evropský sociální fond v ČR

EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

GLM on Count Data and the Poisson Distribution

Juliette Chamagne
Institute of Evolutionary Biology and Environmental Sciences
University of Zurich

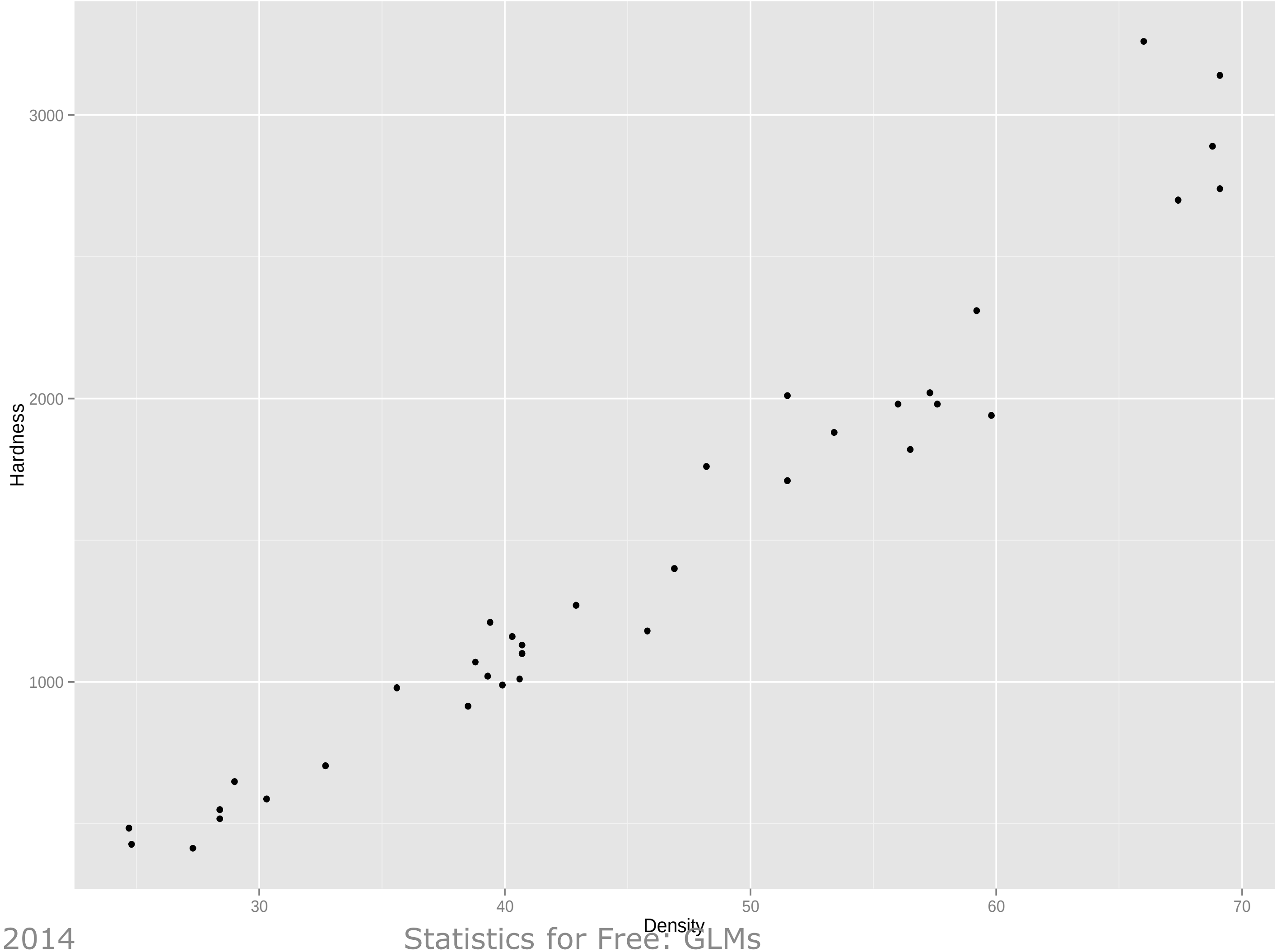
Adapted from:
Statistics for Free:
Generalized Linear Models Using R

