Zero-Inflated GLM and GLMM

Highland Statistics Ltd.

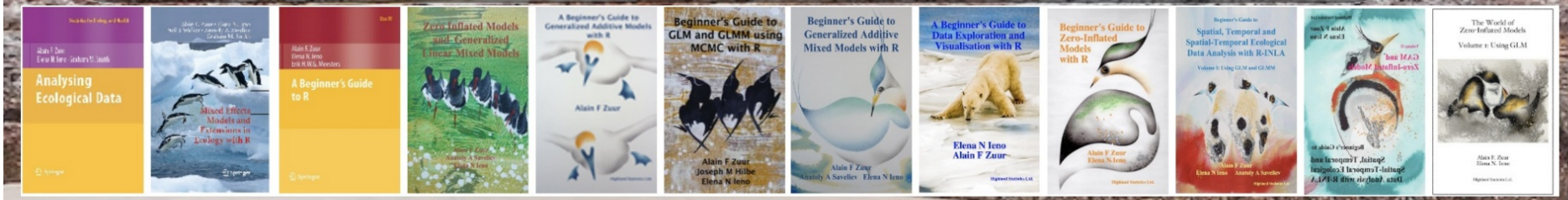
Dr. Alain F. Zuur
highstat@highstat.com
Scotland

Dr. Elena N Ieno
bio@highstat.com
Spain


Introduction
Who are we?

Highland Statistics Ltd.

- We have written 12 statistical text books.
- We teach statistics courses all over the world (and online).
  - Approximately 20 courses per year since 2002.
  - Part of various doctoral programmes.
- We provide statistical consultancy.
- Elena is a biologist.
- Alain is a statistician.
- www.highstat.com

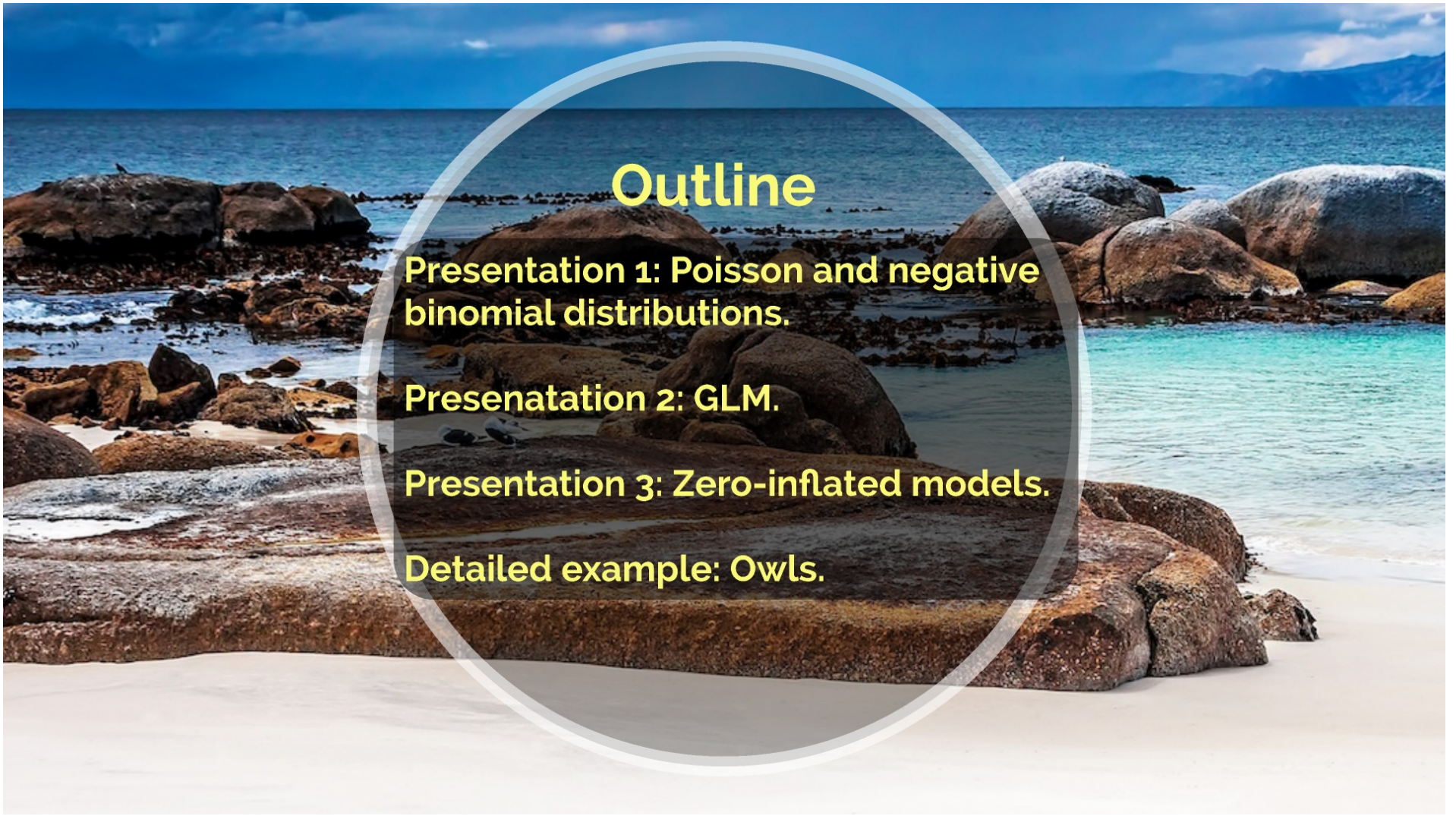


Zero-Inflated GLM and GLMM

Highland Statistics Ltd.

Dr. Alain F. Zuur
highstat@highstat.com
Scotland

Dr. Elena N Ieno
bio@highstat.com
Spain

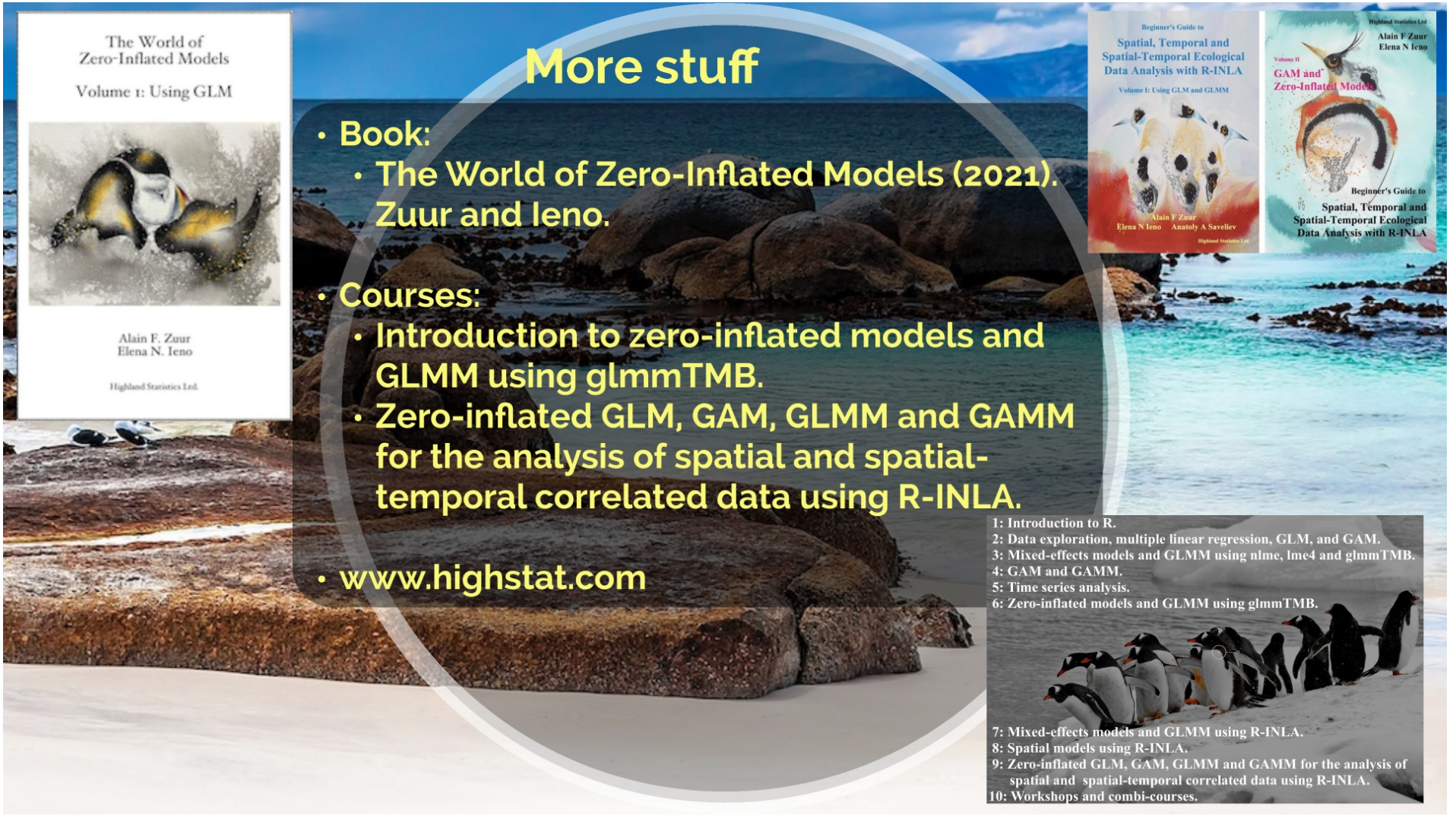

Outline
Presentation 1: Poisson and negative binomial distributions.
Presenatation 2: GLM.
Presentation 3: Zero-inflated models.
Detailed example: Owls.

Zero-Inflated GLM and GLMM

Highland Statistics Ltd.

Dr. Alain F. Zuur
highstat@highstat.com
Scotland

Dr. Elena N Ieno
bio@highstat.com
Spain

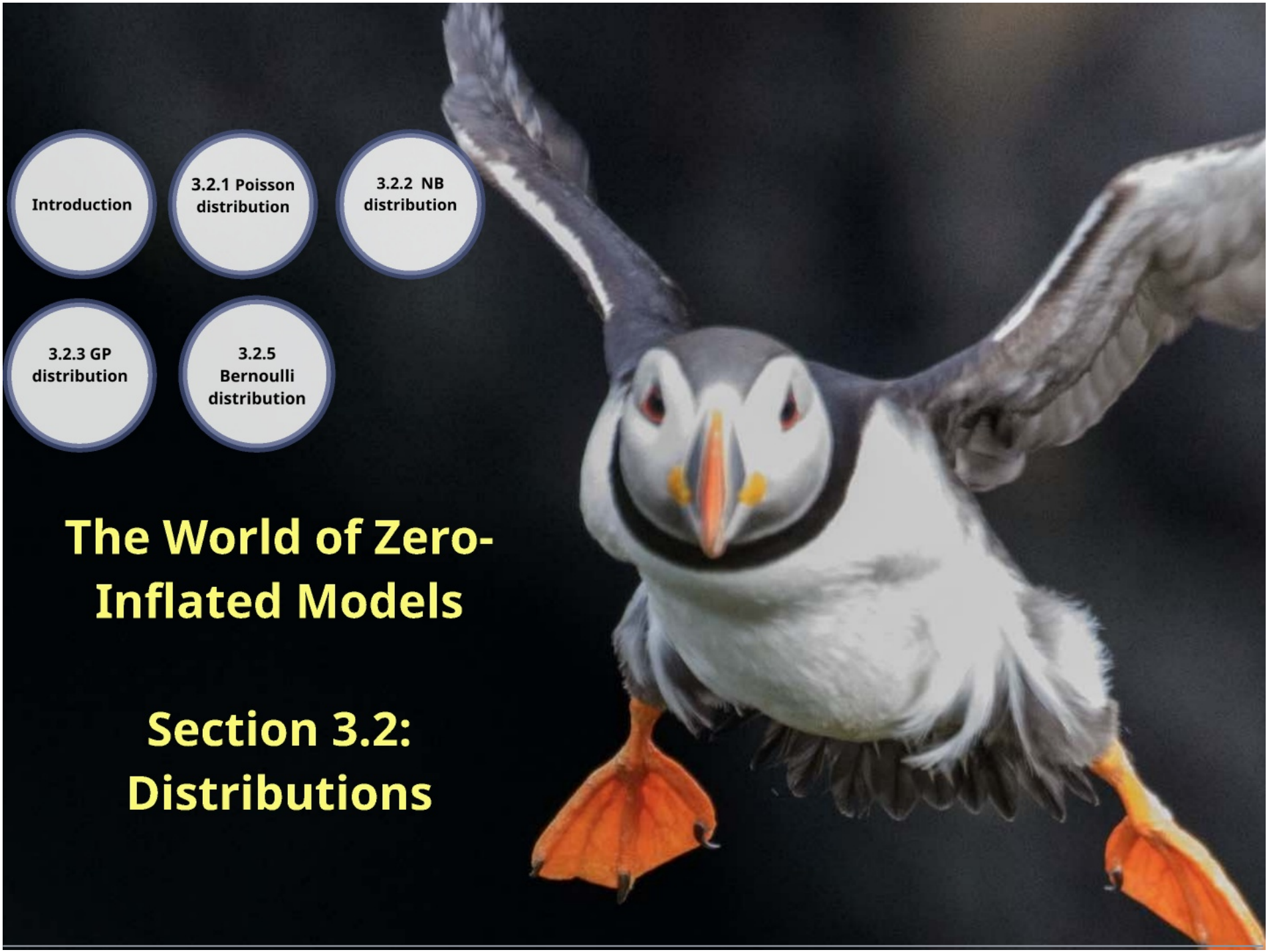

The World of
Zero-Inflated Models
Volume 1: Using GLM
Alain F. Zuur
Elena N. Ieno
Highland Statistics Ltd.
More stuff
• Book:
• The World of Zero-Inflated Models (2021). Zuur and Ieno.
• Courses:
• Introduction to zero-inflated models and GLMM using glmmTMB.
• Zero-inflated GLM, GAM, GLMM and GAMM for the analysis of spatial and spatial-temporal correlated data using R-INLA.
• www.highstat.com
Beginner's Guide to
Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA
Volume I: Using GLM and GLMM
Alain F Zuur
Elena N Ieno
Anatoly A Saveliev
Highland Statistics Ltd
Highland Statistics Ltd
Alain F Zuur
Elena N Ieno
Volume II
GAM and Zero-Inflated Models
Beginner's Guide to
Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA
1: Introduction to R.
2: Data exploration, multiple linear regression, GLM, and GAM.
3: Mixed-effects models and GLMM using nlme, lme4 and glmmTMB.
4: GAM and GAMM.
5: Time series analysis.
6: Zero-inflated models and GLMM using glmmTMB.
7: Mixed-effects models and GLMM using R-INLA.
8: Spatial models using R-INLA.
9: Zero-inflated GLM, GAM, GLMM and GAMM for the analysis of spatial and spatial-temporal correlated data using R-INLA.
10: Workshops and combi-courses.

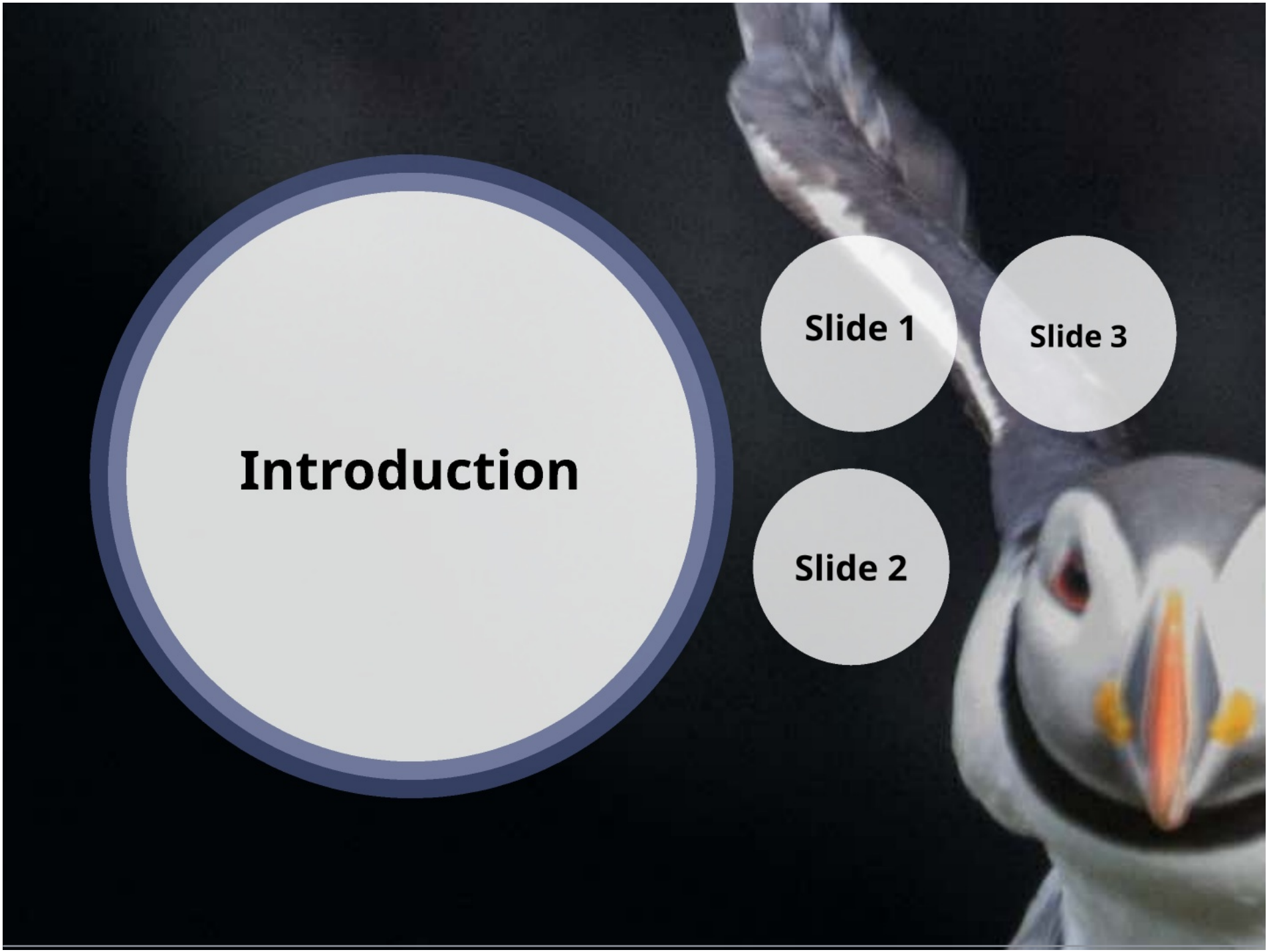
Zero-Inflated GLM and GLMM

Highland Statistics Ltd.

Dr. Alain F. Zuur
highstat@highstat.com
Scotland

Dr. Elena N Ieno
bio@highstat.com
Spain


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


Introduction
Slide 1
Slide 3
Slide 2

In this chapter:

- Revision of the Poisson distribution.
- Revision Poisson GLM for the analysis of count data.
- Discuss 2 other relevant distributions for zero-inflated count data.
  - Negative binomial (NB) models.
  - Generalized Poisson (GP) models.

- The GP can also be used to deal with underdispersion.

Distributions

A GLM consists of 3 steps:

- The distribution for the response variable.
- A predictor function specifying the covariates.
- Link between the expected values of the distribution and the predictor function.

