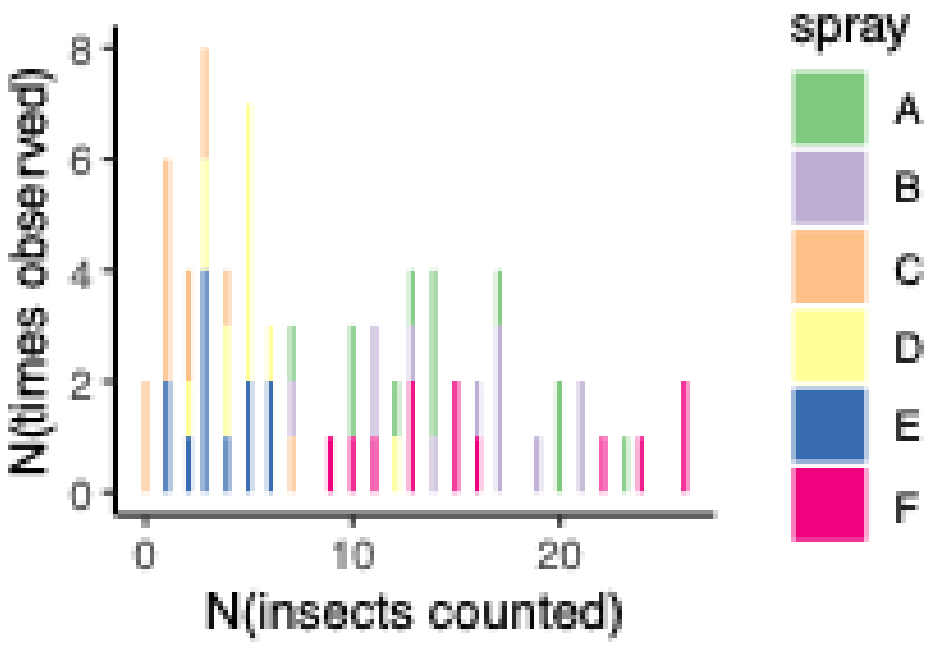
  ( or 14.5 * 0.24 = 3.5 )
- Can get predictions on count scale directly using

  `predict(model, type = "response")`

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models



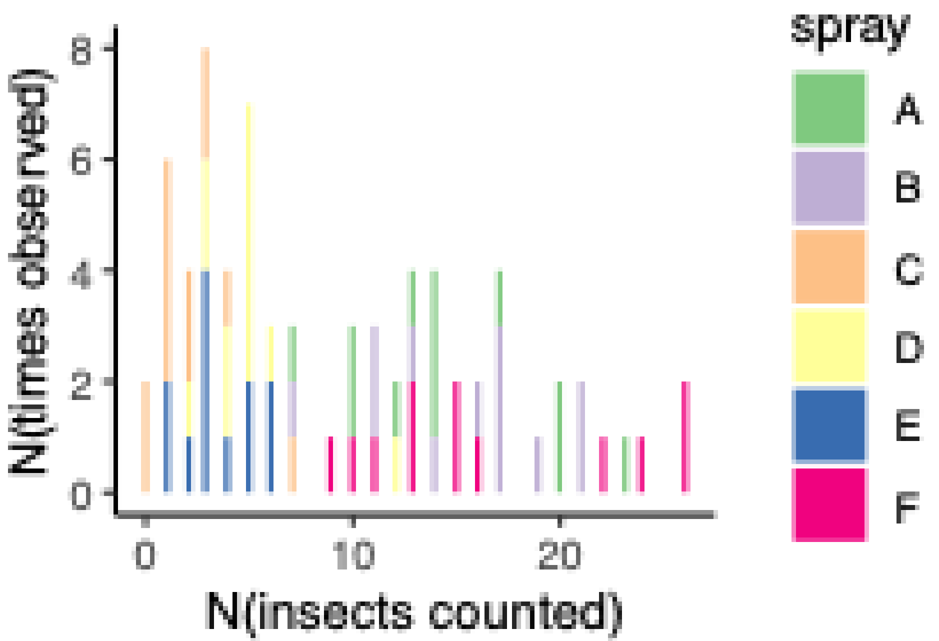

- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used? Yes, $P < 2.2 \times 10^{-16}$

```
> anova(model)
Analysis of Deviance Table

Model: poisson, link: log

Response: count

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     71     409.04
spray  5   310.71        66      98.33 < 2.2e-16 ***
---
```