By Andy Hector
Former University of Zurich
Now University of Oxford

Recap

- Linear Models (lm) assume:
  - Independence
  - Normality
  - Homogeneity
- Generalized Linear Models (glm) allow:
  - Linear predictor (Y ~ a + b*X_1 + c*X_2…)
  - Family distribution (variance)
  - Link function (mean)

Continuous positive data: Gamma Distribution

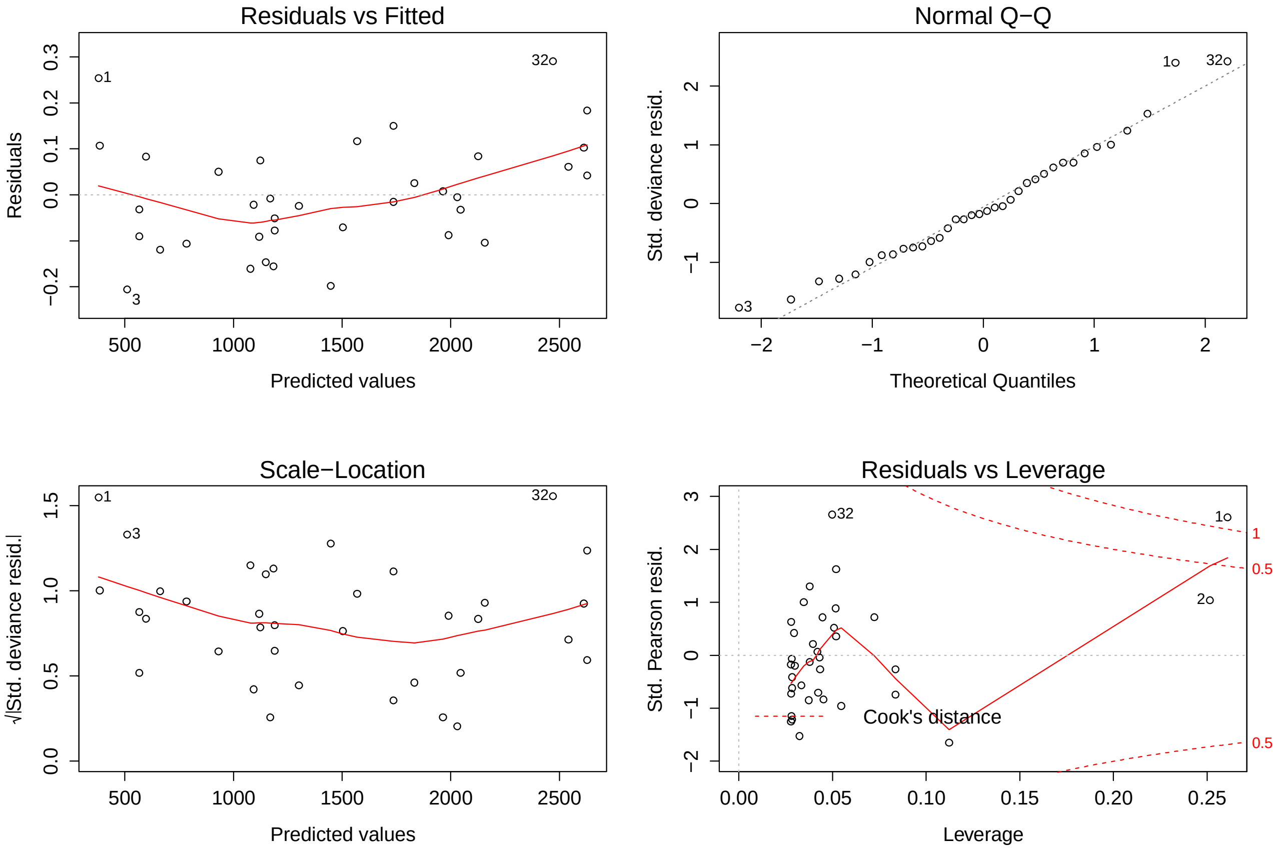


Continuous positive data: Gamma Distribution

```
> Janka.lm <- lm(Hardness ~ Density, data = Janka)
> Janka.glm.Gauss <- glm(Hardness ~ Density, data = Janka, family =