In the second part of the presentation:

- Show how to implement these models in the frequentist package glmmTMB (Magnusson et al., 2021), using a real data set

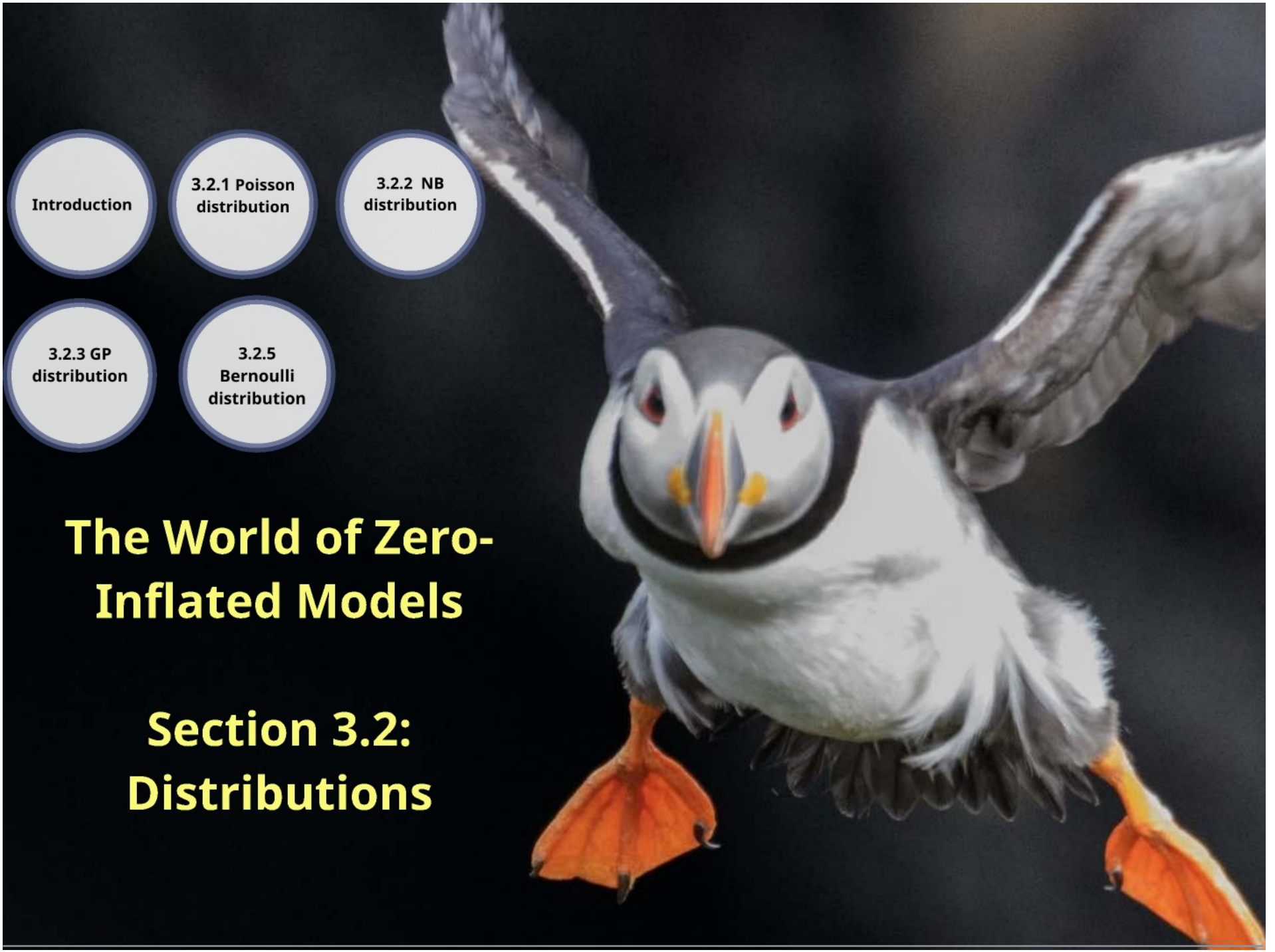
Benefit of applying these models in a frequentist approach:

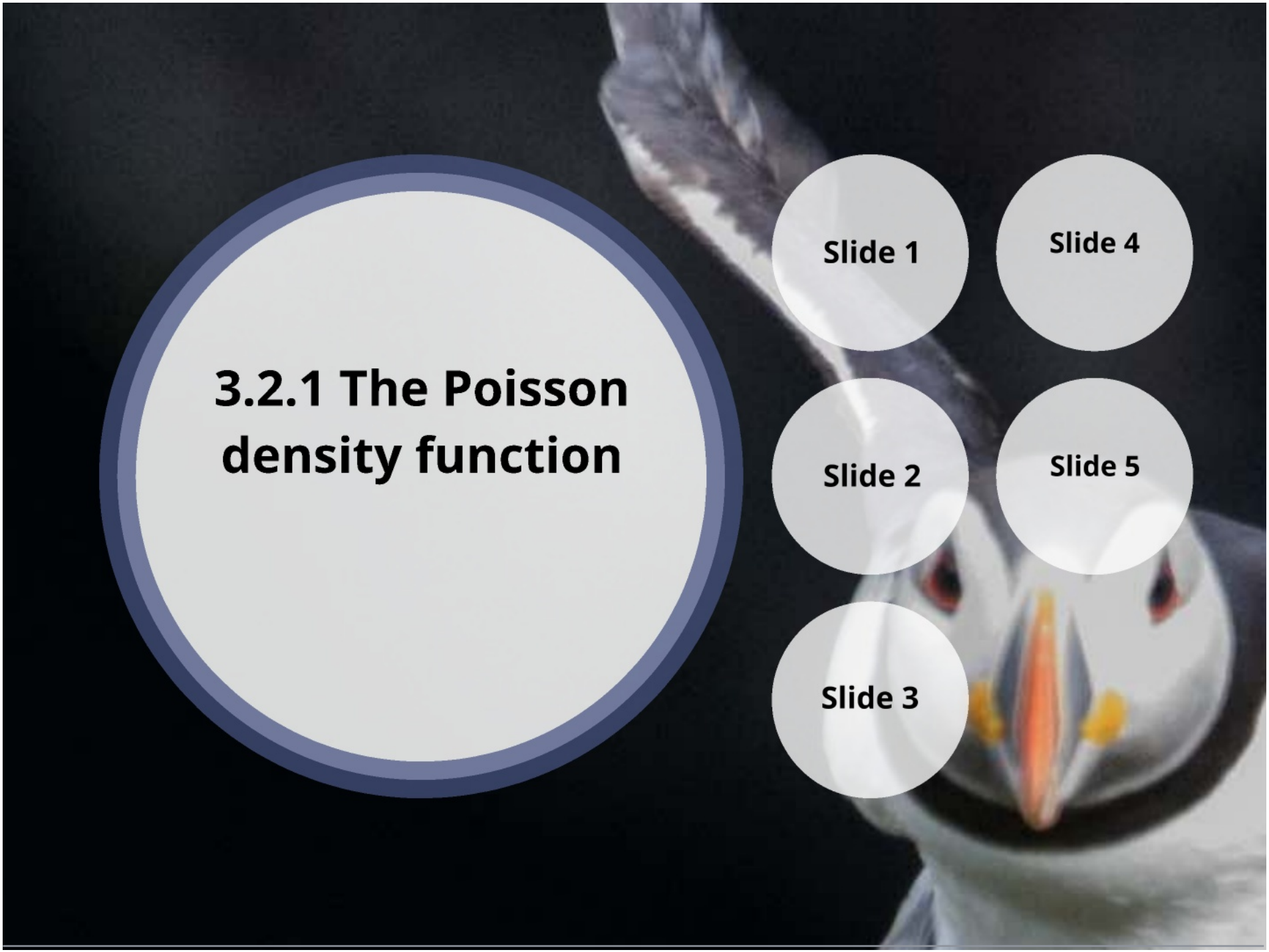
- Relatively easy.
- We can use the glmmTMB package, which is rapidly gaining popularity.

Never start with a zero-inflated model!

- First try an ordinary GLM.
- A covariate might also be able to model the large number of zeros.

- This will be the subject of the next chapter!


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


3.2.1 The Poisson density function
Slide 1
Slide 4
Slide 2
Slide 5
Slide 3

In order to get familiar with zero inflated models:

- Need to know a little bit about the mathematical background.
- Math is not too difficult.

Starting point:

- Poisson density function.

Example:

- Suppose that we go to a puffin colony on the cliffs.
- Using a pair of binoculars, we scan an area that is about 35 meters from the edge.

In order to get familiar with zero inflated models:

- Need to know a little bit about the mathematical background.
- Math is not too difficult.

Starting point:

- Poisson density function.

Example:

- Suppose that we go to a puffin colony on the cliffs.
- Using a pair of binoculars, we scan an area that is about 35 meters from the edge.

Suppose that we also have a magic hat that tells us that the average number of puffin nests in a 3-by-3-meter plot is 5.

We do not expect to observe exactly 5 nests;

- Maybe we will count 0, 1 or 6 nest.

By using a statistical distribution, we can answer the question:

- What is the probability that we will observe 0 nests,
- 1 nest,
- 6 nests,
- or 10 nests,

Given that we know that the average is 5 nests in a 3-by-3 plot that is 35 meters from the cliff edge?

To quantify these probabilities we use a statistical distribution.

Suppose that we also have a magic hat that tells us that the average number of puffin nests in a 3-by-3-meter plot is 5.

We do not expect to observe exactly 5 nests;

- Maybe we will count 0, 1 or 6 nest.

By using a statistical distribution, we can answer the question:

- What is the probability that we will observe 0 nests,
- 1 nest,
- 6 nests,
- or 10 nests,

Given that we know that the average is 5 nests in a 3-by-3 plot that is 35 meters from the cliff edge?

To quantify these probabilities we use a statistical distribution.

Since we are working with counts the obvious choice is the Poisson distribution.

This distribution is given by:

$$f(Y|\mu) = P(Y = y|\mu) = \frac{\mu^y \times e^{-\mu}}{y!} \qquad (3.1)$$

! stands for 'factorial'.

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

$0! = 1$

Y is the number of puffin nests

μ is the mean

' |' means 'for given'

Use the Poisson distribution to calculate the probability that 0, 1, 2 or 10 nests are observed ..... given that the mean μ = 5.

$$
\begin{aligned}
P(Y=0|\mu=5) &= \frac{5^0 \times e^{-5}}{0!} &= 0.006 \\
P(Y=1|\mu=5) &= \frac{5^1 \times e^{-5}}{1!} &= 0.033 \\
P(Y=5|\mu=5) &= \frac{5^5 \times e^{-5}}{5!} &= 0.175 \\
P(Y=10|\mu=5) &= \frac{5^{10} \times e^{-5}}{10!} &= 0.018
\end{aligned}
\qquad (3.2)
$$

What if the mean is μ = 10?

- How does the distribution look in this case?

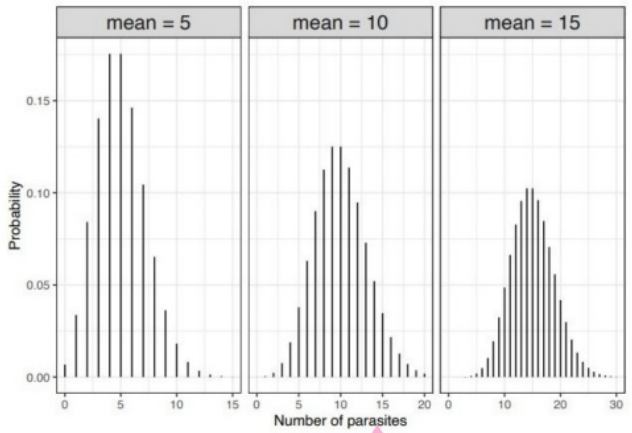
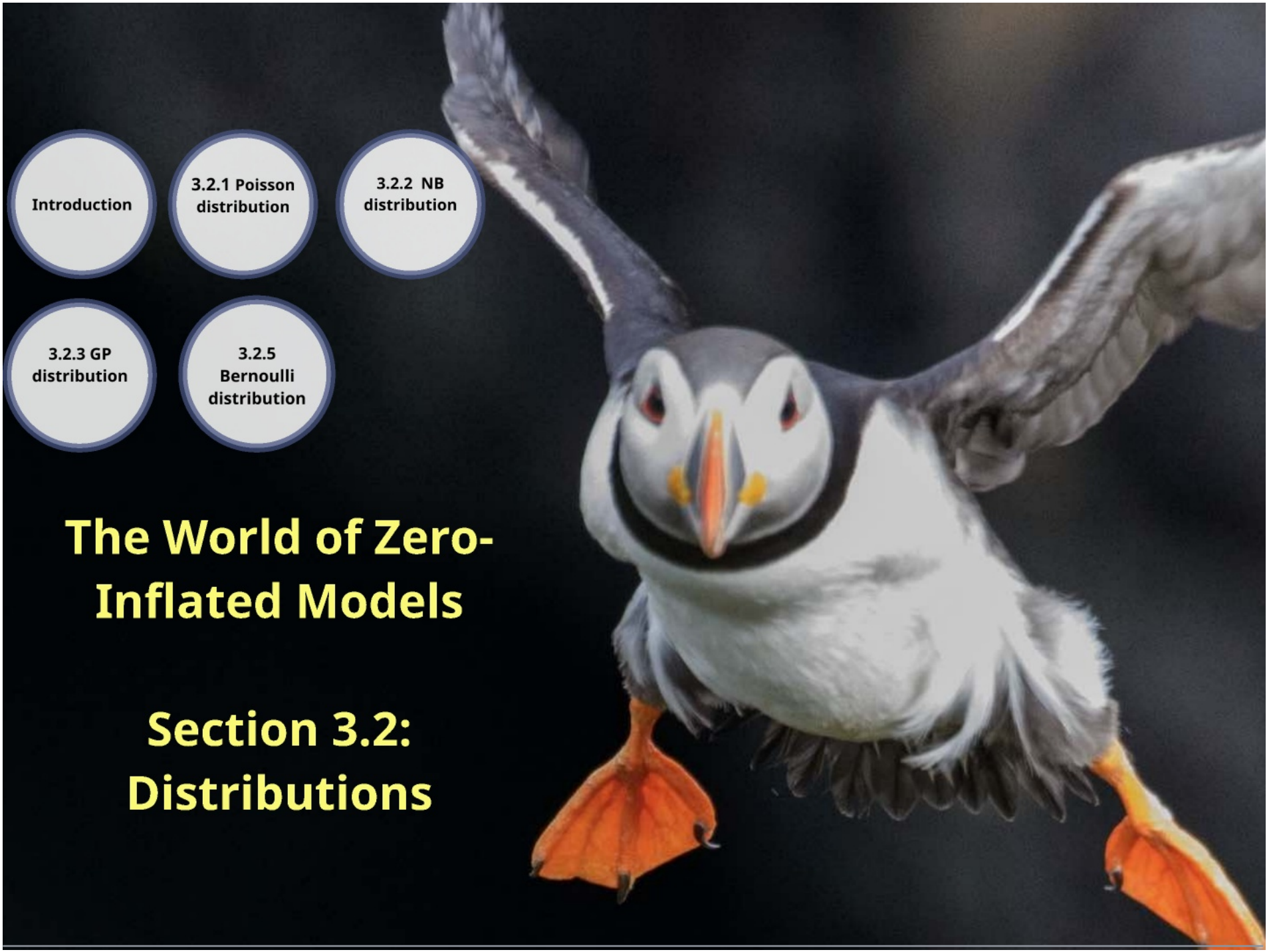


FIGURE 3.2: Poisson distribution for various mean values. Note that the horizontal range differs per panel.

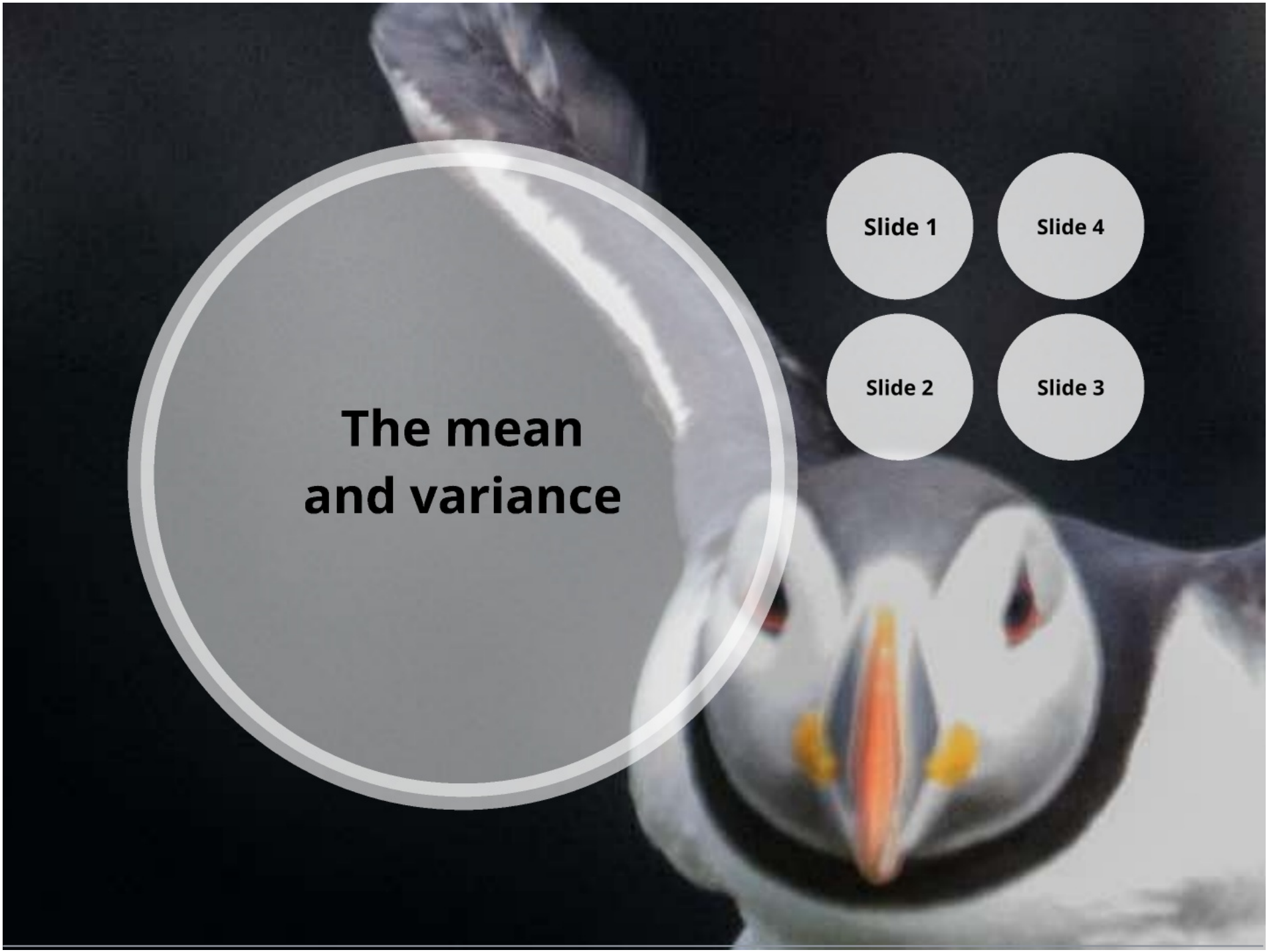
If we assume that the variable Nests_i follows a Poisson distribution, then its mean and variance are as follows.

$$\text{E}\left[\text{Nests}_i\right] = \text{var}\left[\text{Nests}_i\right] = \mu_i$$

The subscript i refers to the plot. In this case $i = 1, \cdots, 38$.


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


3.2.2 The Negative Binomial distribution
NB density function


The mean
and variance
Slide 1
Slide 4
Slide 2
Slide 3

Negative Binomial

- For most ecological data sets the mean–variance relationship as imposed by the Poisson density function does not hold!
- The results obtained with the Poisson distribution are invalid.
- If the variance is larger than the mean, we can use the NB distribution (Hilbe, 2014).
- The mean and variance are given by:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i + \frac{\mu_i^2}{\theta} \end{aligned} \qquad (3.3)$$

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i + \frac{\mu_i^2}{\theta} \end{aligned} \tag{3.3}$$

Negative Binomial

- The variance has an extra variable, namely θ.