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models



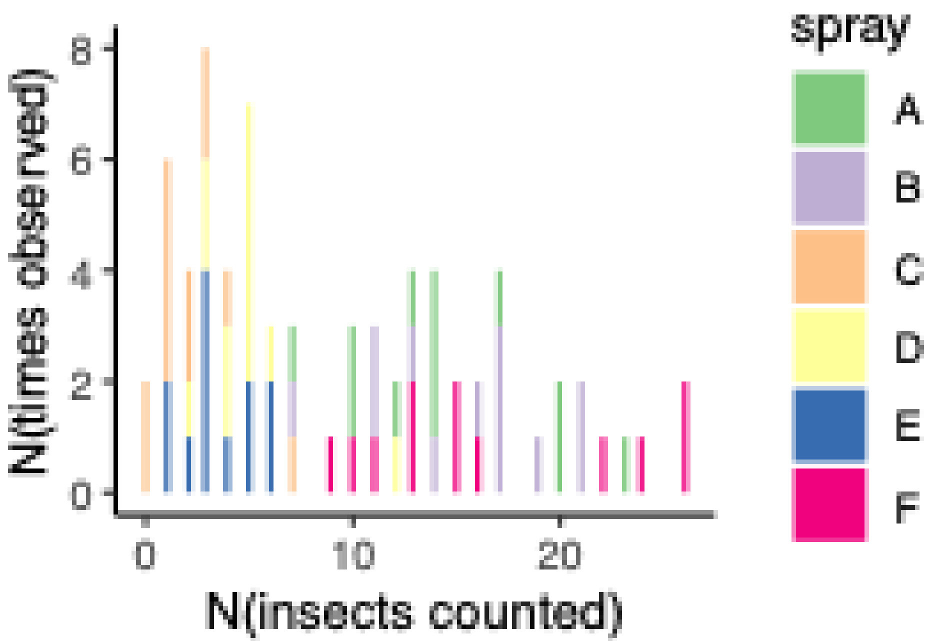

- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used? Yes, $P < 2.2 \times 10^{-16}$
  - How much does insect count change with spray used? Spray E cuts insects to 24% of that in A

```
> exp(coef(model))
(Intercept)      sprayB      sprayC      sprayD      sprayE      sprayF
 14.5000000   1.0574713   0.1436782   0.3390805   0.2413793   1.1494253
```