gaussian(link="identity"))
> summary(Janka.glm.Gauss)

Call:
glm(formula = Hardness ~ Density, family = gaussian(link = "identity"),
    data = Janka)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-338.40   -96.98   -15.71    92.71   625.06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1160.500    108.580  -10.69 2.07e-12 ***
Density        57.507      2.279   25.24  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

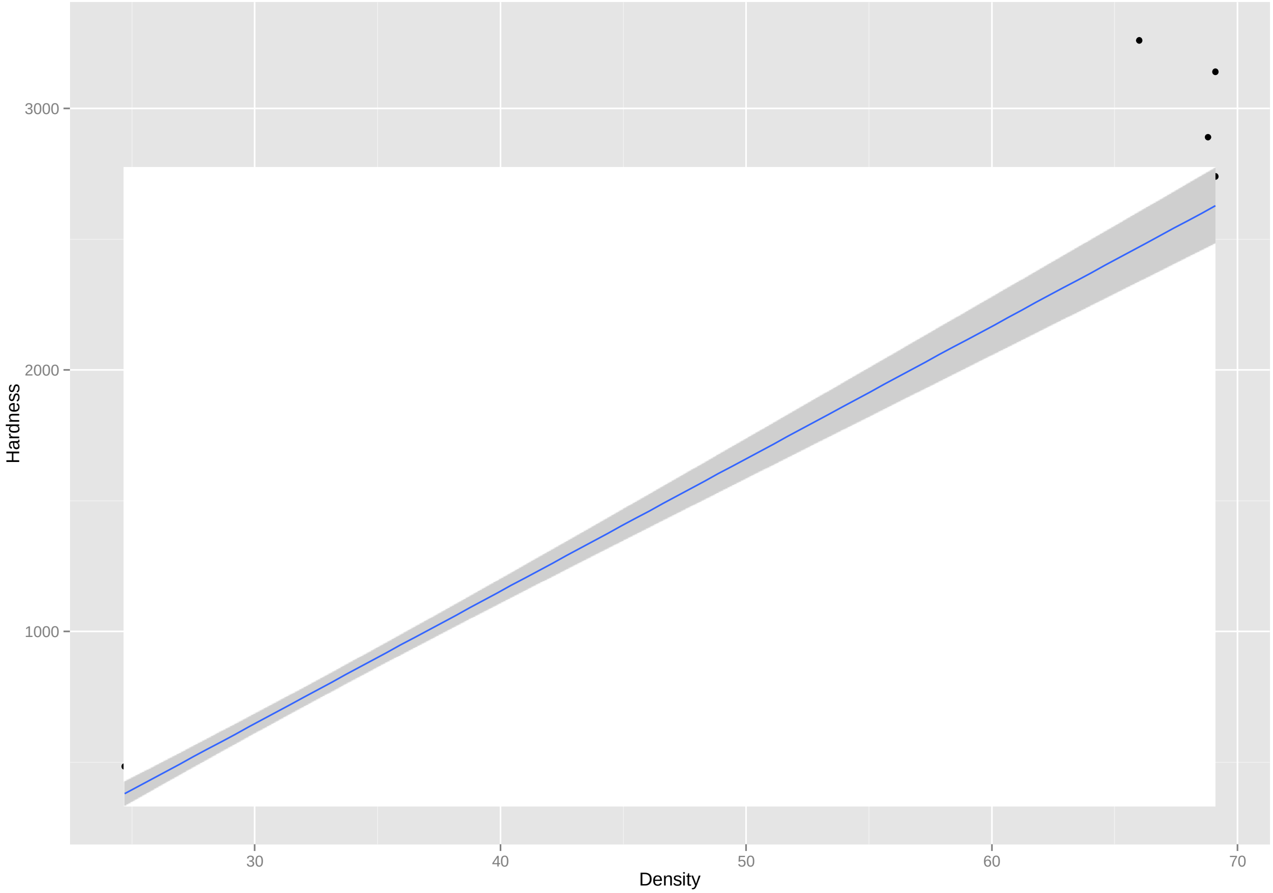
Continuous positive data: Gamma Distribution