- **If θ is large relative to μ_i^2:**
  - The term μ_i^2 / θ is close to 0.
  - The variance of Nest_i is close to μ_i .
  - NB converges to the Poisson distribution

NB distributions

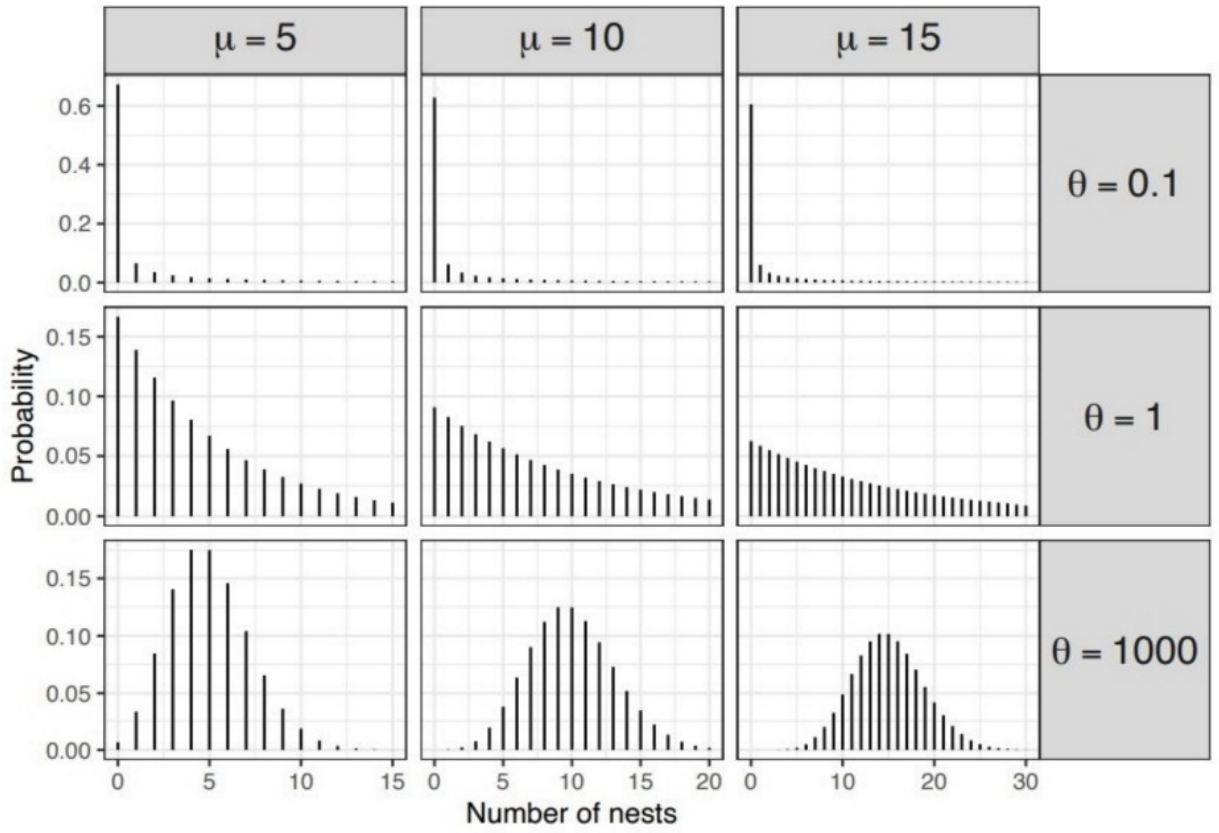


FIGURE 3.3: NB distribution functions for various mean values μ and θ.

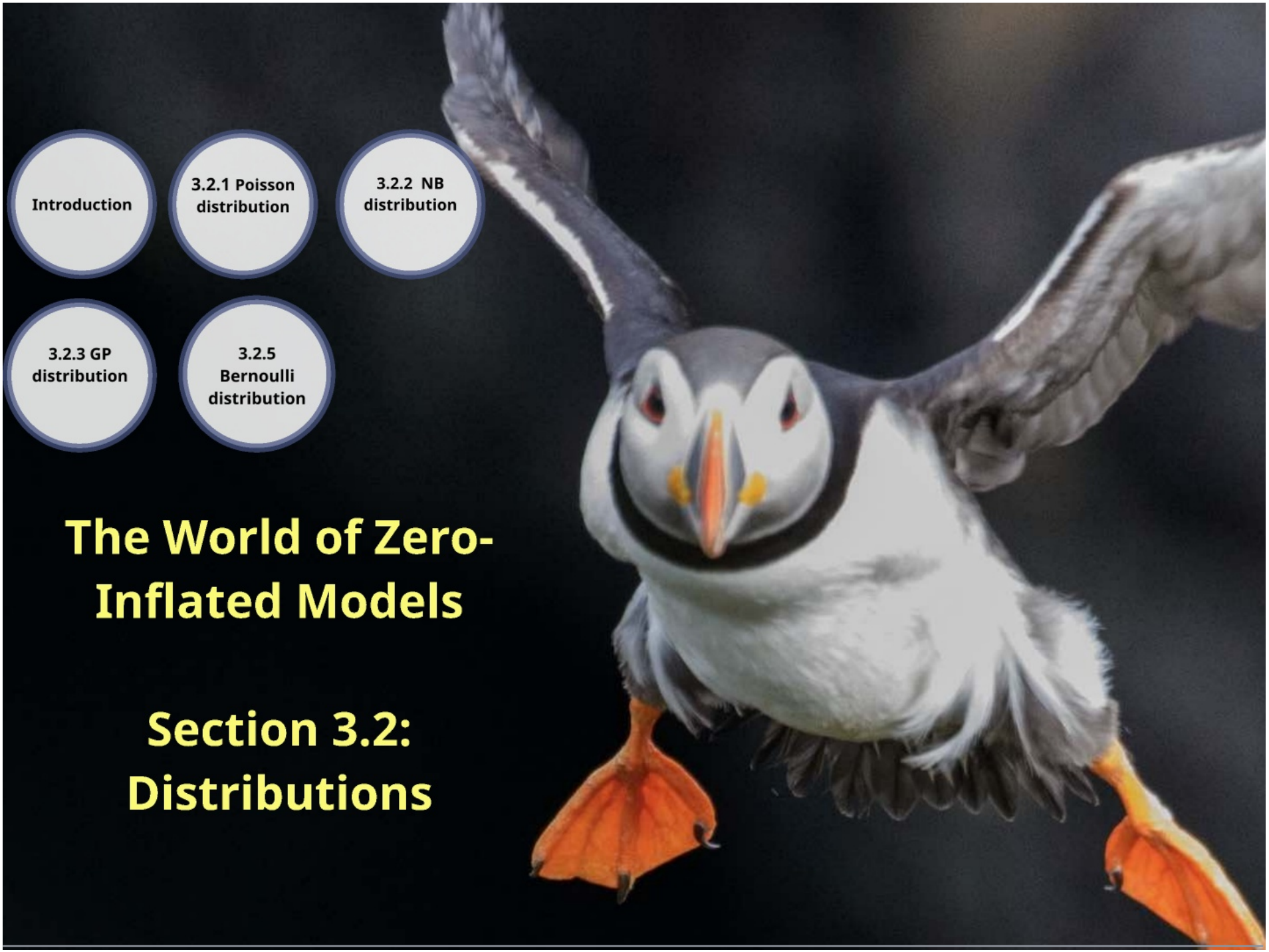
Recall...

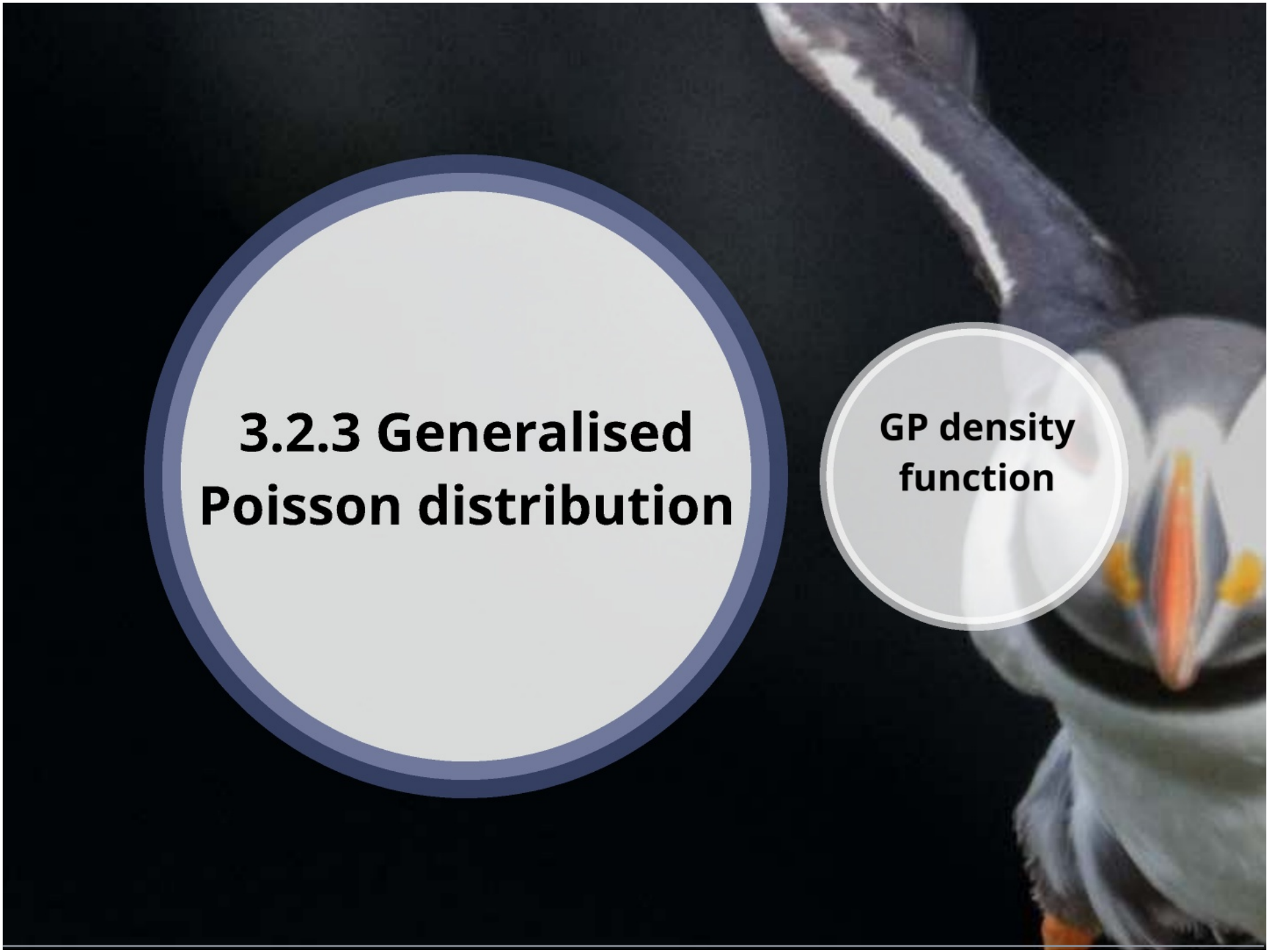
- For bivariate linear regression model with a normal distribution, the observed data should be scattered around the regression line, with approximately equal numbers of observations above and below this line

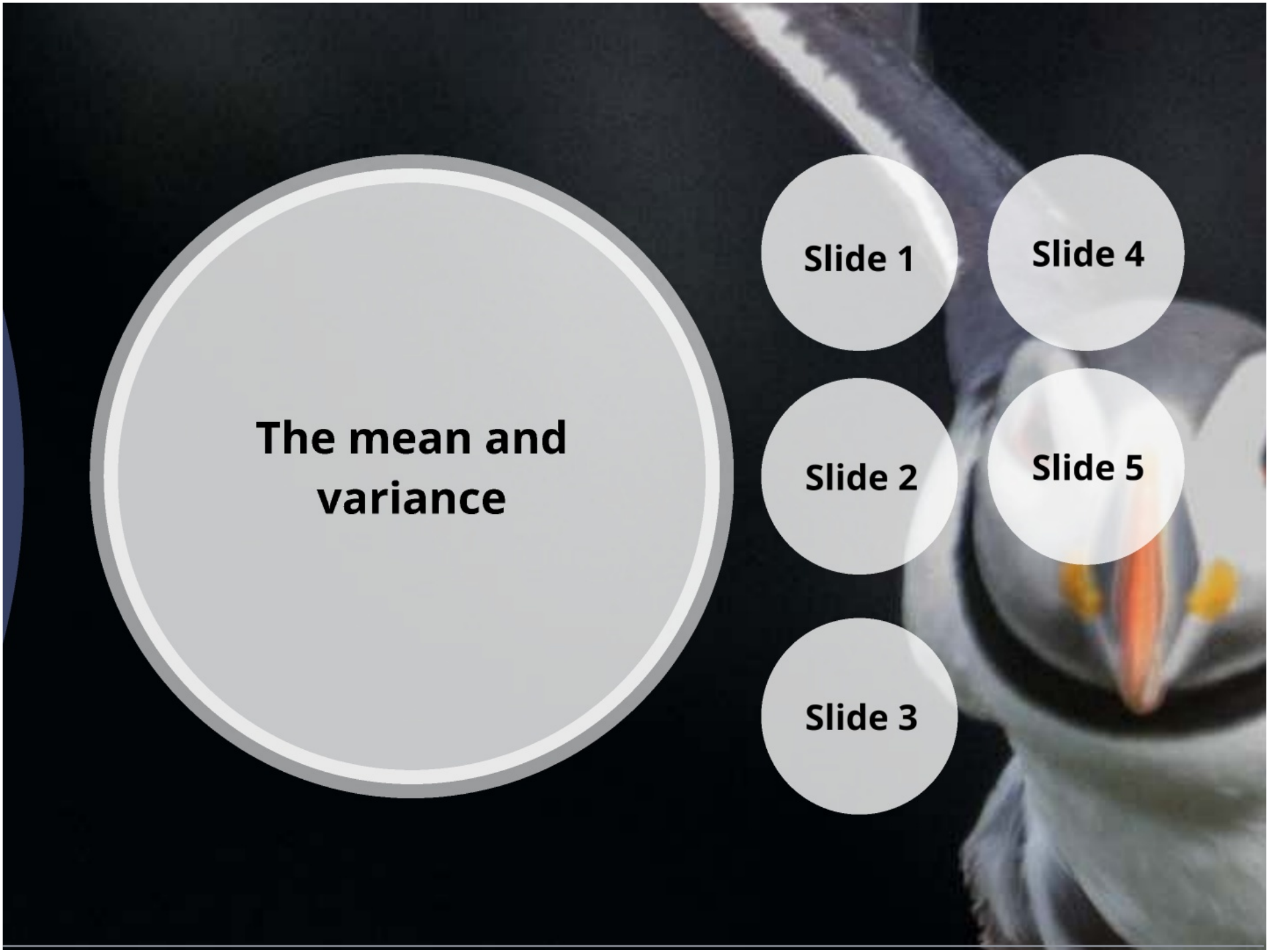
This is not necessarily for the NB distribution.

The data are not necessarily centred around their mean value.

The NB distribution allows for the situation that the majority of the observations are below the average and some observations that are considerably larger than the mean.


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


3.2.3 Generalised Poisson distribution
GP density function


The mean and variance
Slide 1
Slide 2
Slide 3
Slide 4
Slide 5

Generalized Poisson distribution (Consul 1989)

- A highly undervalued distribution.
- A good competitor for the NB distribution.
- Easy to implement in the package glmmTMB.
- Also available for (zero-inflated) GLMs and GLMMs .

The GP distribution can be used if:

- The variation in the data is larger than allowed for by the Poisson mean–variance relationship (overdispersion).
- The variation is smaller than the mean (underdispersion).

Generalized Poisson distribution (Consul 1989)

- A highly undervalued distribution.
- A good competitor for the NB distribution.
- Easy to implement in the package glmmTMB.
- Also available for (zero-inflated) GLMs and GLMMs .

The GP distribution can be used if:

- The variation in the data is larger than allowed for by the Poisson mean–variance relationship (overdispersion).
- The variation is smaller than the mean (underdispersion).

There are various ways to express the mean and variance of a variable that follows a GP distribution.

The glmmTMB package uses:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

There are various ways to express the mean and variance of a variable that follows a GP distribution.

The glmmTMB package uses:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

GP Distribution

- Just as for the Poisson GLM, the mean is *μ_i*.
- The variance term is different.
- GP allows for overdispersion if λ > 1.
- GP allows for underdispersion if 1/(1 − λ)^2 is smaller than 1.

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

GP Distribution

- Just as for the Poisson GLM, the mean is *μ_i*.
- The variance term is different.
- GP allows for overdispersion if λ > 1.
- GP allows for underdispersion if 1/(1 − λ)^2 is smaller than 1.

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

glmmTMB reports $\phi = 1/(1-\lambda)^2$

If $\phi > 1$, it assumes overdispersion.

If $\phi < 1$, it allows underdispersion.

If $\phi = 1$, then we obtain the Poisson.

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \tfrac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

glmmTMB reports $\phi = 1/(1-\lambda)^2$

If $\phi > 1$, it assumes overdispersion.

If $\phi < 1$, it allows underdispersion.

If $\phi = 1$, then we obtain the Poisson.

GP density functions

dgenpois1, VGAM package (Yee, 2021)

This function does not allow for $\phi < 1$ (cannot use it to visualise underdispersion)

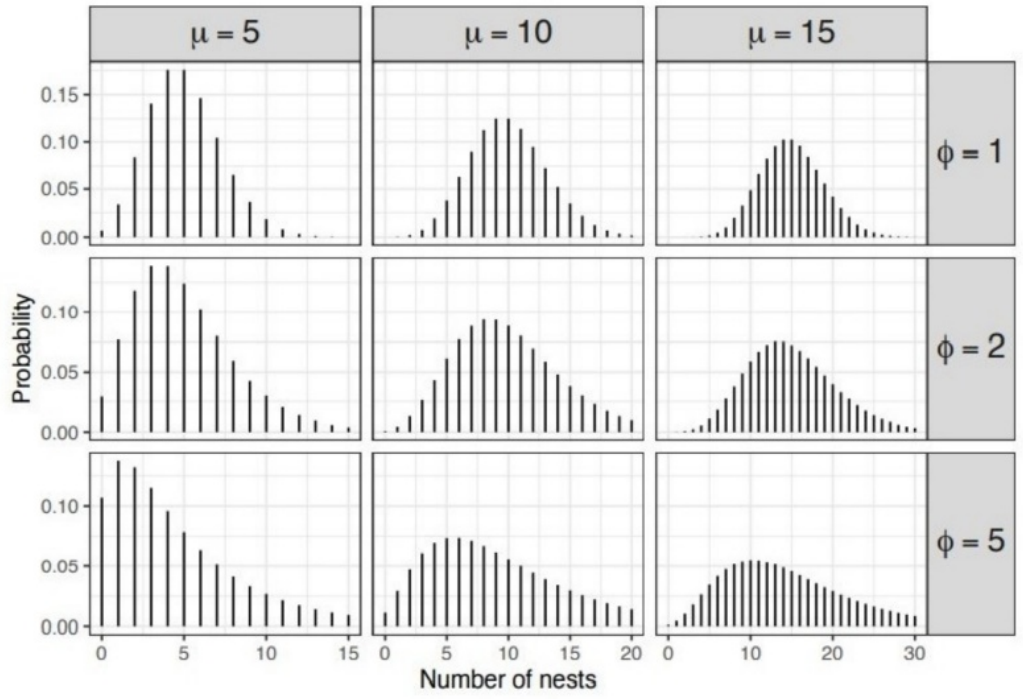


FIGURE 3.4: GP distribution functions for various mean values μ and ϕ values.