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models



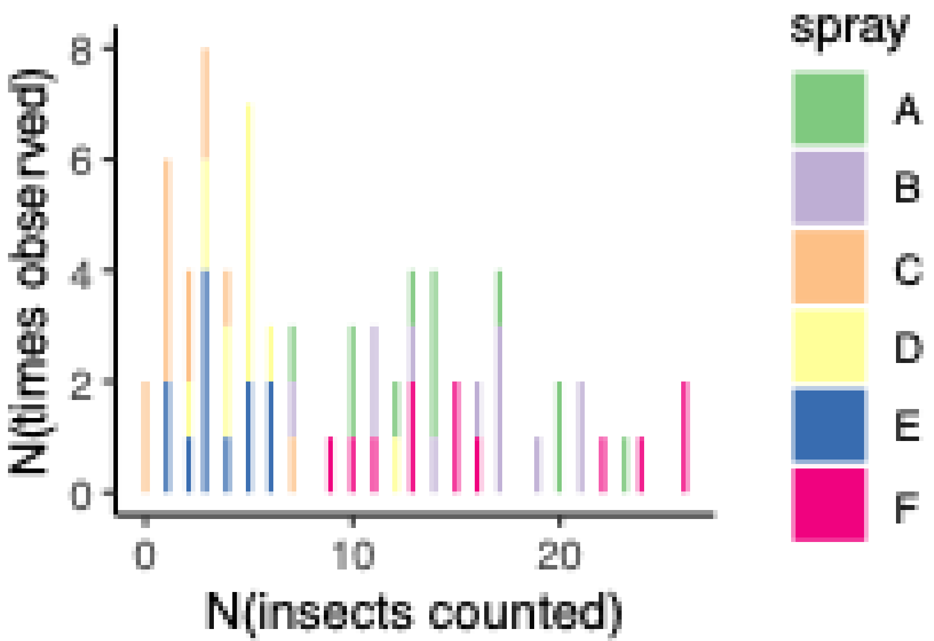

- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used? Yes, $P < 2.2 \times 10^{-16}$
  - How much does insect count change with spray used ? Spray E cuts insects to 24% of that in A
  - What's the expected count for spray A? 14.5 insects (can only count integers, but this is the expected average)

```
> exp(coef(model))
(Intercept)      sprayB      sprayC      sprayD      sprayE      sprayF
 14.5000000   1.0574713   0.1436782   0.3390805   0.2413793   1.1494253
```

# Generalized linear models

- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models




- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used? Yes, $P < 2.2 \times 10^{-16}$
  - How much does insect count change with spray used ? Spray E cuts insects to 24% of that in A
  - What's the expected count for spray A? 14.5 insects (can only count integers, but this is the expected average)
  - How sure are we of the answers? SE and CI on the link scale so CI is exp(2.5) – exp(2.8) i.e. 14.5 (12.5 – 16.8 95% CI)

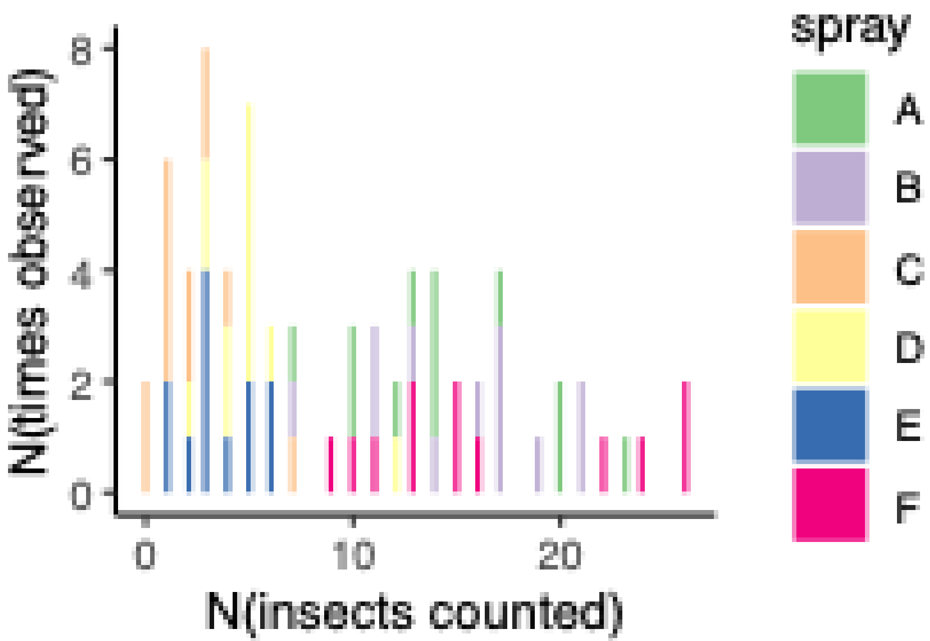
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
```