```
> Janka.glm.Gamma <- glm(Hardness ~ Density, data = Janka , family =
Gamma(link = "identity"))
```



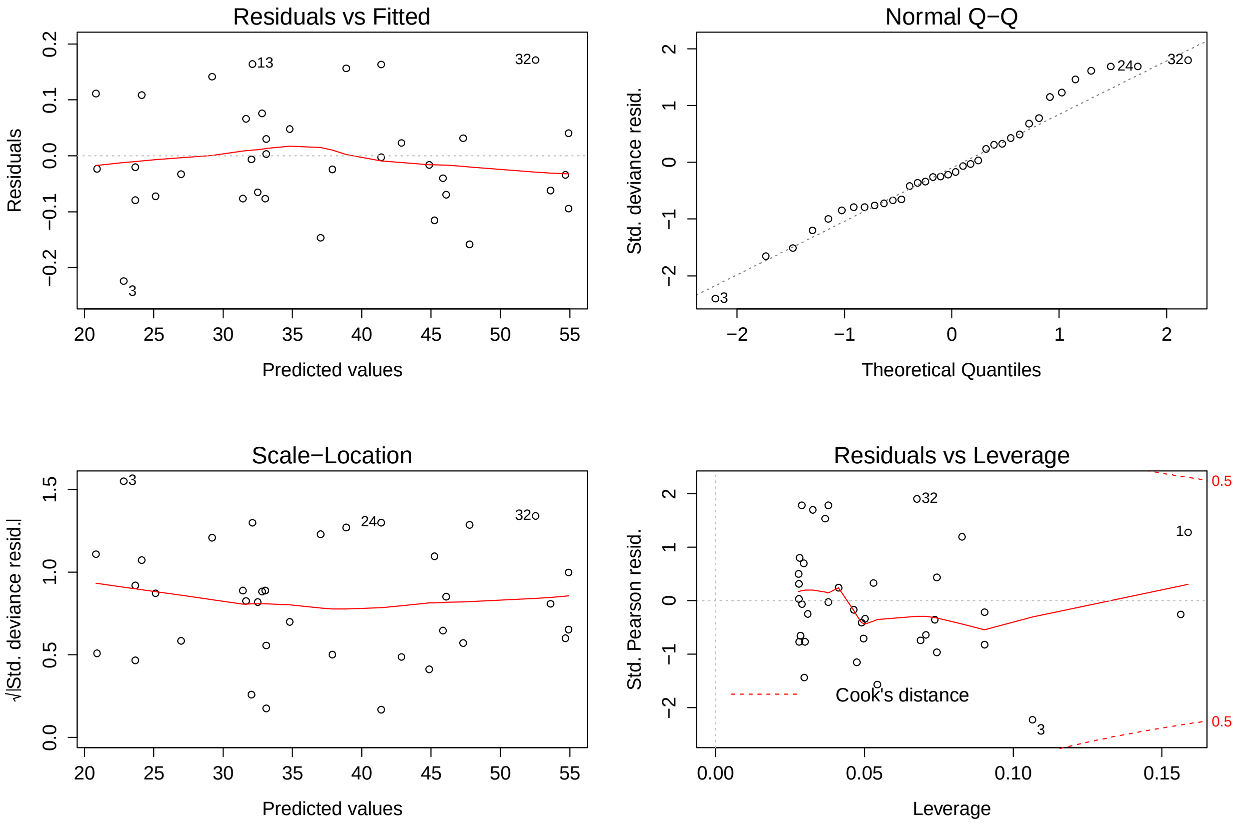
Continuous positive data: Gamma Distribution