GP density functions

dgenpois1, VGAM package (Yee, 2021)

This function does not allow for $\phi < 1$ (cannot use it to visualise underdispersion)

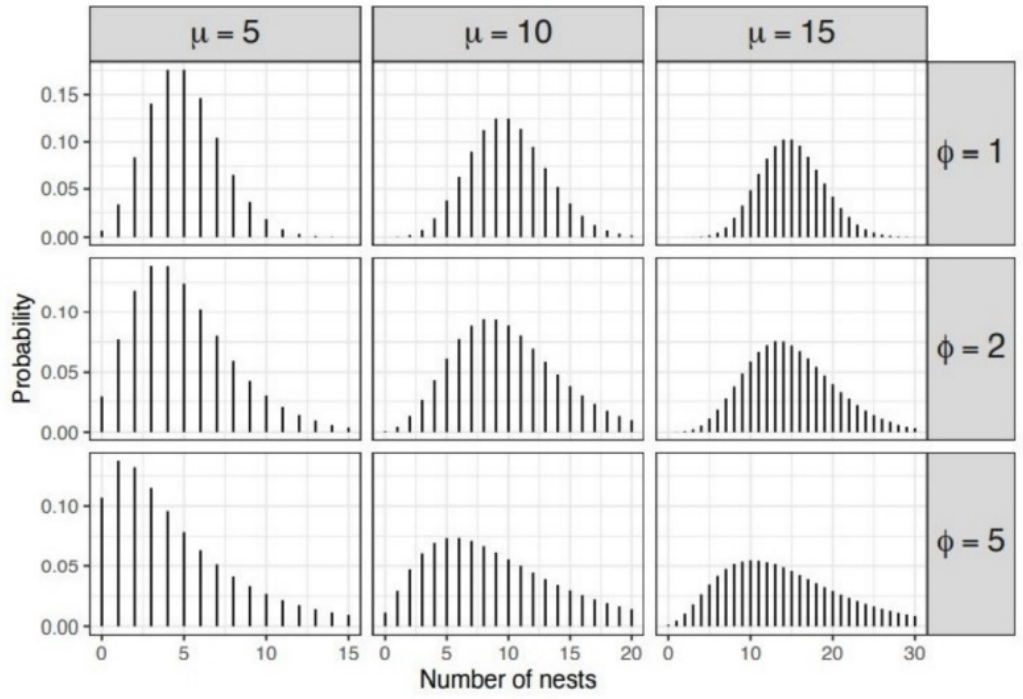
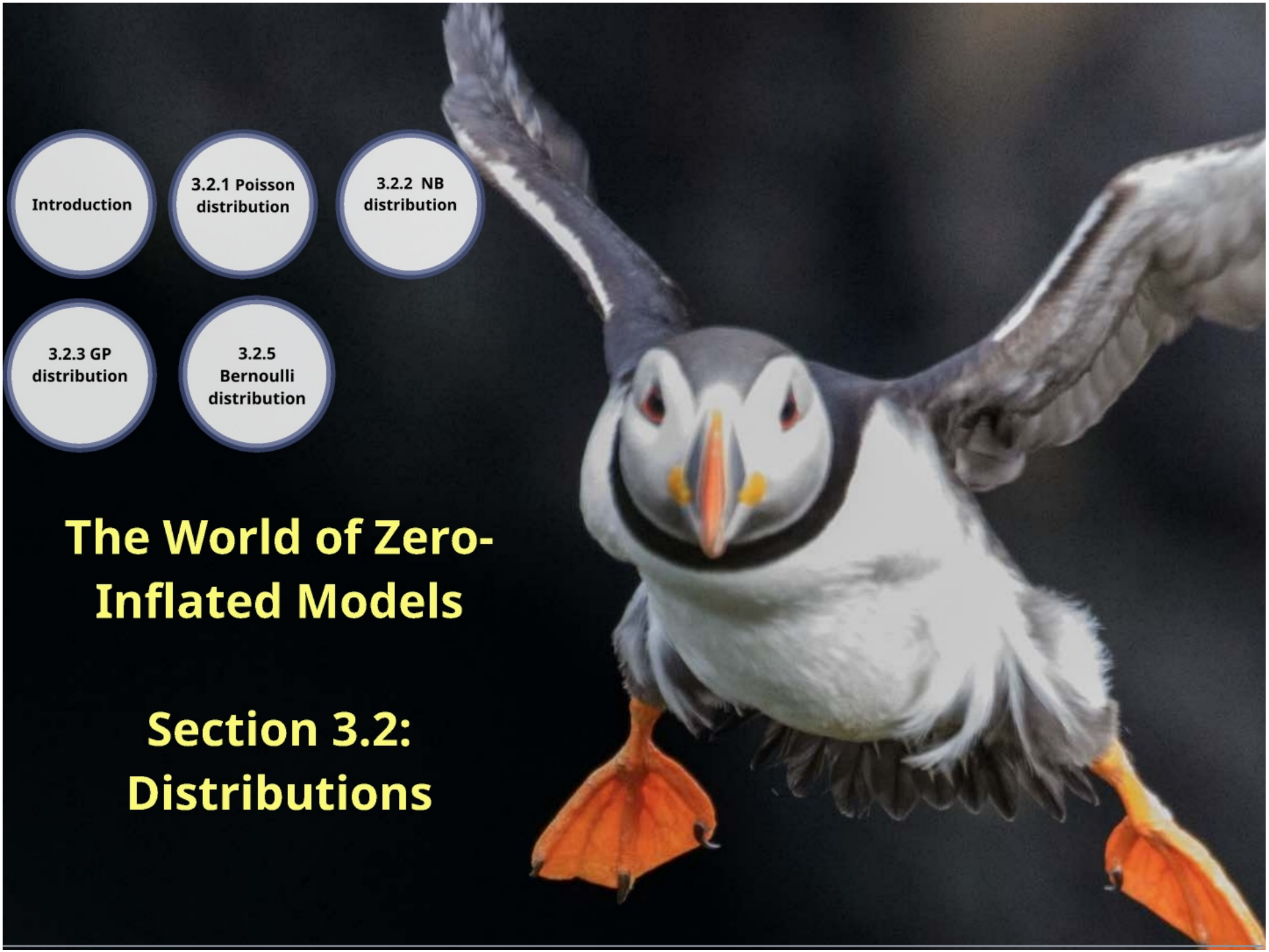


FIGURE 3.4: GP distribution functions for various mean values μ and ϕ values.


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


presence/absence
Data
Bernoulli
dist
function


The Mean
and variance
Slide 1

Suppose that we are not interested in the number of nests in a plot

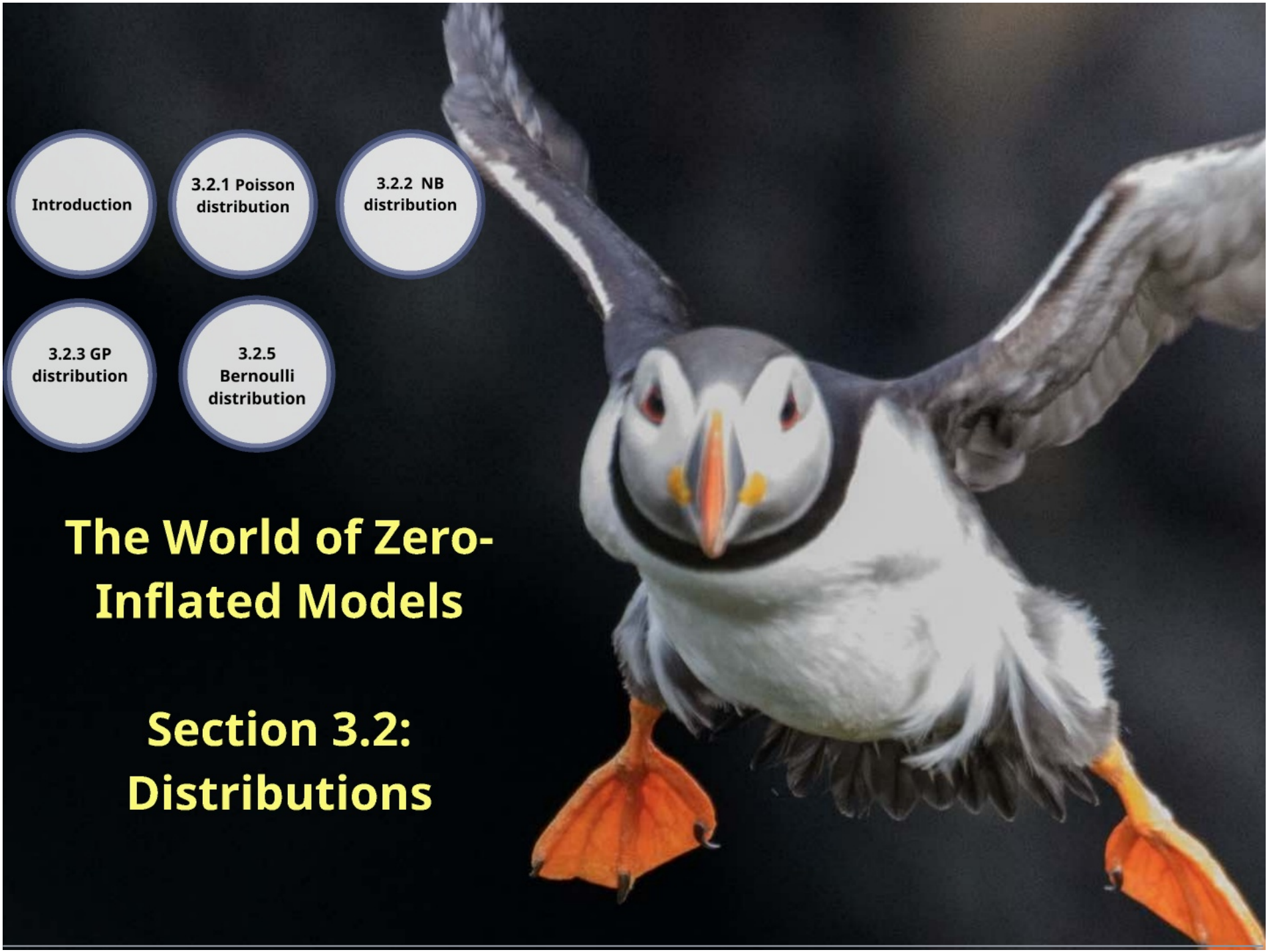
We only want to know whether there are any nests in a plot, yes or no.

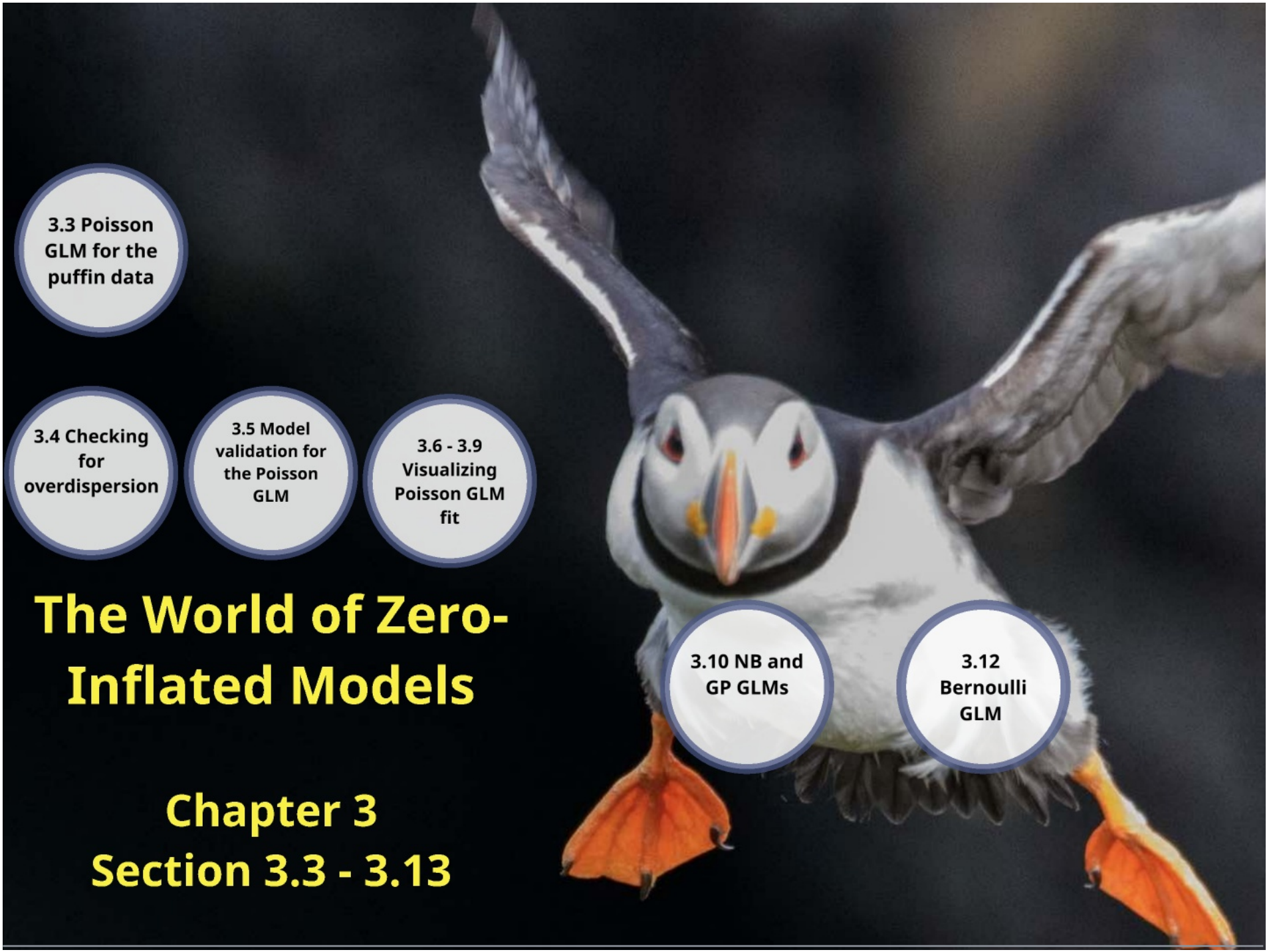
Then a Bernoulli distribution function can be used (coded as 1/0)
It is given by:

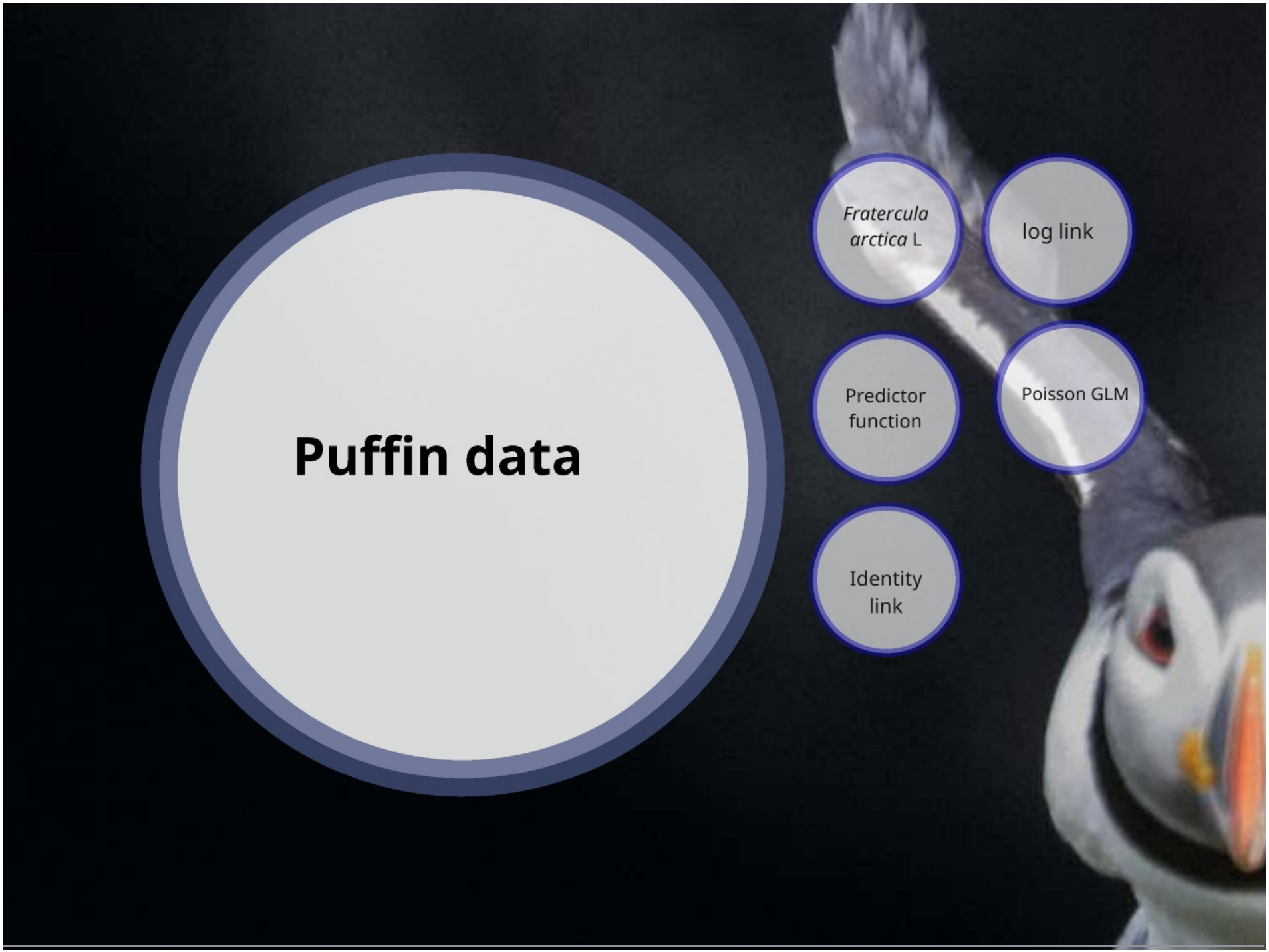
$$P(Y = y|\pi) = y^{\pi} \times (1 - y)^{1-\pi}$$

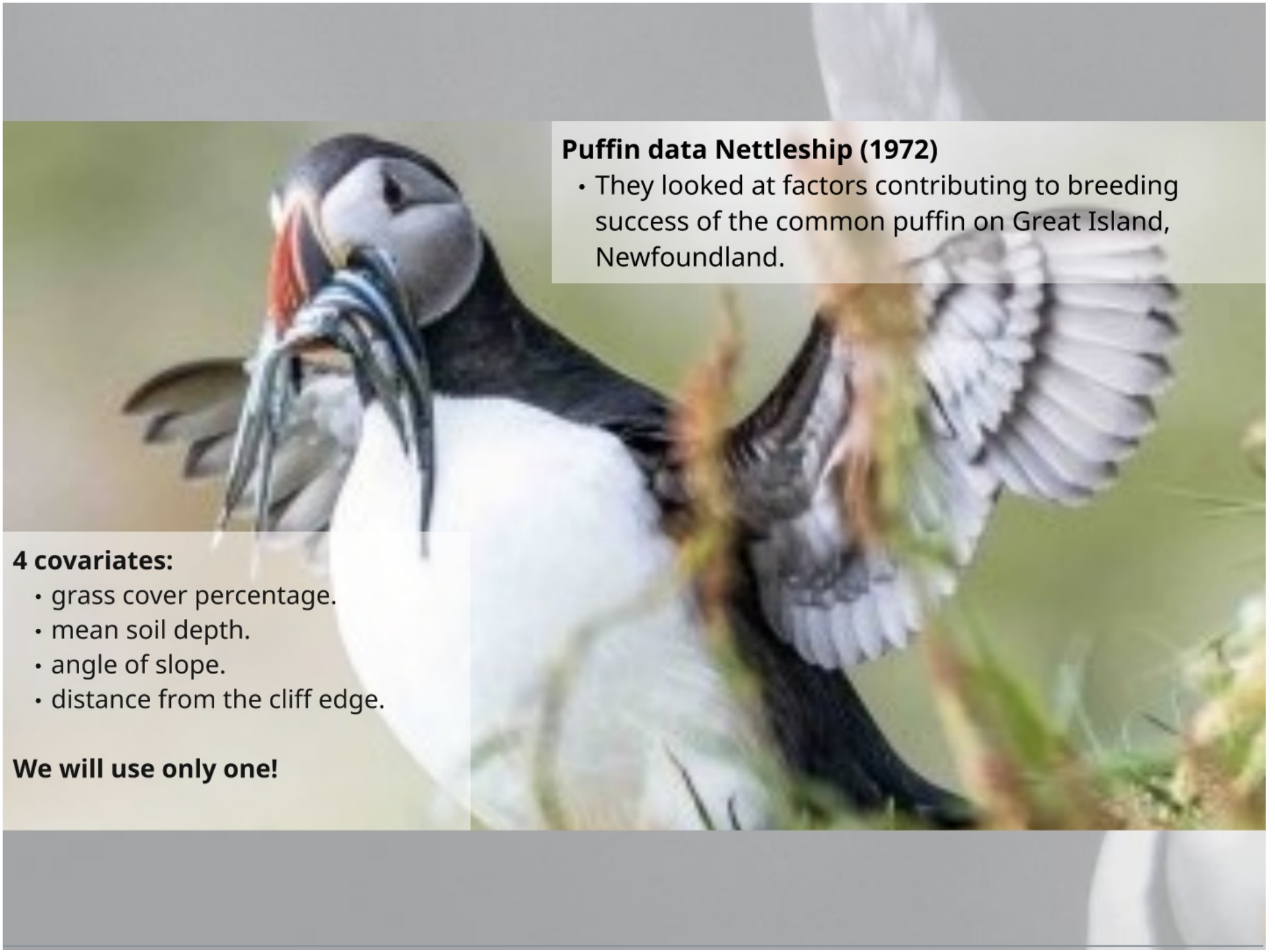
The mean and variance of a variable that follows a Bernoulli distribution are as follows:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \pi_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \pi_i \times (1 - \pi_i) \end{aligned} \tag{3.5}$$


Introduction
3.2.1 Poisson distribution
3.2.2 NB distribution
3.2.3 GP distribution
3.2.5 Bernoulli distribution
The World of Zero-Inflated Models
Section 3.2: Distributions


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
3.10 NB and GP GLMs
3.12 Bernoulli GLM
Chapter 3
Section 3.3 - 3.13


Puffin data
Fratercula arctica L
log link
Predictor function
Poisson GLM
Identity link



Puffin data Nettleship (1972)

- They looked at factors contributing to breeding success of the common puffin on Great Island, Newfoundland.