```
> confint(model)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept)  2.52178693  2.8191381
sprayB      -0.15136099  0.2635229
sprayC      -2.38324541 -1.5411710
sprayD      -1.38456987 -0.7928587
sprayE      -1.77084619 -1.0951362
sprayF      -0.06362048  0.3431038
```

# Generalized linear models

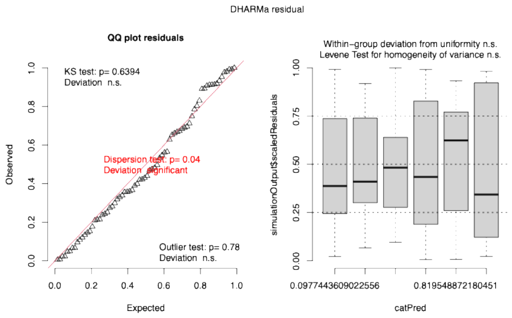
- What if I'm counting insects in a field?
  - Counts: Numeric, categorical – integers only
  - Don't know how many insects there could be
  - Never negative
  - Is an average number (*don't know what it is*)

Poisson models




- Two variables (response and explanatory):
  - Number of insects (categorical numeric)
  - Type of spray (categorical)
- Questions (typical for a 1-way ANOVA)
  - Does number of insects depend on spray used? Yes, $P < 2.2 \times 10^{-16}$
  - How much does insect count change with spray used ? Spray E cuts insects to 24% of that in A
  - What's the expected count for spray A? 14.5 insects (can only count integers, but this is the expected average)
  - How sure are we of the answers? SE and CI on the link scale so CI is exp(2.5) – exp(2.8) i.e. 14.5 (12.5 – 16.8 95% CI)

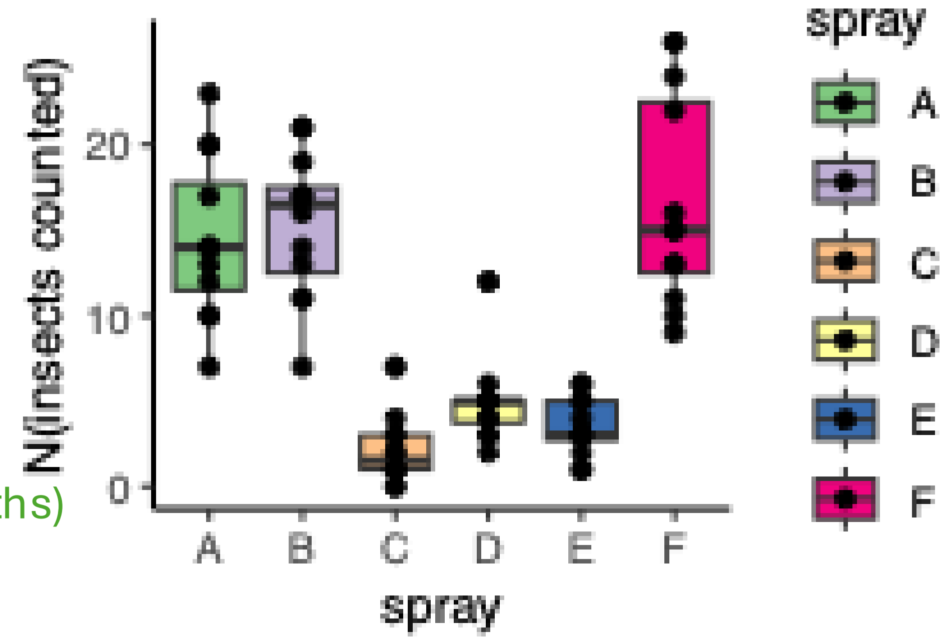
- Take the very similar approach to other families
  - e.g. binomial
- But what if checks fail?
  - (as they did, marginally, here)