```
> Janka.glm.Gamma <- glm(Hardness ~ Density, data = Janka , family =
Gamma(link = "identity"))
```



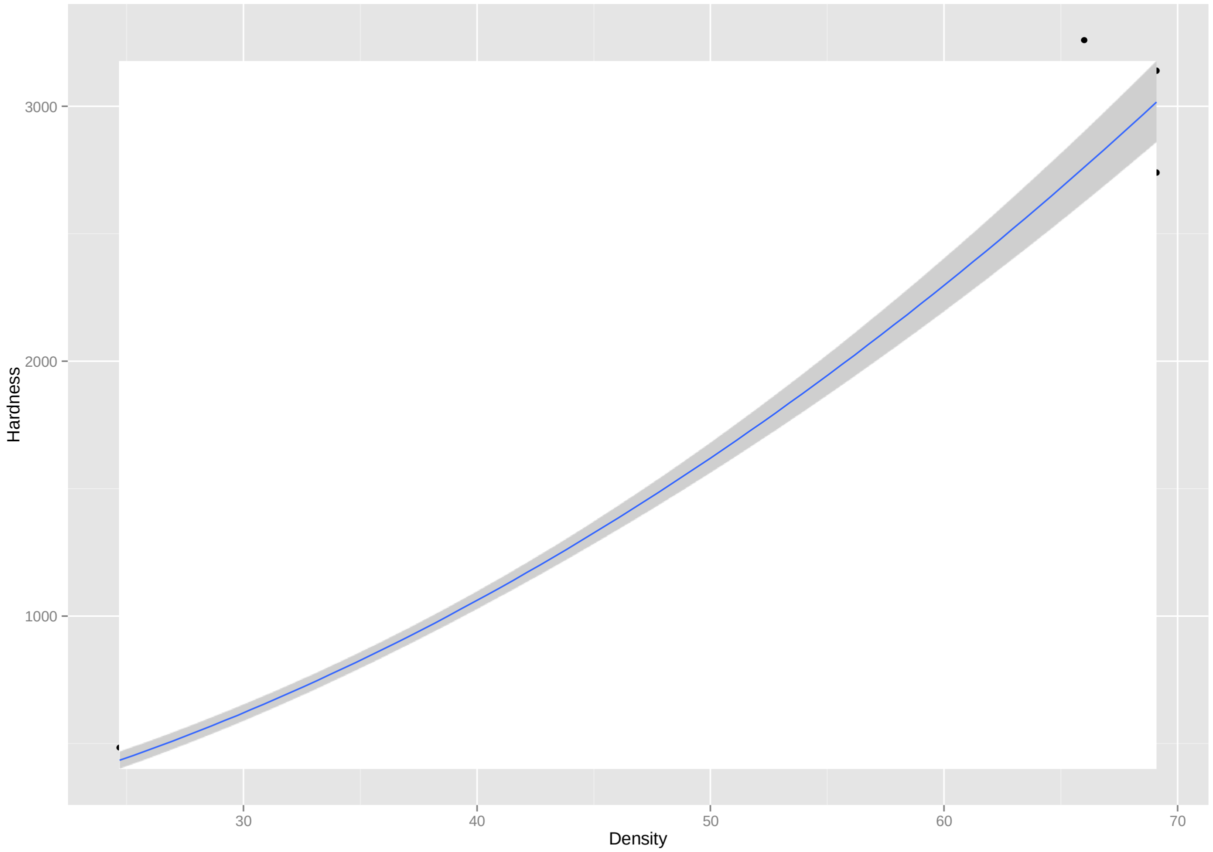
Continuous positive data: Gamma Distribution

```
> Janka.glm.Gamma <- glm(Hardness ~ Density, data = Janka , family =
Gamma(link = ”sqrt"))
```



Continuous positive data: Gamma Distribution