4 covariates:

- grass cover percentage.
- mean soil depth.
- angle of slope.
- distance from the cliff edge.

We will use only one!

Predictor function:

- Specifies the covariate part of the model.

$$\eta_i = \text{Intercept} + \beta \times \text{Distance}_i$$

Link function:

- Specifies the relationship between the mean of the distribution, μ_i, and the prediction function η_i.

Identity link results in:

$$\mu_i = \eta_i = \text{Intercept} + \beta \times \text{Distance}_i$$

This is only a valid option when the fitted values are not negative.

In a Poisson GLM:

- log link function.
- Ensures that the expected values are strictly positive (Hilbe, 2014).

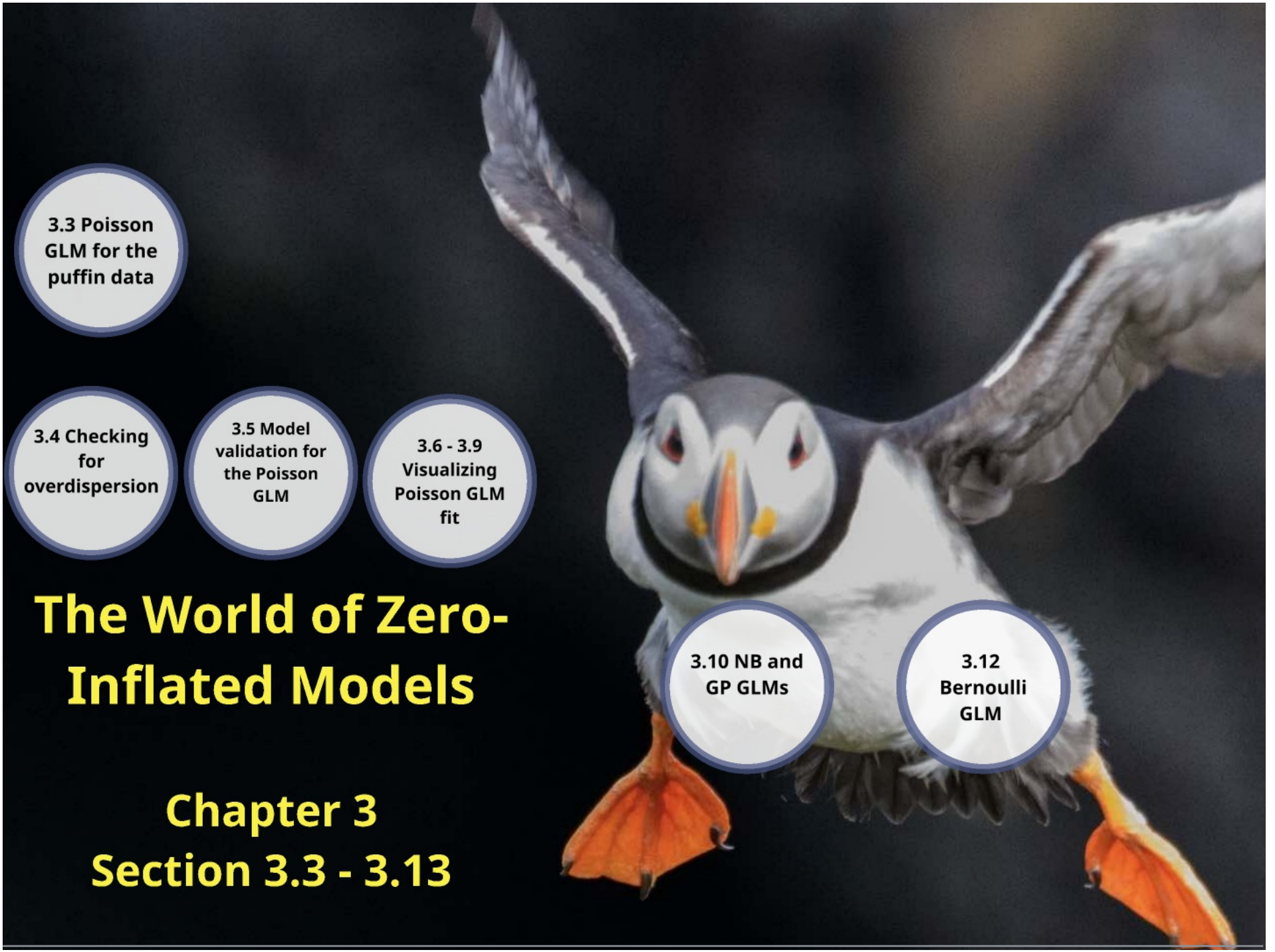
$$\mu_i = e^{\text{Intercept}+\beta\times\text{Distance}_i}$$

Poisson GLM for the puffin data:

$$\begin{aligned} \text{Nests}_i &\sim \text{Poisson}\left(\mu_i\right) \\ \text{E}\left[\text{Nests}_i\right] &= \text{var}\left[\text{Nests}_i\right] = \mu_i \\ \log\left(\mu_i\right) &= \text{Intercept} + \beta \times \text{Distance}_i \end{aligned} \tag{3.6}$$

To estimate the parameters of the model we use the glm function.

```
M1 < - glm(Nesting ~ 1 + Distance,
           family = poisson,
           data = PF)
```


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
Chapter 3
Section 3.3 - 3.13
3.10 NB and GP GLMs
3.12 Bernoulli GLM


Overdispersion
Overdispersion values
Conclusion
Dispersion statistics
Summary
Why simulation

3.4 Checking for overdispersion

We need to verify wether the 'mean = variance' relationship holds.

Calculate the dispersion statistic:

- If 1: Then it is ok.
- If > 1: Overdispersion.
- If < 1: Underdispersion

Dispersion statistic formula:

$$E_i = \frac{\text{Nesting}_i - \text{E}\left[\text{Nesting}_i\right]}{\sqrt{\text{var}\left[\text{Nesting}_i\right]}} = \frac{\text{Nesting}_i - \mu_i}{\sqrt{\mu_i}}$$

```
E1   <- resid(M1, type = "pearson") #Pearson residuals
N    <- nrow(PF)                    #Sample size
k    <- length(coef(M1))            #Number of parameters
Disp <- sum(E1^2) / (N - k)         #Dispersion statistic
Disp
```

```
## [1] 1.975952
```

How much larger than 1 is acceptable?

Answer:

- We don’t know.

What you can do:

- Simulate a large number of truly Poisson distributed data sets from the model.
- Each time calculate the dispersion statistic.
- Compare the dispersion statistics for these 1000 simulated data sets with the dispersion statistic of the model applied on the observed data.

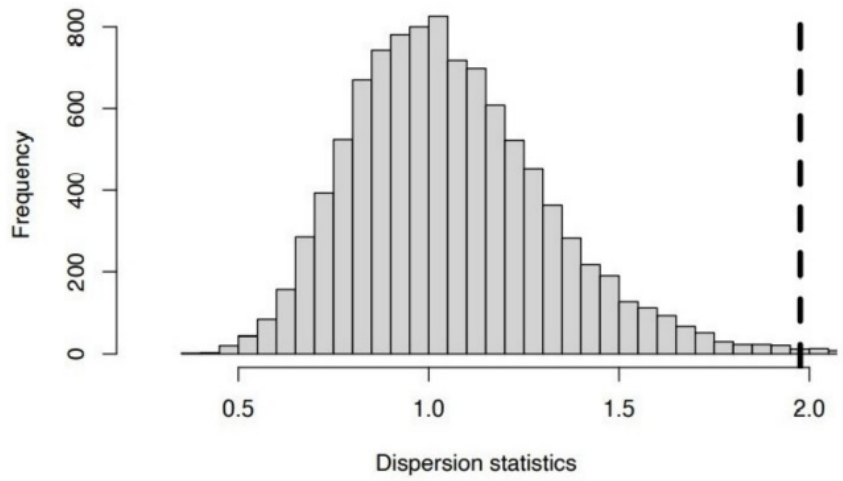
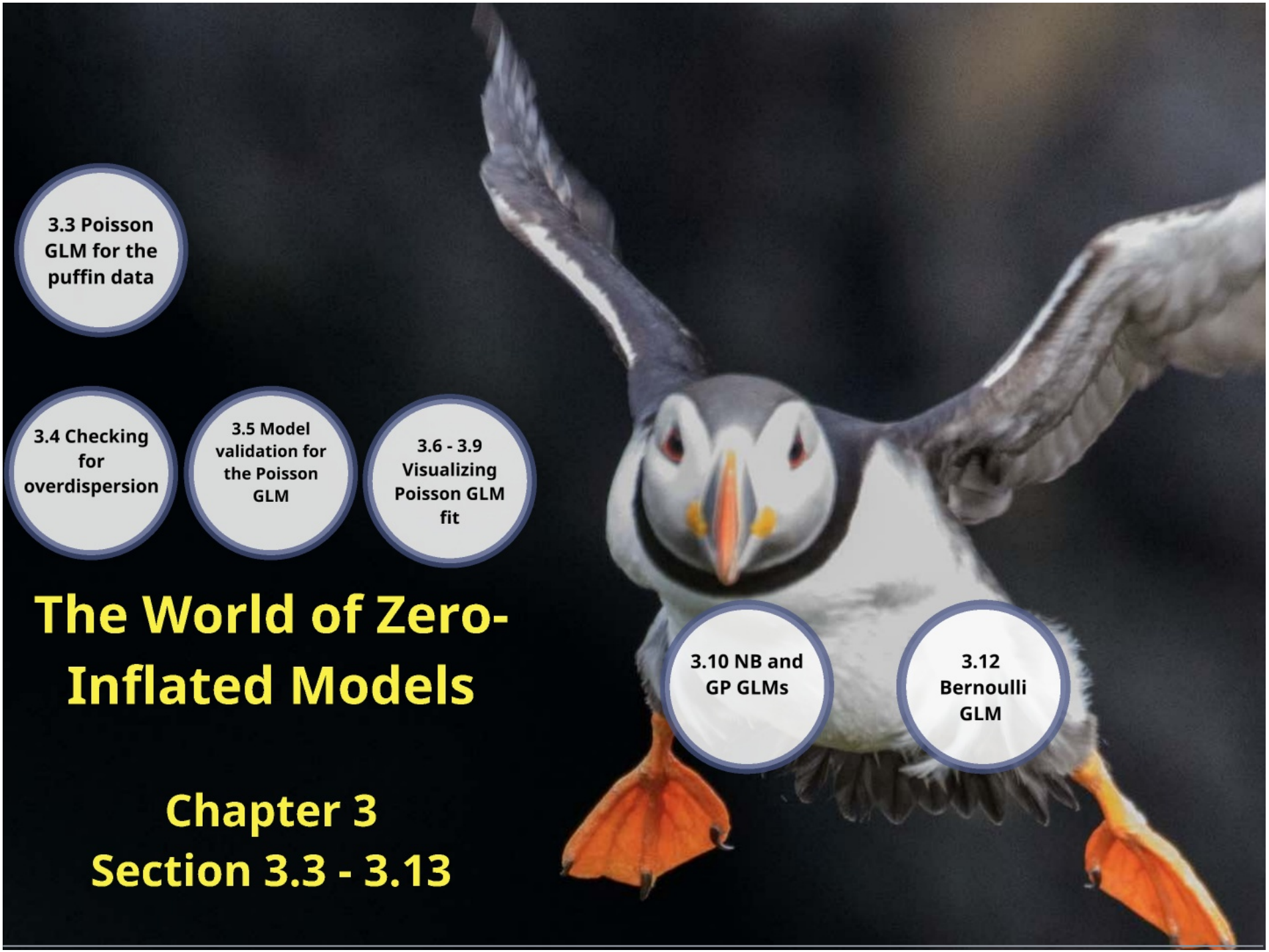
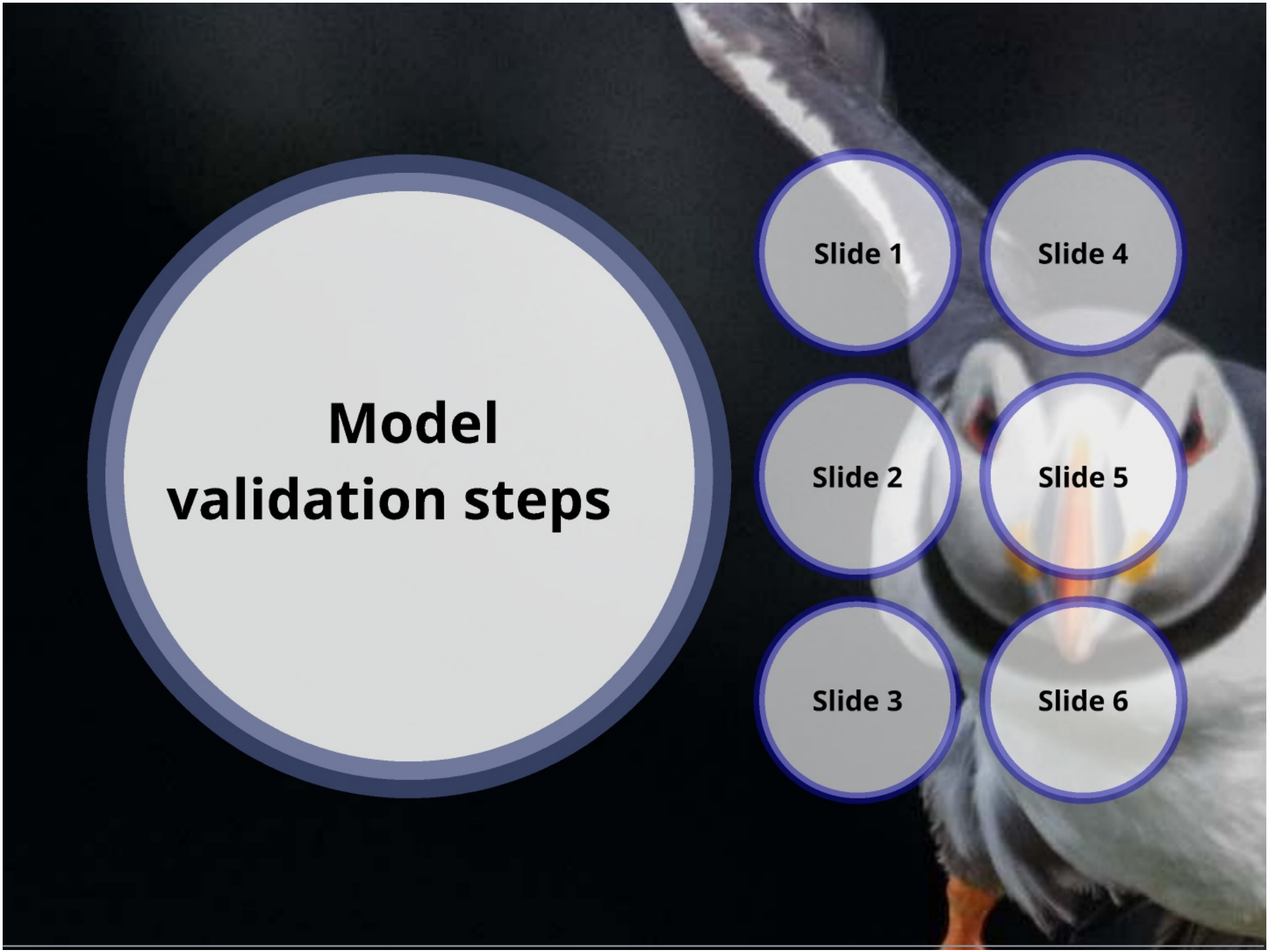


FIGURE 3.6: Results of the simulation study for the Poisson GLM to investigate the range of dispersion statistics. The dotted line is the dispersion statistic obtained by applying the Poisson GLM on the data.

The Poisson GLM applied on the puffin data shows that there is overdispersion. If there is overdispersion then we need to figure out the cause of the overdispersion and adjust the model accordingly. If we select the wrong follow-up model, then the estimated parameters may be biased.


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
Chapter 3
Section 3.3 - 3.13
3.10 NB and GP GLMs
3.12 Bernoulli GLM


Model validation steps
Slide 1
Slide 4
Slide 2
Slide 5
Slide 3
Slide 6

Model validation

Even if the dispersion statistic of a Poisson GLM indicates that there is no overdispersion, we should still apply a detailed model validation.

Model validation steps

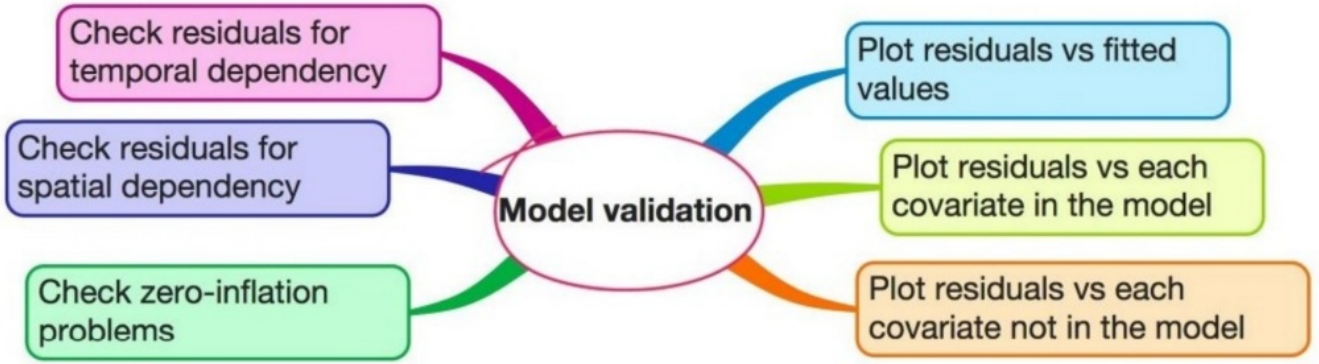


FIGURE 3.7: Model validation steps for a regression-type analysis.

Working with residuals

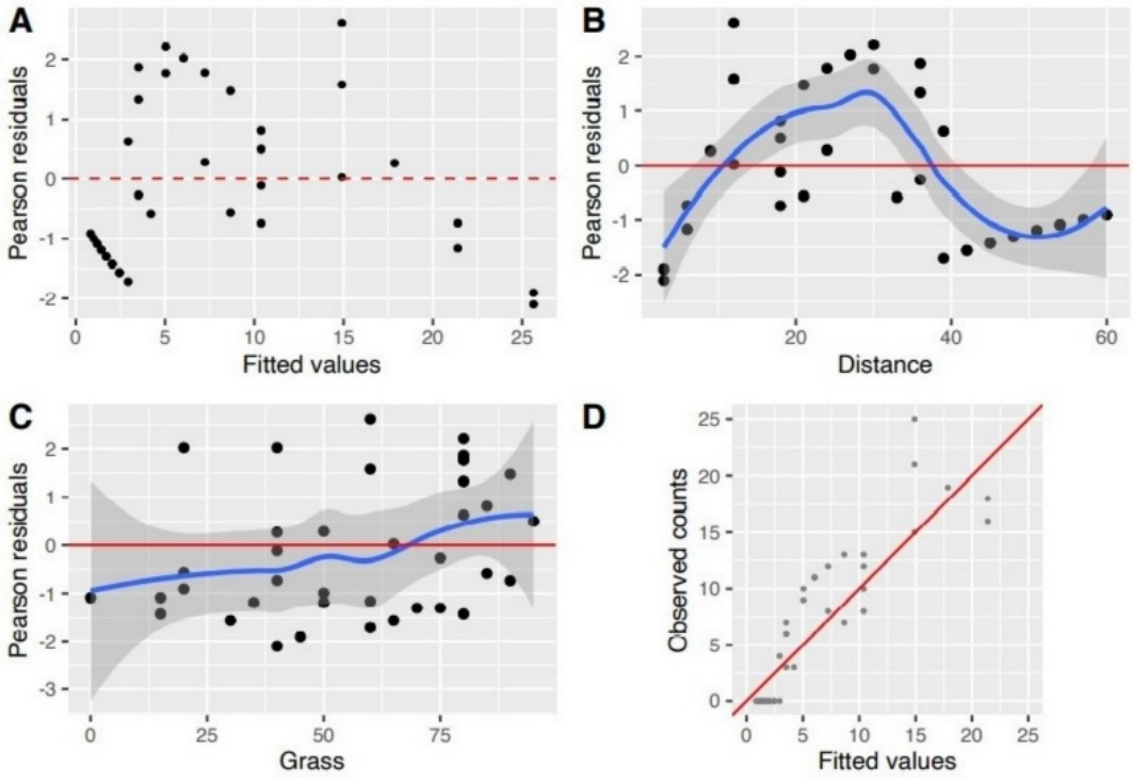


FIGURE 3.8: A: Pearson residuals plotted versus fitted values obtained by the Poisson GLM applied on the puffin data. B: Pearson residuals versus the covariate Distance. A scatterplot smoother is added to aid visual interpretation. C: Pearson residuals versus the covariate Grass. D: Observed nesting data plotted versus the fitted values of the Poisson GLM.