```
> confint(model)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept)  2.52178693  2.8191381
sprayB      -0.15136099  0.2635229
sprayC      -2.38324541 -1.5411710
sprayD      -1.38456987 -0.7928587
sprayE      -1.77084619 -1.0951362
sprayF      -0.06362048  0.3431038
```

# Generalized linear models

Poisson models



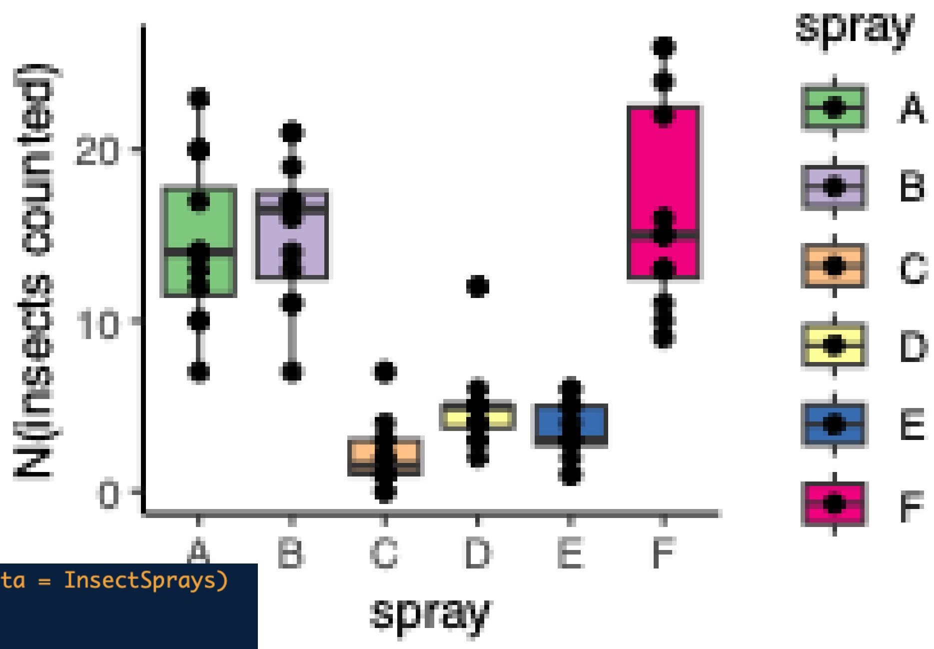

- But what if checks fail?
  - (as they did, marginally, here)
- Think again what your model knows about
  - Is there something missing?
  - Known explanatory variables (covariates) that aren't modelled
- Possibilities here
  - Different insect species have different sensitivity (aphids versus moths)
  - Some fields are windier than others
  - Some fields are closer to insect-rich habitats than others
- Adding in such covariates likely to deal with over-dispersion
- Under-dispersion rarer – could be something problematic with the experiment (non-independence, confounding)
  - All the fields with one spray were either close or far away from a source of insects
- *Only* when done your best with the model and it's *still* over/under dispersed
  - Fit an extra parameter to account for it
- Various options
  - family = quasipoisson
  - Negative binomial

# Generalized linear models

- Various options
  - family = quasipoisson
  - Negative binomial

Poisson models

```
glm(count ~ spray, family = "poisson")
glm(count ~ spray, family = "quasipoisson")
```




- Very similar outcome

```
> model <- glm(count ~ spray, family = poisson, data = InsectSprays)
> summary(model)

Call:
glm(formula = count ~ spray, family = poisson, data = InsectSprays)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67415    0.07581  35.274  < 2e-16 ***
sprayB       0.05588    0.10574   0.528    0.597
sprayC      -1.94018    0.21389  -9.071  < 2e-16 ***
sprayD      -1.08152    0.15065  -7.179 7.03e-13 ***
sprayE      -1.42139    0.17192  -8.268  < 2e-16 ***
sprayF       0.13926    0.10367   1.343    0.179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: 376.59

Number of Fisher Scoring iterations: 5
```

```
> mqp <- glm(count ~ spray, family = quasipoisson, data = InsectSprays)
> summary(mqp)

Call:
glm(formula = count ~ spray, family = quasipoisson, data = InsectSprays)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.67415    0.09309  28.728  < 2e-16 ***
sprayB       0.05588    0.12984   0.430    0.668
sprayC      -1.94018    0.26263  -7.388 3.30e-10 ***
sprayD      -1.08152    0.18499  -5.847 1.70e-07 ***
sprayE      -1.42139    0.21110  -6.733 4.82e-09 ***
sprayF       0.13926    0.12729   1.094    0.278
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.507713)

    Null deviance: 409.041  on 71  degrees of freedom
Residual deviance:  98.329  on 66  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```