```
> Janka.glm.Gamma <- glm(Hardness ~ Density, data = Janka , family =
Gamma(link = ”sqrt"))
```



Count data

- Data are integers (whole numbers): 0, 1, 2, 3…
- Data are never negative.
- Residuals are restricted in value (can get lines of residuals in residual plots).
- Zeros are often common.

Count data

We know how many times something happened but not how many times it did not. Examples:

number of children per family

number of doctor visits per year

number of species per area

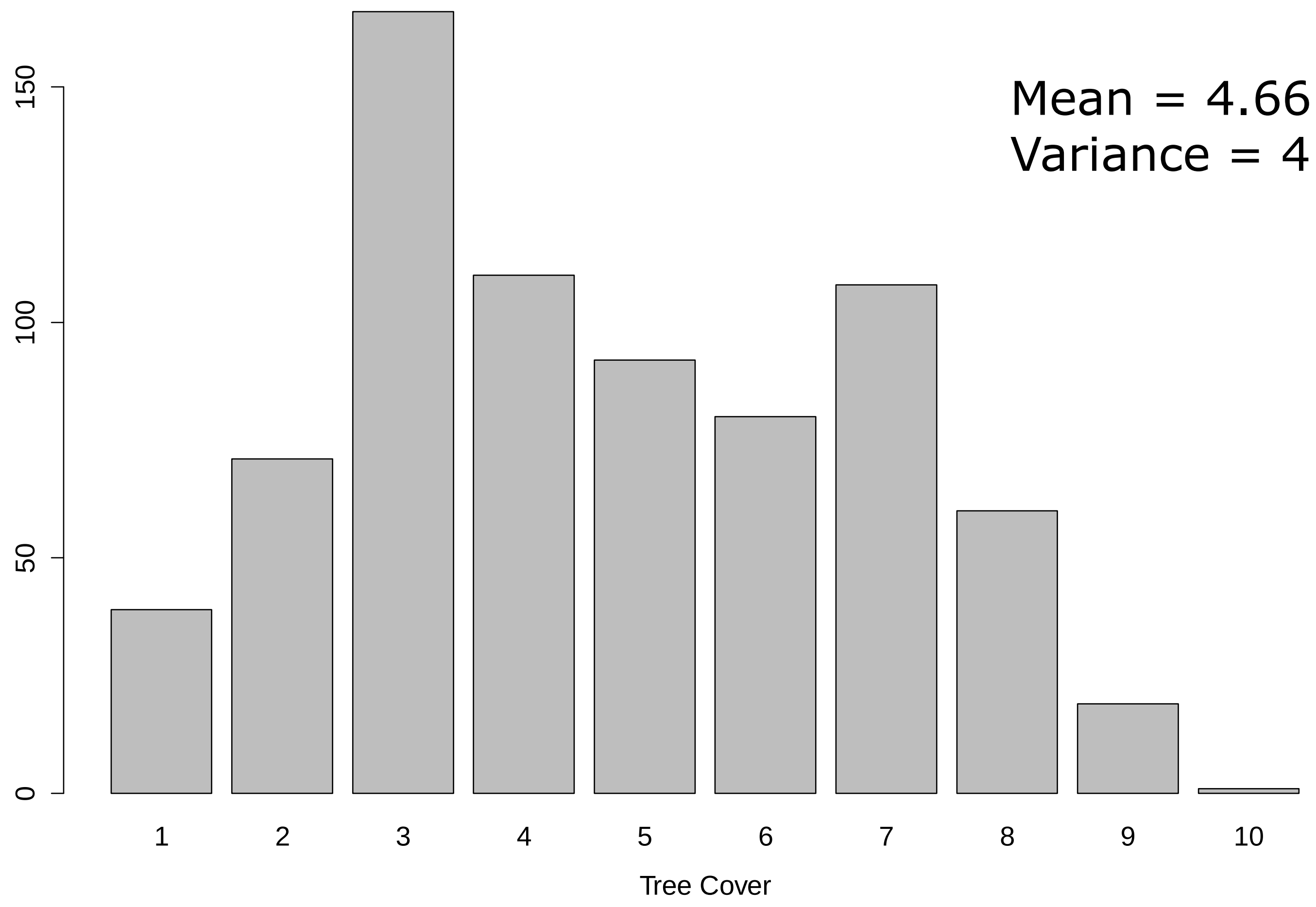
number of individuals from one species per area

tree cover (from 1 to 10) of *Tsuga canadensis*

Count data

Tree cover	Occurrence
1	39
2	71
3	166
4	110
5	92
6	80
7	108
8	60
9	19
10	1

Count data



Mean = 4.66
Variance = 4.47

Count data

- Log-linear models: GLMs with a Poisson errors and log link function.

Mean

Variance

Count data

Mean

- Log link function prevents negative counts since the fitted values are antilogs (exp) and must be positive.

Count data

Variance