Zero-inflation problems:

- Calculate the numbers of zeros in each simulated data set.
- Compare these with the number of zeros in the observed nesting data.

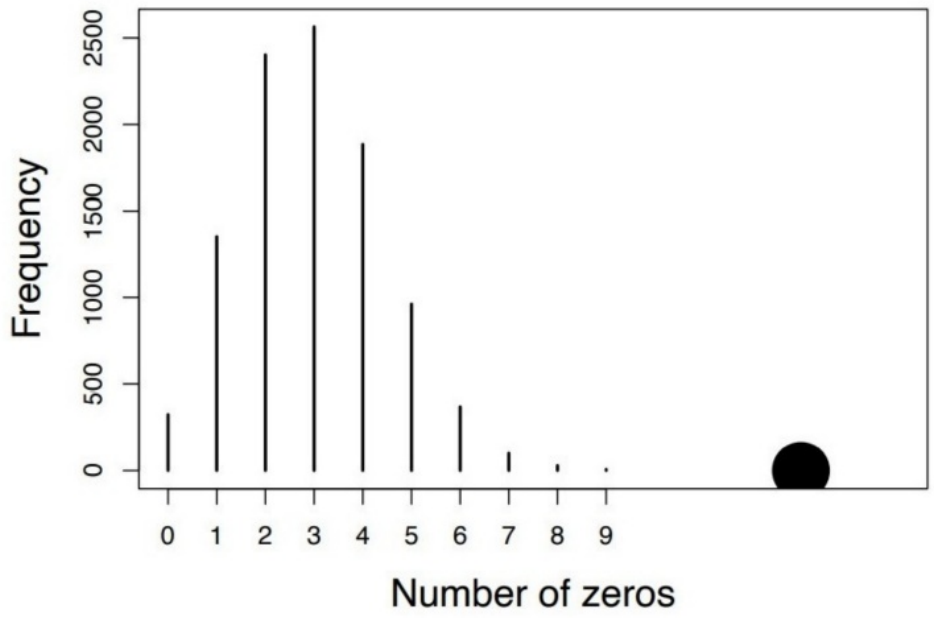
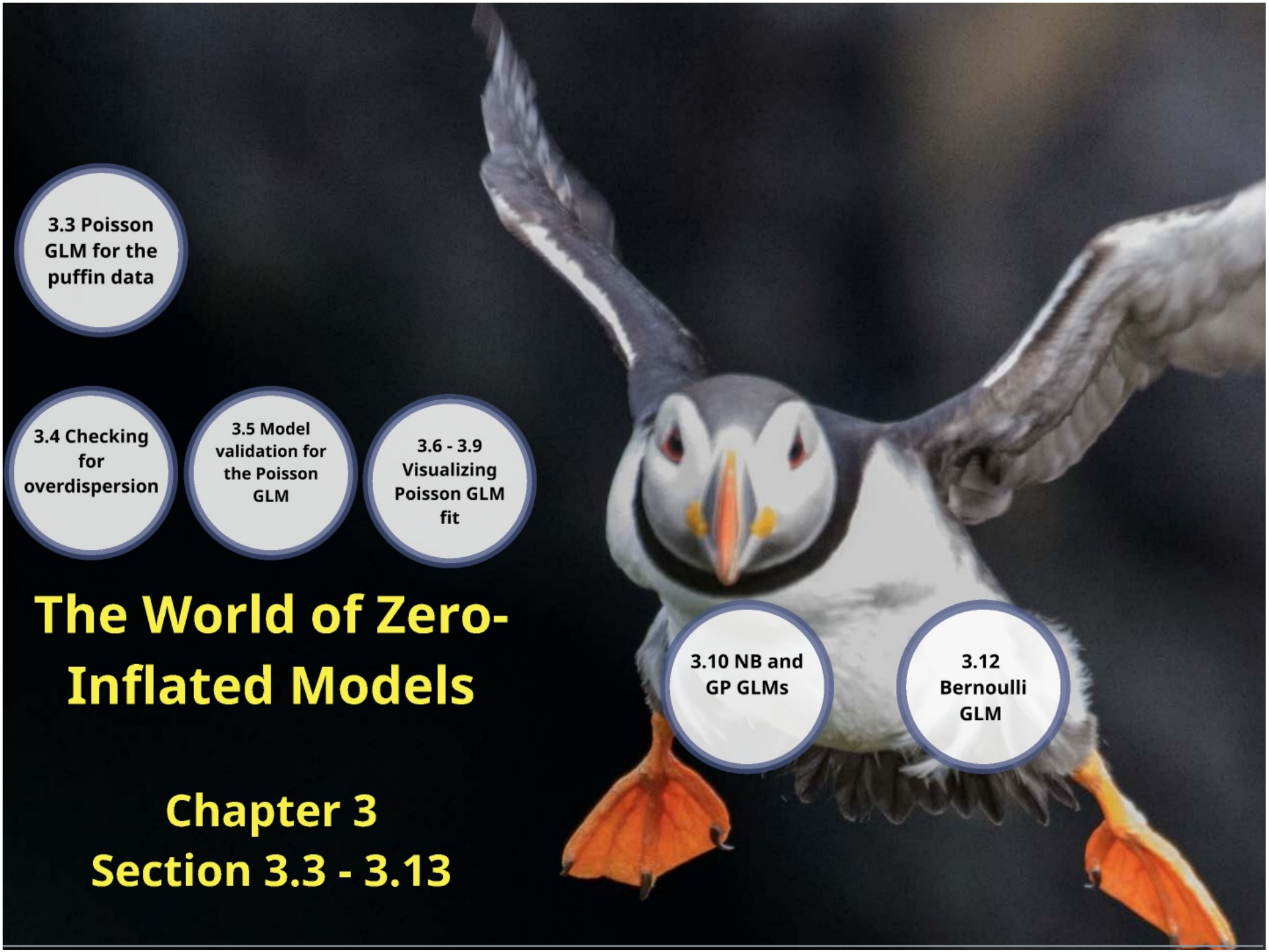


FIGURE 3.9: Frequency plot showing in how many of the 10000 simulated data sets we have a data set with 0 zeros, 1 zero, 2 zeros, etc. The dot is the number of zeros in the observed data

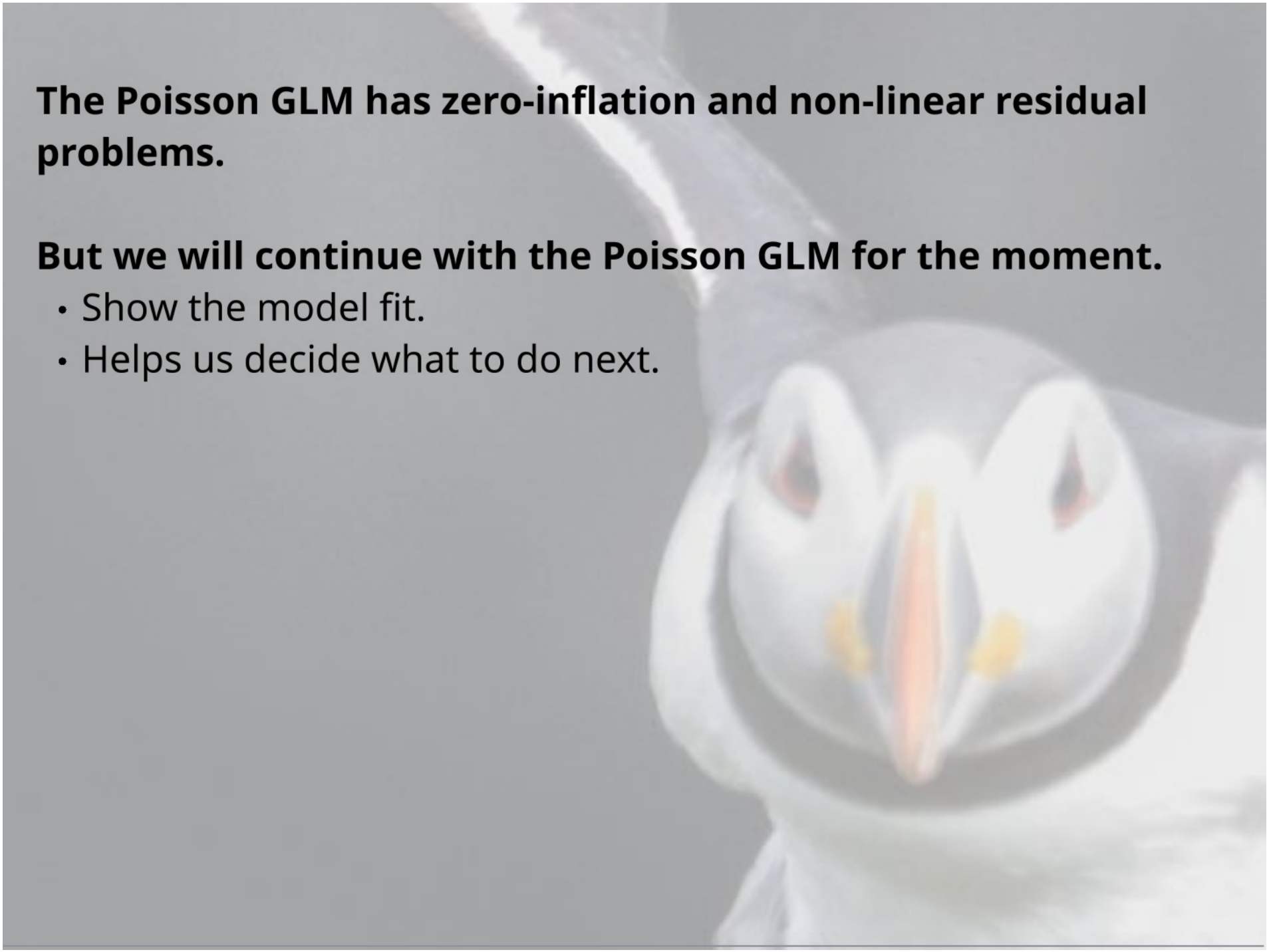
Conclusions:

The Poisson GLM applied on the puffin data is overdispersed. Model validation indicates two main problems. The model cannot cope with the excessive number of zeros in the nesting data, and there is a non-linear distance to the edge effect.

and now what?


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
Chapter 3
Section 3.3 - 3.13
3.10 NB and GP GLMs
3.12 Bernoulli GLM


Poisson GLM fit
Slide 1
Slide 4
Slide 2
Slide 5
Slide 3



The Poisson GLM has zero-inflation and non-linear residual problems.

But we will continue with the Poisson GLM for the moment.

- Show the model fit.
- Helps us decide what to do next.

Model fit

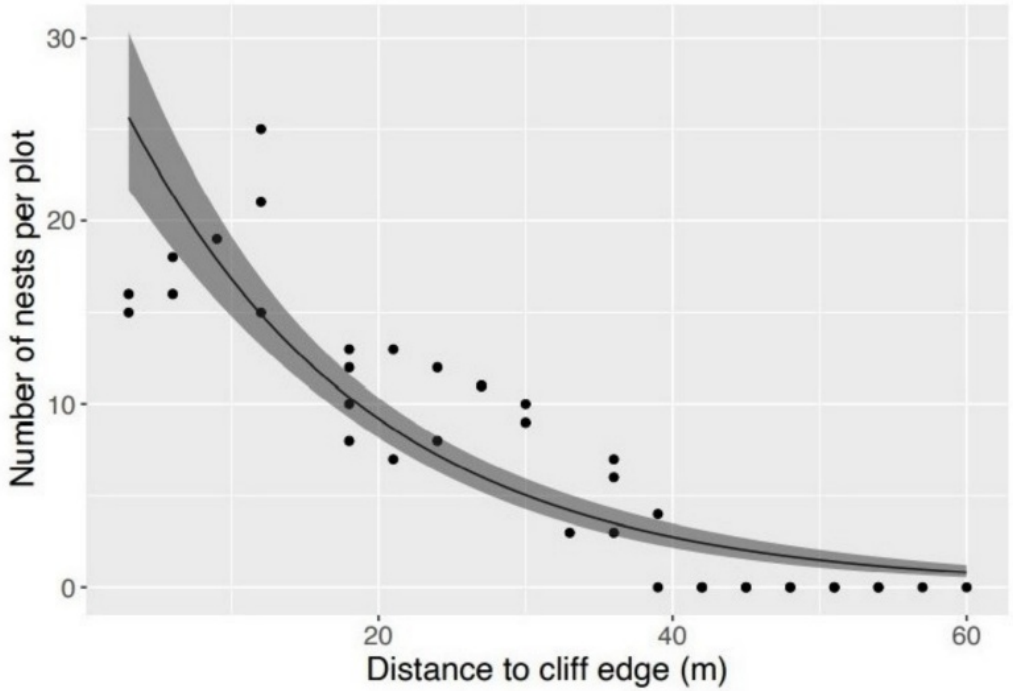


FIGURE 3.10: Model fit of the Poisson GLM applied on the puffin data. The grey polygon around the fitted values represents a 95% confidence interval for the mean.

Fitted values

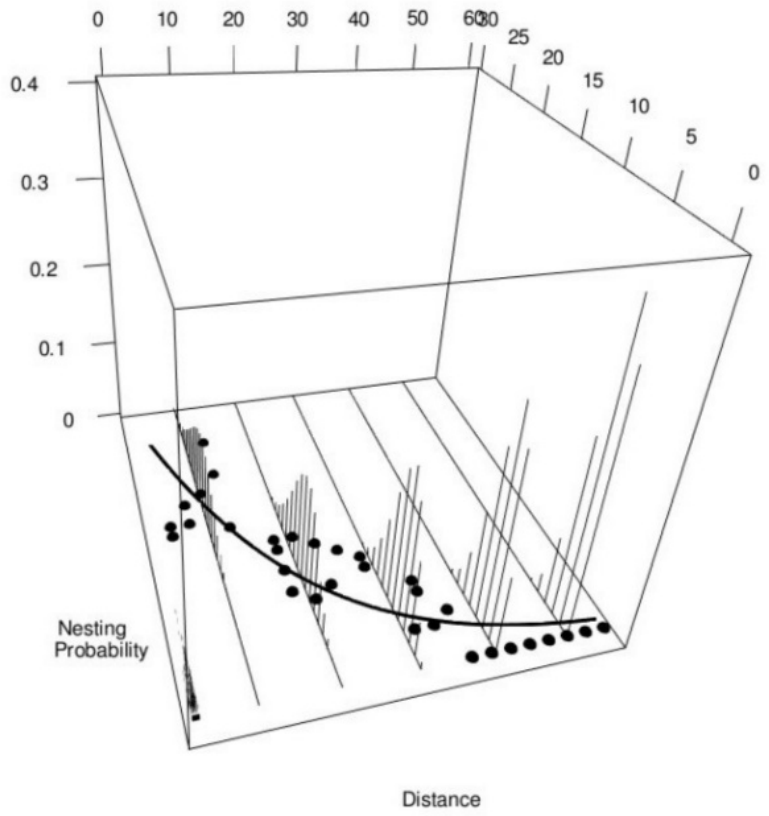


FIGURE 3.11: 3-D presention of the fitted Poisson GLM applied on the puffin data. The x-y space shows nestings versus distance. The solid black line is the model fit of the Poisson GLM. At 5 equi-distant values along the distance axis we show Poisson density curves (z-axis). These density curves have a mean value that is given by the fitted value.

What is the Poisson distribution doing?

- For 10 specific Distance values we have simulated 50 nesting values via rpois.

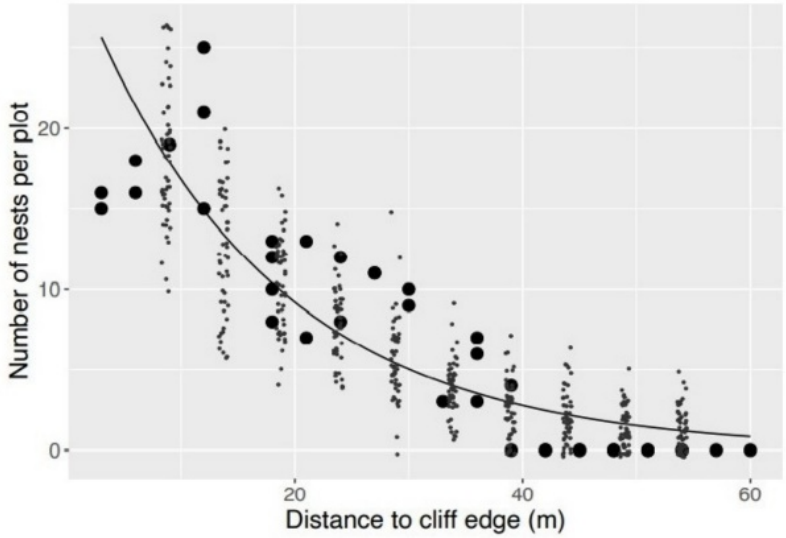
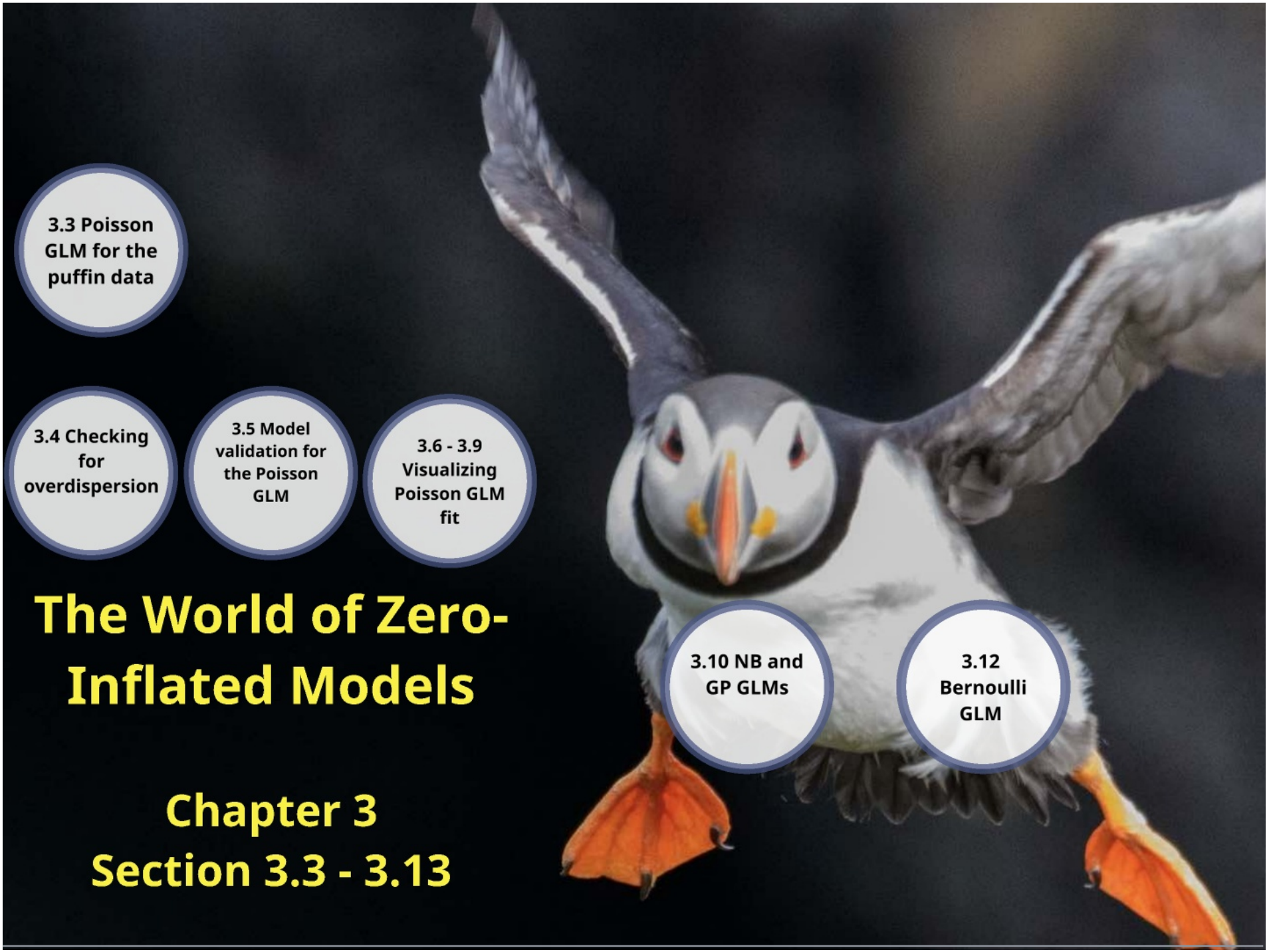
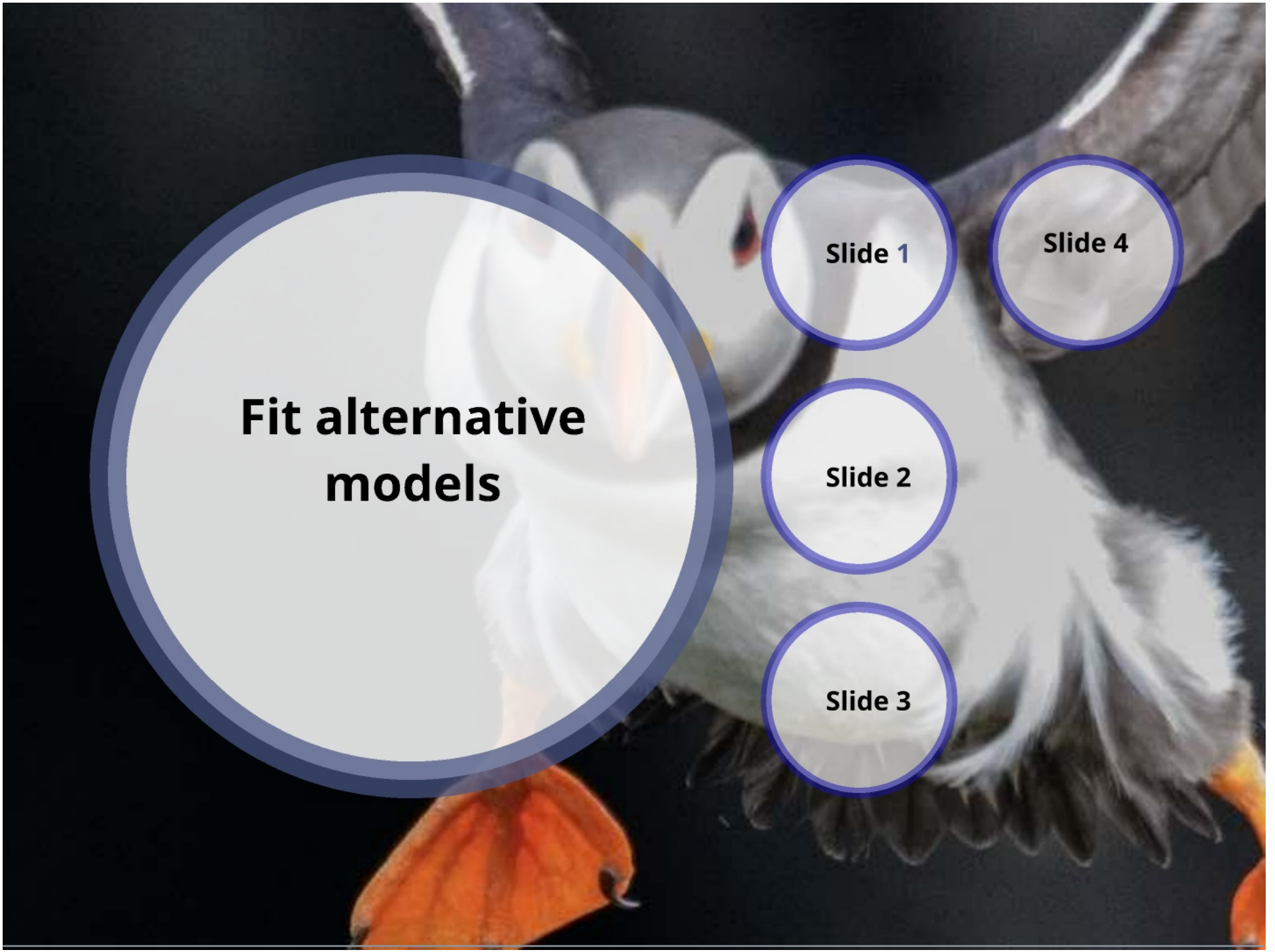
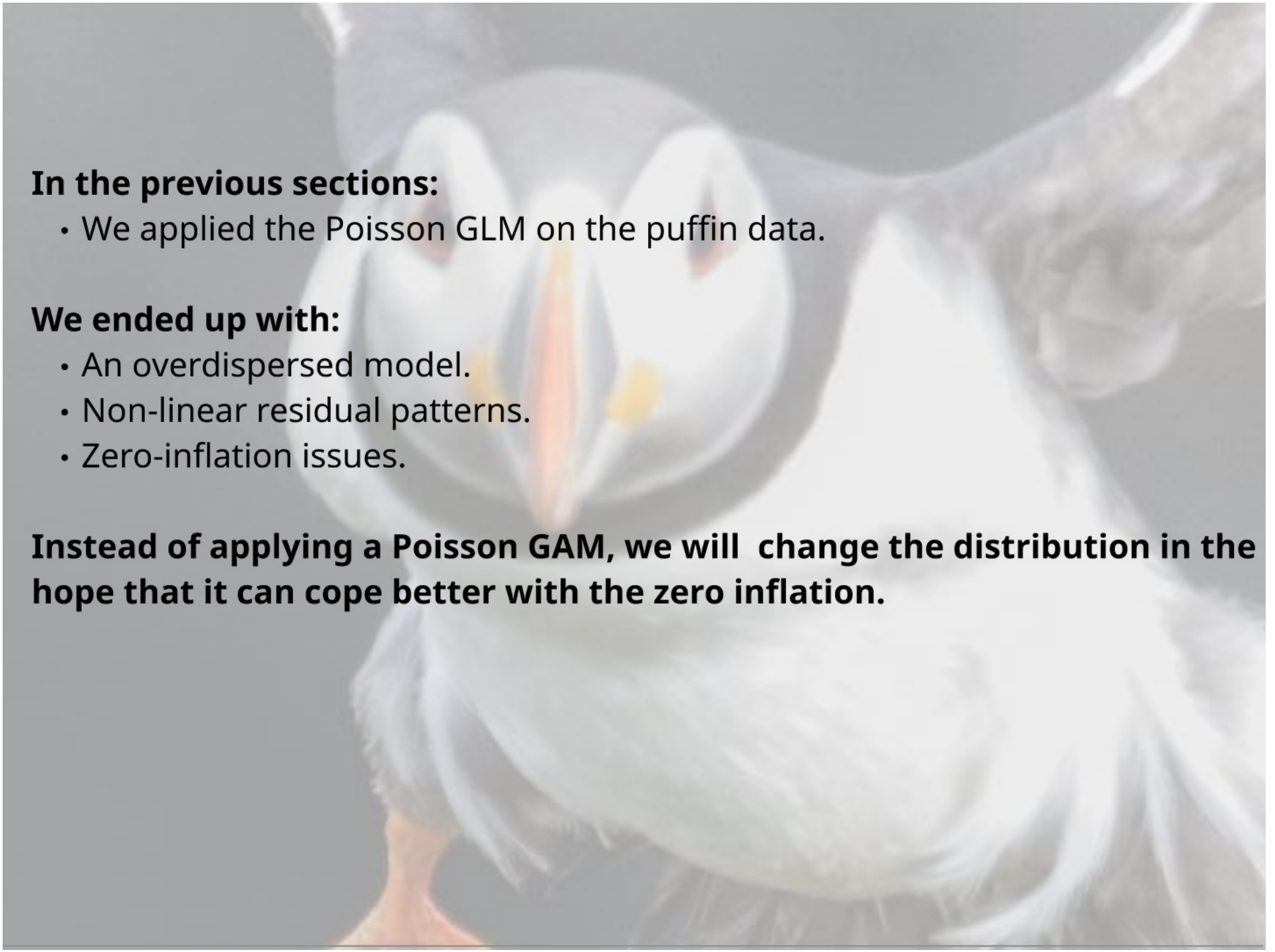


FIGURE 3.12: Fitted values of the Poisson GLM (solid line), observed data (larger black dots) and 50 simulated nesting values (smaller dots) at 10 equidistant distance values.

Time for a change!!


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
3.10 NB and GP GLMs
3.12 Bernoulli GLM
Chapter 3
Section 3.3 - 3.13


Fit alternative models
Slide 1
Slide 4
Slide 2
Slide 3



In the previous sections:

- We applied the Poisson GLM on the puffin data.

We ended up with:

- An overdispersed model.
- Non-linear residual patterns.
- Zero-inflation issues.

Instead of applying a Poisson GAM, we will change the distribution in the hope that it can cope better with the zero inflation.

Alternative distributions:

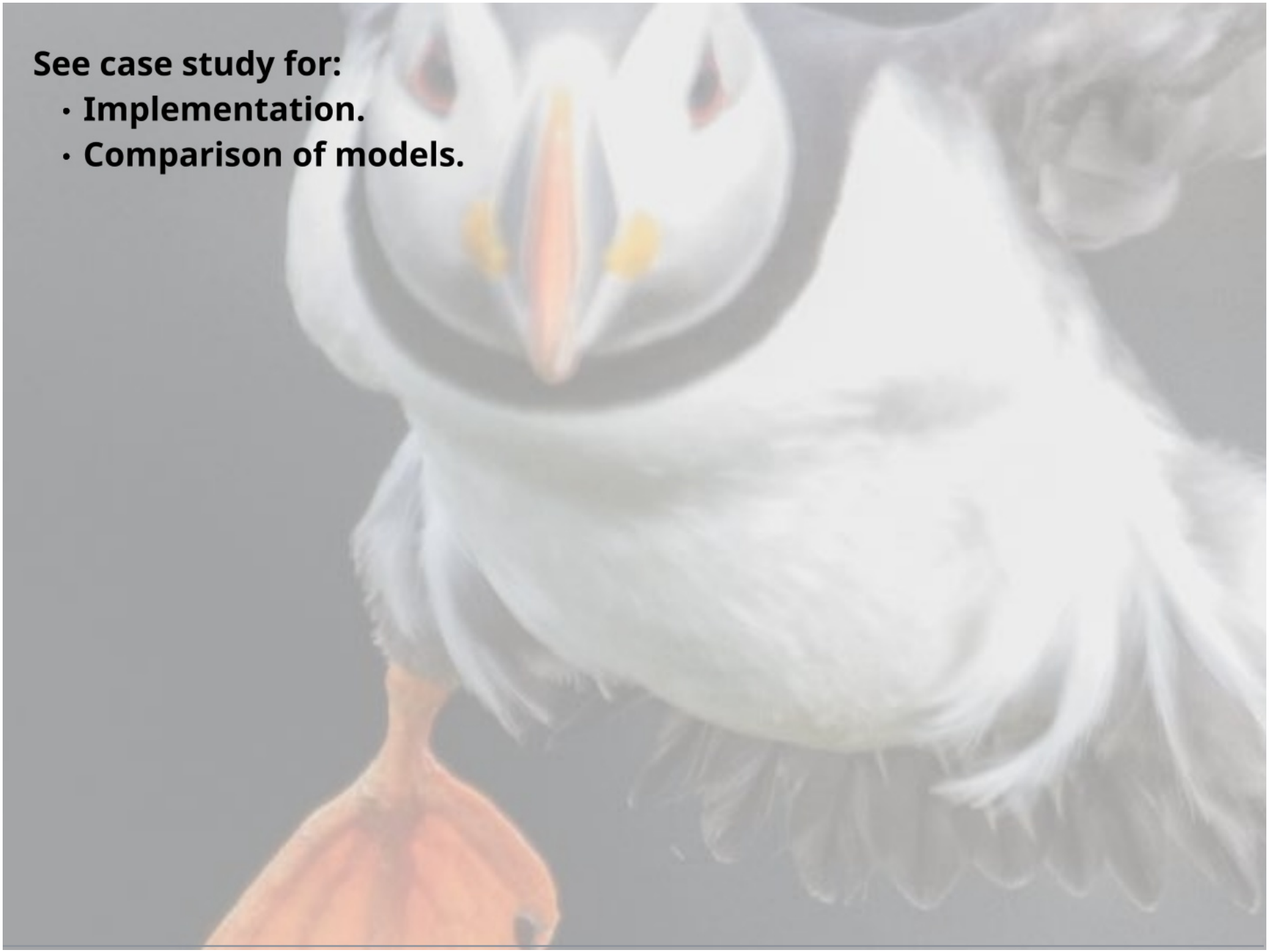
- NB GLM:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i + \frac{\mu_i^2}{\theta} \end{aligned} \tag{3.3}$$

- GP GLM:

$$\begin{aligned} \mathrm{E}\left[\mathrm{Nest}_i\right] &= \mu_i \\ \mathrm{var}\left[\mathrm{Nest}_i\right] &= \mu_i \times \frac{1}{(1-\lambda)^2} \\ &= \mu_i \times \phi \end{aligned} \tag{3.4}$$

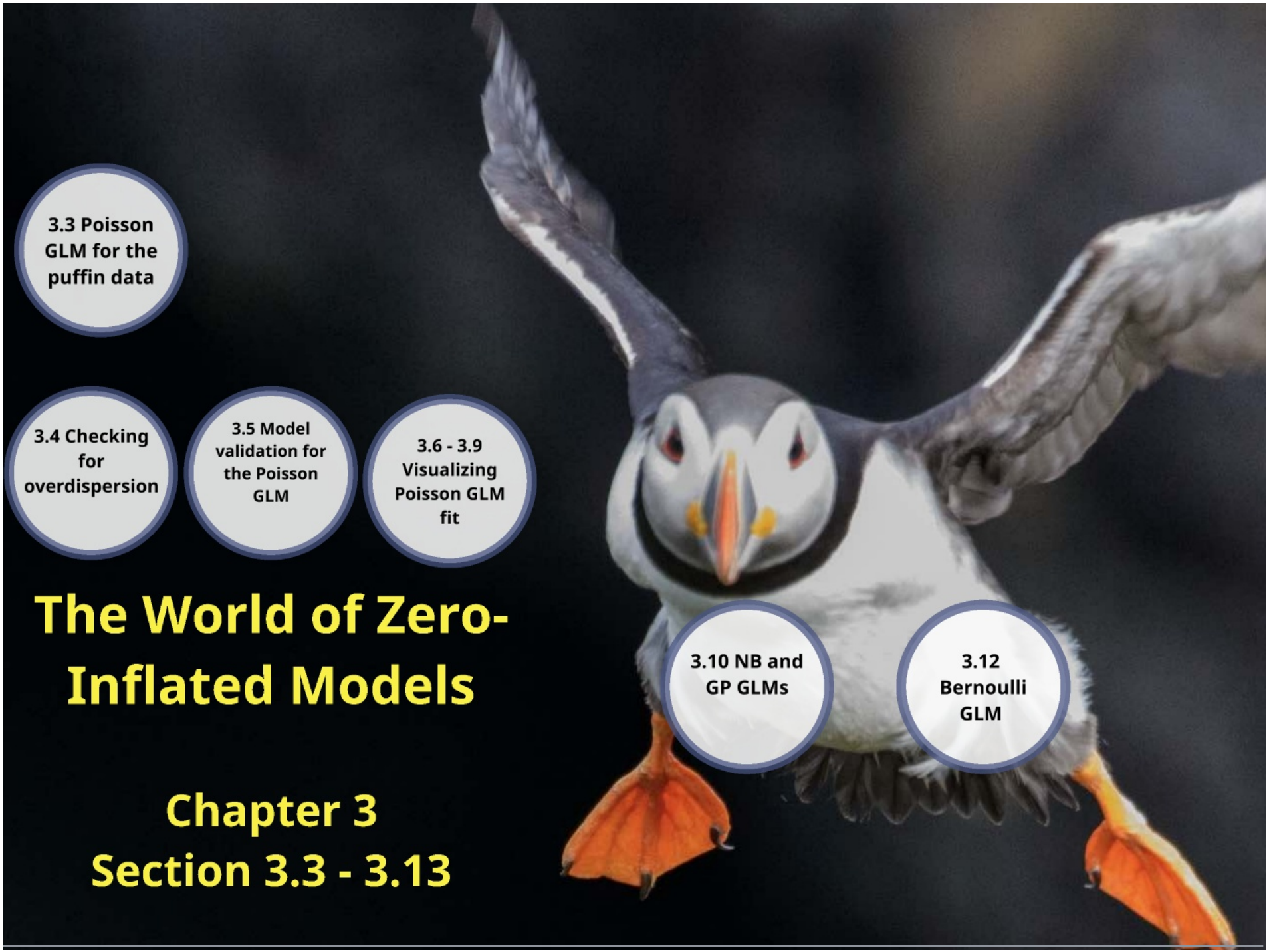
$$\mu_i = e^{\mathrm{Intercept} + \beta \times \mathrm{Distance}_i}$$



See case study for:

- **Implementation.**
- **Comparison of models.**

You now have a choice!!


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
Chapter 3
Section 3.3 - 3.13
3.10 NB and GP GLMs
3.12 Bernoulli GLM


Presence/absence data
Slide 1
Slide 4
Slide 2
Slide 5
Slide 3

Model formulation

Are there puffin nests in a plot yes or no?

```
PF$Nesting01 <- ifelse(test = PF$Nesting > 0,
                       yes = 1,
                       no = 0)
table(PF$Nesting01)
```

A Bernoulli GLM is defined as follows:

$$
\begin{aligned}
\text{Nests01}_i &\sim \text{Bernoulli}\left(\pi_i\right) \\
\text{E}\left[\text{Nests01}_i\right] &= \pi_i \\
\text{var}\left[\text{Nests01}_i\right] &= \pi_i \times (1 - \pi_i) \\
\pi_i &= \frac{\exp(\text{Intercept} + \beta \times \text{Grass}_i)}{1 + \exp(\text{Intercept} + \beta \times \text{Grass}_i)}
\end{aligned}
\tag{3.10}
$$

Fitted probabilities are always between 0 and 1
we use the ratio of the 2 exponential functions
in Equation (3.10)

$$\text{logit}\,(\pi_i) = \text{Intercept} + \beta \times \text{Grass}_i \qquad (3.11)$$

Applying the Bernoulli GLM

```
B1 <- glm(Nesting01 ~ Grass,
          family = binomial,
          data = PF)
print(summary(B1)$coefficients, digits = 3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1613     0.8791   -1.32   0.1865
## Grass         0.0338     0.0155    2.18   0.0292
```

The numerical output indicates that we can write the model as follows.

$$\text{logit}\left(\pi_i\right) = -1.161 + 0.038 \times \text{Grass}_i$$

The effect of grass cover percentage is positive and significant at the 5% level.