Poisson distribution is a one parameter distribution, variance is defined as equal to the mean – when using the Poisson we make this assumption for our data.

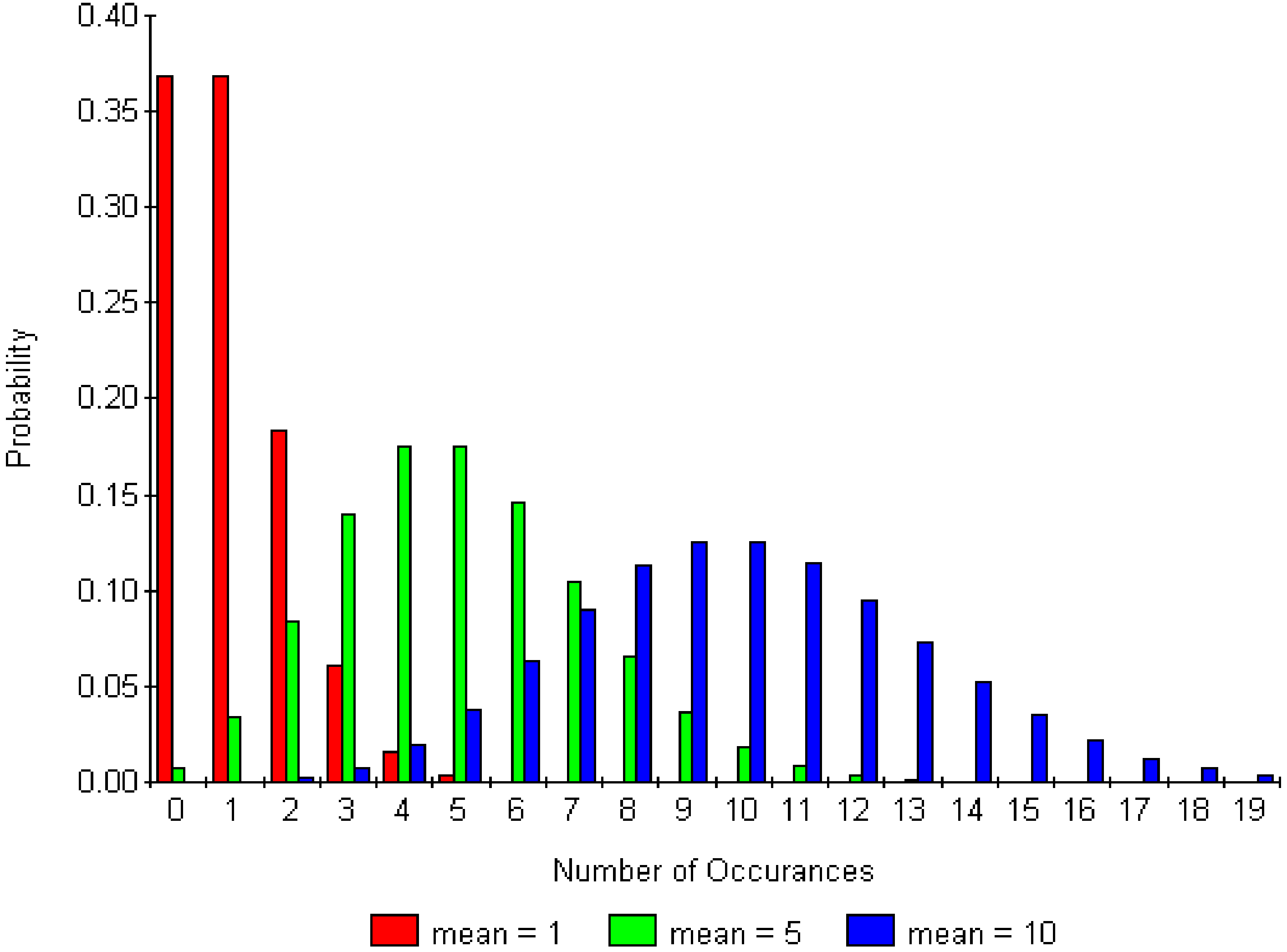
The Poisson Distribution

- The the variance, σ^2 is equal to the mean, μ (mu)

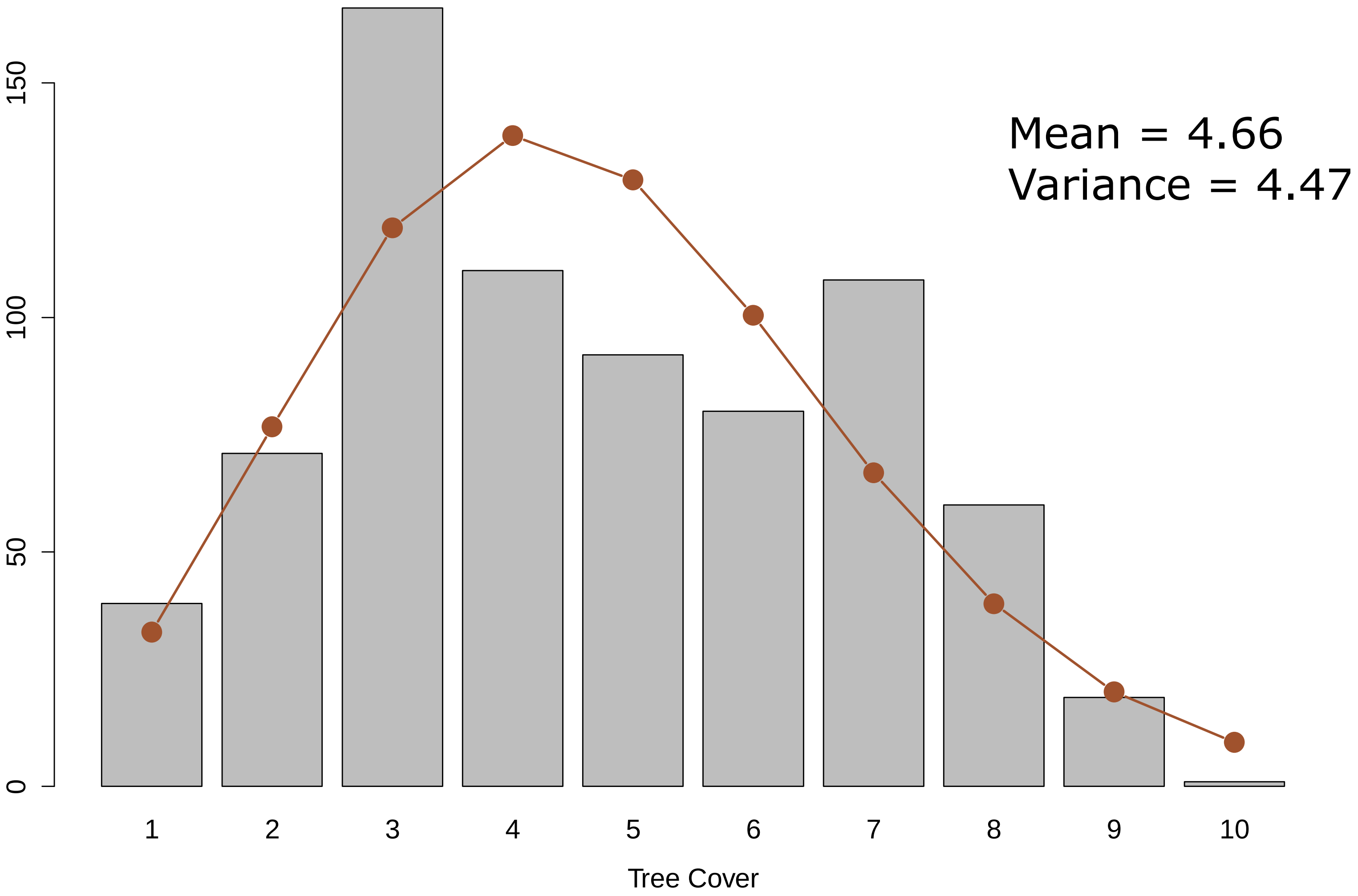
$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

- Zero term: P(0) = $e^{-\mu}$

The Poisson Distribution



The Poisson Distribution



Count data

- Residual deviance is assumed to equal the residual degrees of freedom and the scale parameter is set as one
- Check for over-dispersion and deal with it using QML (Quasi Maximum Likelihood)
- Deviance is once again estimated by an iterative weighted least squares maximum likelihood procedure with its distribution approximately following the chi-squared distribution.

$$2\sum O\ln\left(\frac{O}{E}\right)$$

Count data in R

```
glm(Y ~ X, family=poisson (link = log))
```

Count data in R

If overdispersion (Residual deviance higher than residual degrees of freedom)

glm(Y ~ X, family=**quasipoisson**)

Count data in R

```
> glm3 = glm(cover~elev,data=dat2,family=poisson)
> summary(glm3)

Call:
glm(formula = cover ~ elev, family = poisson, data = dat2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0673  -0.8250  -0.3048   0.9991   2.1347

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.546e+00  5.135e-02  30.115   <2e-16 ***
elev        -8.448e-06  5.471e-05  -0.154    0.877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 749.25  on 745  degrees of freedom
Residual deviance: 749.23  on 744  degrees of freedom
AIC: 3214.2
```

Exercise: the parkgrass experiment

- Counts of species in plots of the Park Grass experiment

```
glm(species ~ biomass, poisson (link = log))
```

Counts of species in plots of the Park Grass experiment

The Park Grass Experiment

Harvesting in 1941

Counts of species in plots of the Park Grass experiment

The Park Grass Experiment

One of the longest running experiment: since 1856

Rothamsted experimental station (England)

Effects of fertilizers on Crop productivity