Model interpretation

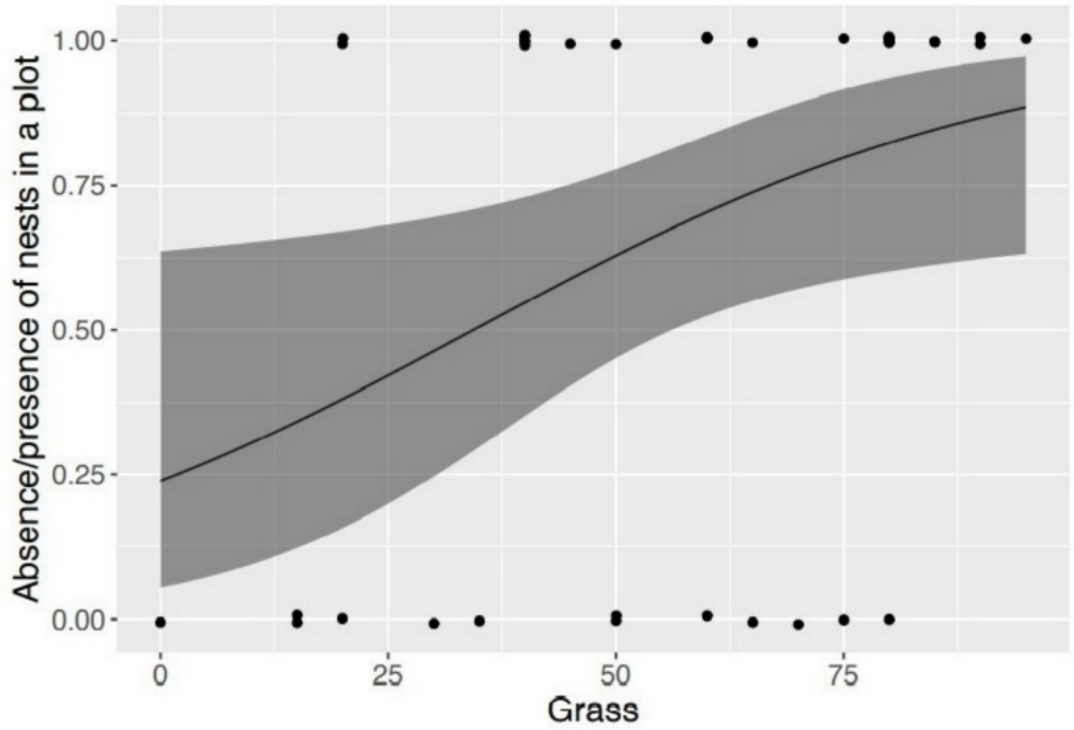
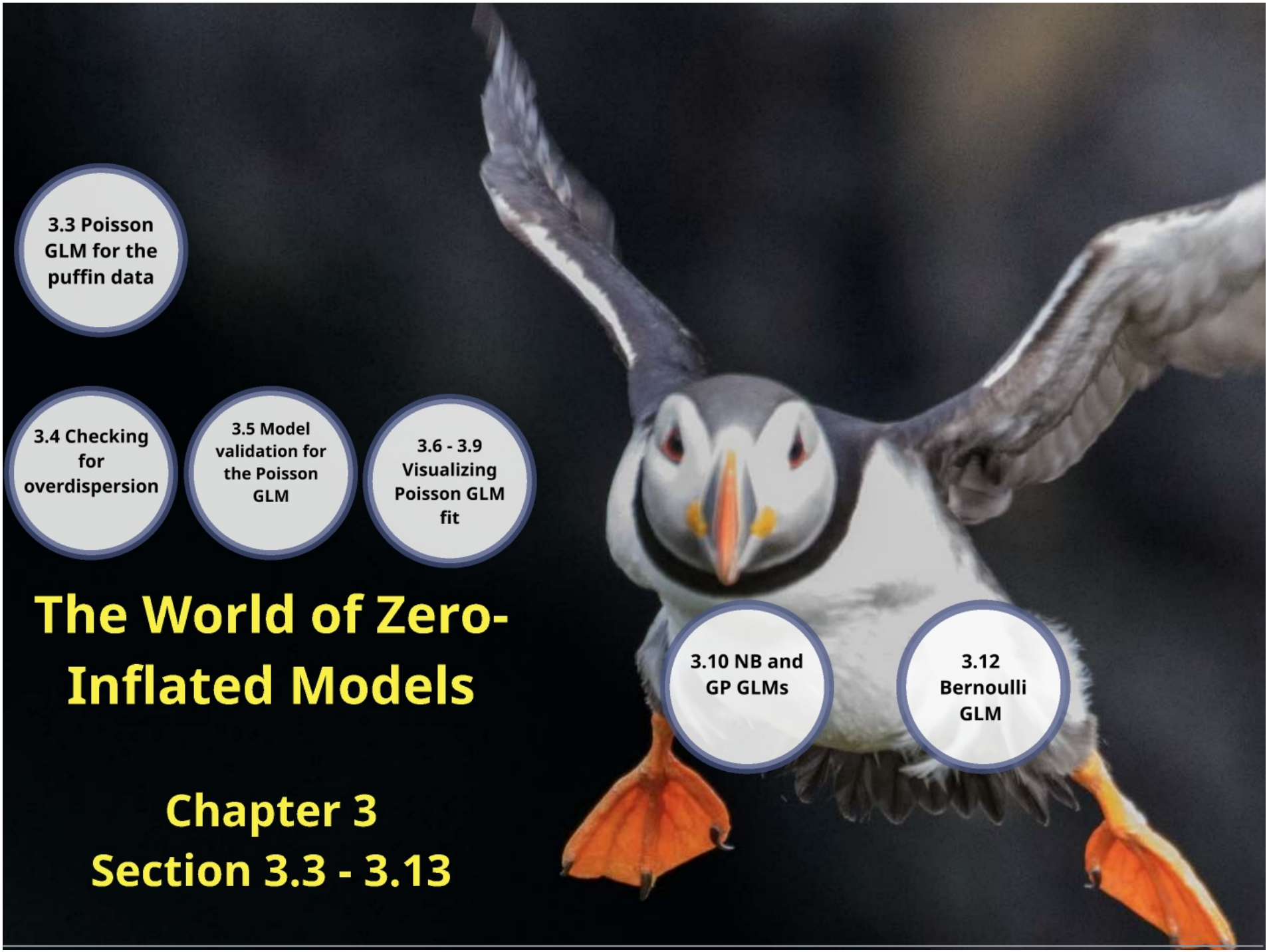


FIGURE 3.28: Model fit of the Bernoulli GLM applied on the absence–presence of puffin nests. The grey polygon around the fitted values represents a 95% confidence interval for the mean. The dots are the observed absence–presence data, and a small amount of random noise was added in order to be able to make a distinction between observations with the same values.


3.3 Poisson GLM for the puffin data
3.4 Checking for overdispersion
3.5 Model validation for the Poisson GLM
3.6 - 3.9 Visualizing Poisson GLM fit
The World of Zero-Inflated Models
Chapter 3
Section 3.3 - 3.13
3.10 NB and GP GLMs
3.12 Bernoulli GLM


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


Zero-inflated models for count data
Slide 1
Slide 2

In this chapter/presentation:

- Analyse count data with an excessive number of zeros.

Various types of models exist for the analysis of such data:

- Zero-inflated Poisson (ZIP) models
- Zero-altered Poisson (ZAP) models.

The latter ones are also called hurdle models

- Not discussed here.

Alternatives for the ZIP model:

- Zero-inflated negative binomial (ZINB) model.
- Zero-inflated generalised poisson (ZIGP) model.

It is important to fully comprehend what these models do.

- Explain them using simulated data.

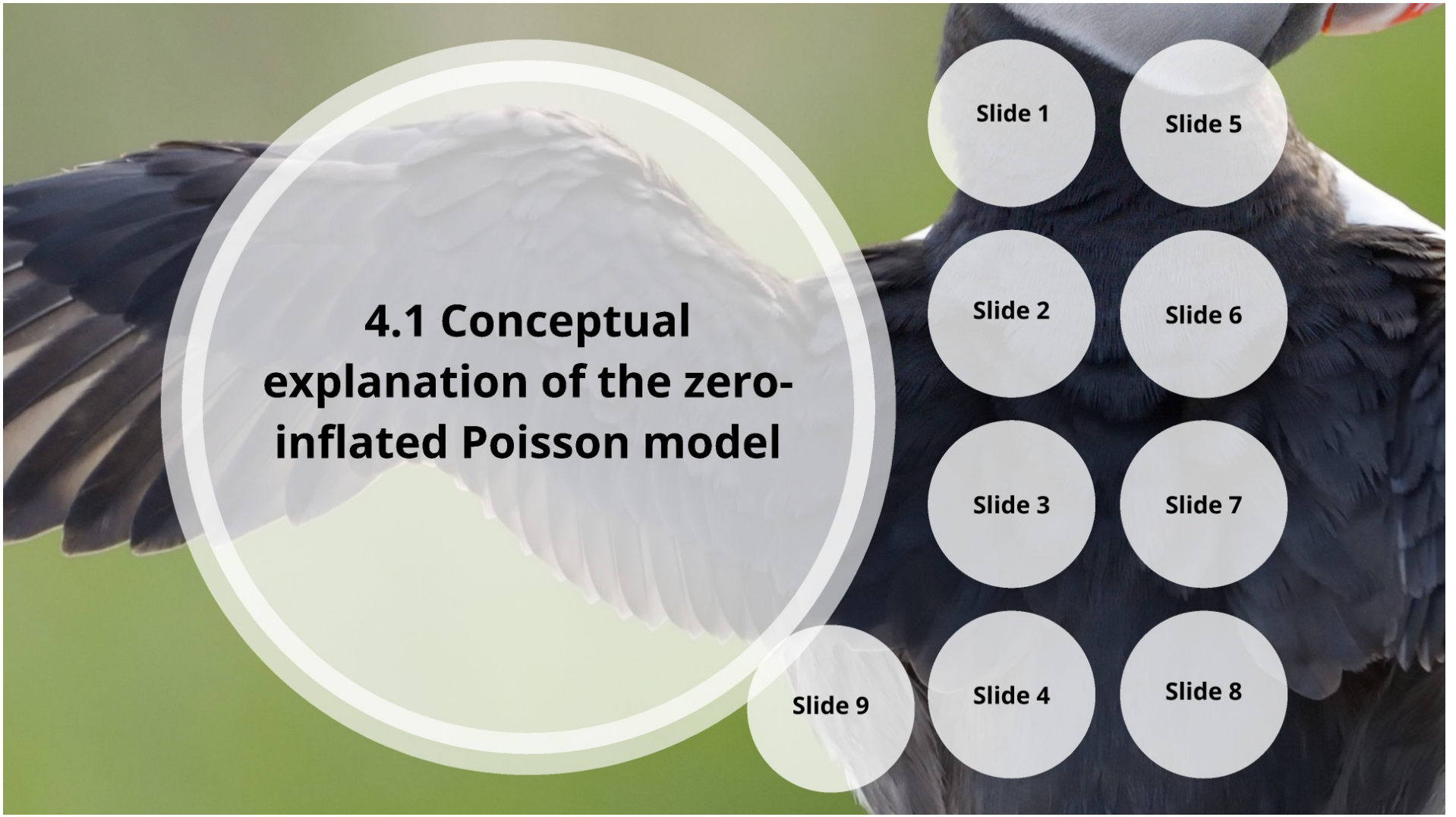
Some general comments:

- Never start with a zero-inflated model
  - First try an ordinary GLM.
  - A covariate might also be able to model the large number of zeros.

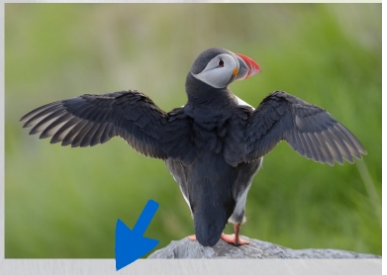
- It is generally recommended to apply GLMs on count data instead of transforming the data followed by a linear regression analysis.
  - Warton (2018).

- A transformation on the response variable does not solve the problem of zero inflation.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.1 Conceptual explanation of the zero-inflated Poisson model
Slide 1
Slide 5
Slide 2
Slide 6
Slide 3
Slide 7
Slide 9
Slide 4
Slide 8

Suppose that we are going to the cliffs and sample multiple plots for the presence or absence of puffin nests.



Bernoulli GLM:

- Consider 1 as success.
- Consider 0 as failure.
- π is the probability of success.
- A large value of π means that most likely we will see puffin nests.

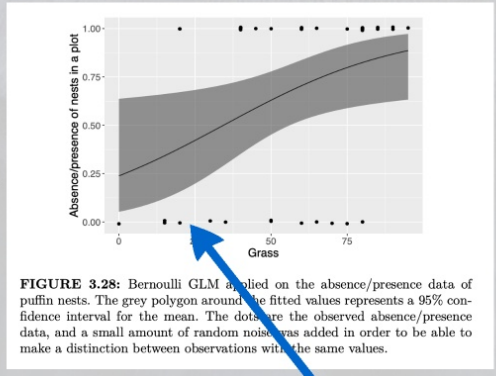
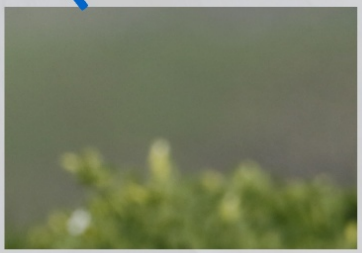


FIGURE 3.28: Bernoulli GLM applied on the absence/presence data of puffin nests. The grey polygon around the fitted values represents a 95% confidence interval for the mean. The dots are the observed absence/presence data, and a small amount of random noise was added in order to be able to make a distinction between observations with the same values.

In this section, we will flip the definition of π:

- Consider 0 as success.
- Consider 1 as failure.
- A large π means that most likely we will not see any puffin nests in the plot.



Option 1: Use the logit link function:

$$\text{logit}\left(\pi_i\right) = \text{Intercept} + \beta \times \text{Grass}_i$$

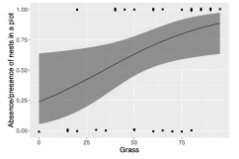


FIGURE 3.28: Bernoulli GLM applied on the absence/presence data of puffin nests. The grey polygon around the fitted values represents a 95% confidence interval for the mean. The dots are the observed absence/presence data, and a small amount of random noise was added in order to be able to make a distinction between observations with the same values.

Option 2: Keep it simple and only use an intercept:

- The probability π is constant.

$$\text{logit}\left(\pi_i\right) = \text{Intercept}$$

The term π_i is the probability of measuring no puffin nests in plot i. The logit link function will be used to model it. This link function may (or may not) contain covariates. Using covariates greatly increases the complexity of the ZIP model.

Let us simulate some puffin data in R

- Suppose that the probability of seeing no puffin nests in a 3-by-3-meter plot is 0.62.
- This corresponds to an intercept of 0.5:

$$\pi_i = \frac{\exp(0.5)}{1 + \exp(0.5)} = 0.62$$

Interpretation:

- The probability of absence (0) is 0.62.
- It is more likely to see no puffin nests (0) in a plot as compared to seeing puffin nests in a plot (1).

We will use the rbinom function to simulate either a 0 or a 1

```
Pi <- exp(0.5) / (1 + exp(0.5))
Pi
```

```
## [1] 0.6224593
```

```
set.seed(12345) #Set the random seed
W <- rbinom(n = 1000, size = 1, prob = 1 - Pi)
table(W)
```

```
## W
##   0   1
## 598 402
```

Suppose that we not only sample the absence and presence of puffin nests but also the actual numbers of nests per plot.

Use a Poisson distribution for this process

- To generate zero-inflated count data we apply a small trick!
- Use u_i x W_i as the mean value of the Poisson distribution.

- W_i is the outcome of the Bernoulli draw.

$$
\begin{aligned}
W_i &\sim \text{Bernoulli}\,(\pi_i) \\
\text{Nests}_i &\sim \text{Poisson}\,(\mu_i \times W_i) \\
\log(\mu_i) &= \beta_1 + \beta_2 \times \text{Distance}_i \\
\text{logit}\,(\pi_i) &= \gamma_1
\end{aligned}
\tag{4.1}
$$

β_1 is the intercept for the count part of the model.
γ_1 is the intercept for the Bernoulli part of the model.

$$
\begin{aligned}
W_i &\sim \text{Bernoulli}(\pi_i) \\
\text{Nests}_i &\sim \text{Poisson}(\mu_i \times W_i) \\
\log(\mu_i) &= \beta_1 + \beta_2 \times \text{Distance}_i \\
\text{logit}(\pi_i) &= \gamma_1
\end{aligned}
\tag{4.1}
$$

This is the mu_i part:

```
Distance <- seq(from = 3,       #Generate Distance values
                to = 60,        #that have similar values as
                length = 1000) #the real Distance values.
mu <- exp(-0.1 + 0.05 * Distance)
```

We already have the W_i which is the vector with zeros and ones.

We obtain zero-inflated numbers of nests per plot by drawing data from a Poisson distribution with mean: u_i x W_i.

```
Nests <- rpois(1000, lambda = mu * W)
```

The main trick to generate zero-inflated data is to sample counts via `rpois(..., lambda = W*mu)`. Every value in `W` that equals 0 will result in a simulated count value of 0.

Combine the relevant variables in a data frame:

```
ZipData <- data.frame(Distance = Distance,
                      Nests    = Nests)
```

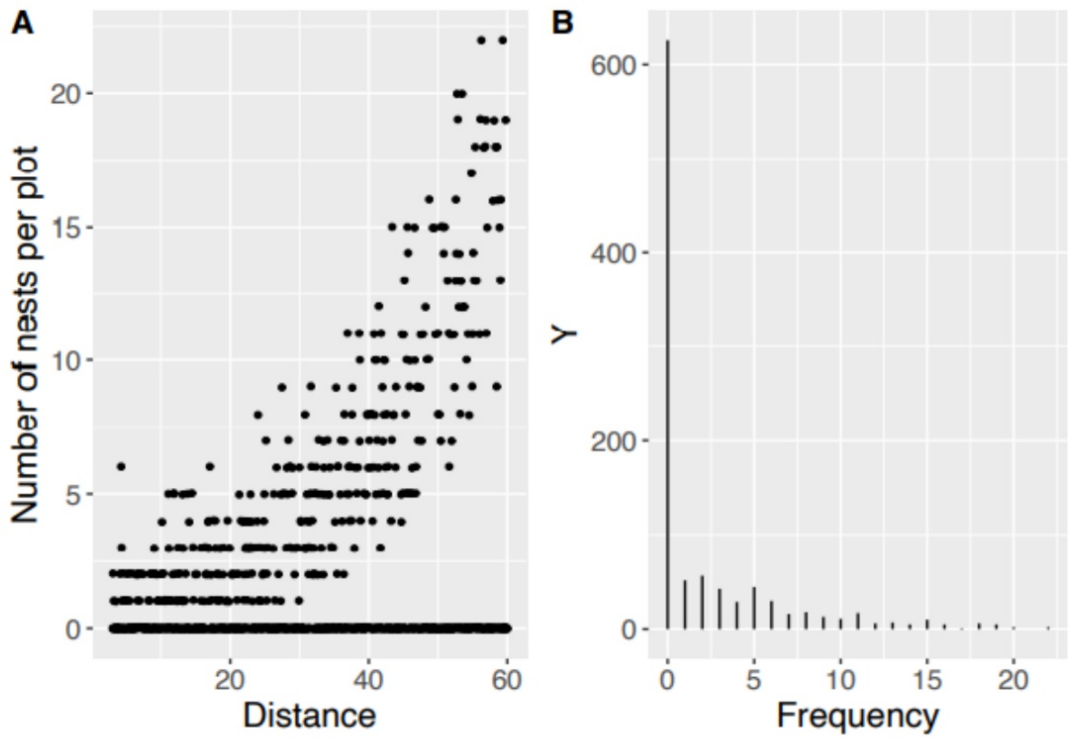


FIGURE 4.1: A: Simulated zero-inflated number of nests in a plot. A small amount of random noise was added to aid visualisation. B: Frequency plot of the simulated number of nests in a plot.

Let us combine all the components that we created in an object with the name Fantasy

```
Fantasy <- data.frame(PlotID   = 1:1000,
                      Distance = Distance,
                      W        = W,
                      mu       = mu,
                      W.mu     = W * mu,
                      Nests    = Nests)
tail(Fantasy, 10)
```

```
##      PlotID Distance W       mu     W.mu Nests
## 991     991 59.48649 0 17.71345  0.00000     0
## 992     992 59.54354 0 17.76406  0.00000     0
## 993     993 59.60060 0 17.81481  0.00000     0
## 994     994 59.65766 0 17.86570  0.00000     0
## 995     995 59.71471 0 17.91674  0.00000     0
```

Distance and Nests:

- Sampled in the field.
- Will be used for the statistical analyses.

We don't sample, see or know the W and mu components.

- We can only estimate them with the ZIP model.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.2 Fitting the zero-inflated Poisson model in R
Slide 1
Slide 2
Slide 3

Mathematical expression for the ZIP distribution

$$\begin{aligned} \text{Nest}_i &\sim \text{ZIP}(\mu_i, \pi_i) \\ \text{E}[\text{Nests}_i] &= (1-\pi_i) \times \mu_i \\ \text{var}[\text{Nests}_i] &= (1-\pi_i) \times (\mu_i + \pi_i \times \mu_i^2) \end{aligned} \tag{4.2}$$

$$\begin{aligned} \log(\mu_i) &= \beta_1 + \beta_2 \times \text{Distance}_i \\ \text{logit}(\pi) &= \gamma_1 \end{aligned} \tag{4.3}$$

If $\pi = 0$, then we obtain the Poisson GLM.

Option 1: Fit the model with zeroinfl from pscl.

```
library(pscl)
M1 <- zeroinfl(Nests ~ 1 + Distance | 1, data = ZipData)
```

Option 2: Use the glmmTMB package.

```
library(glmmTMB)
M2 <- glmmTMB(Nests ~ 1 + Distance,
              ziformula =~ 1,
              family = "poisson",
              data = ZipData)
```

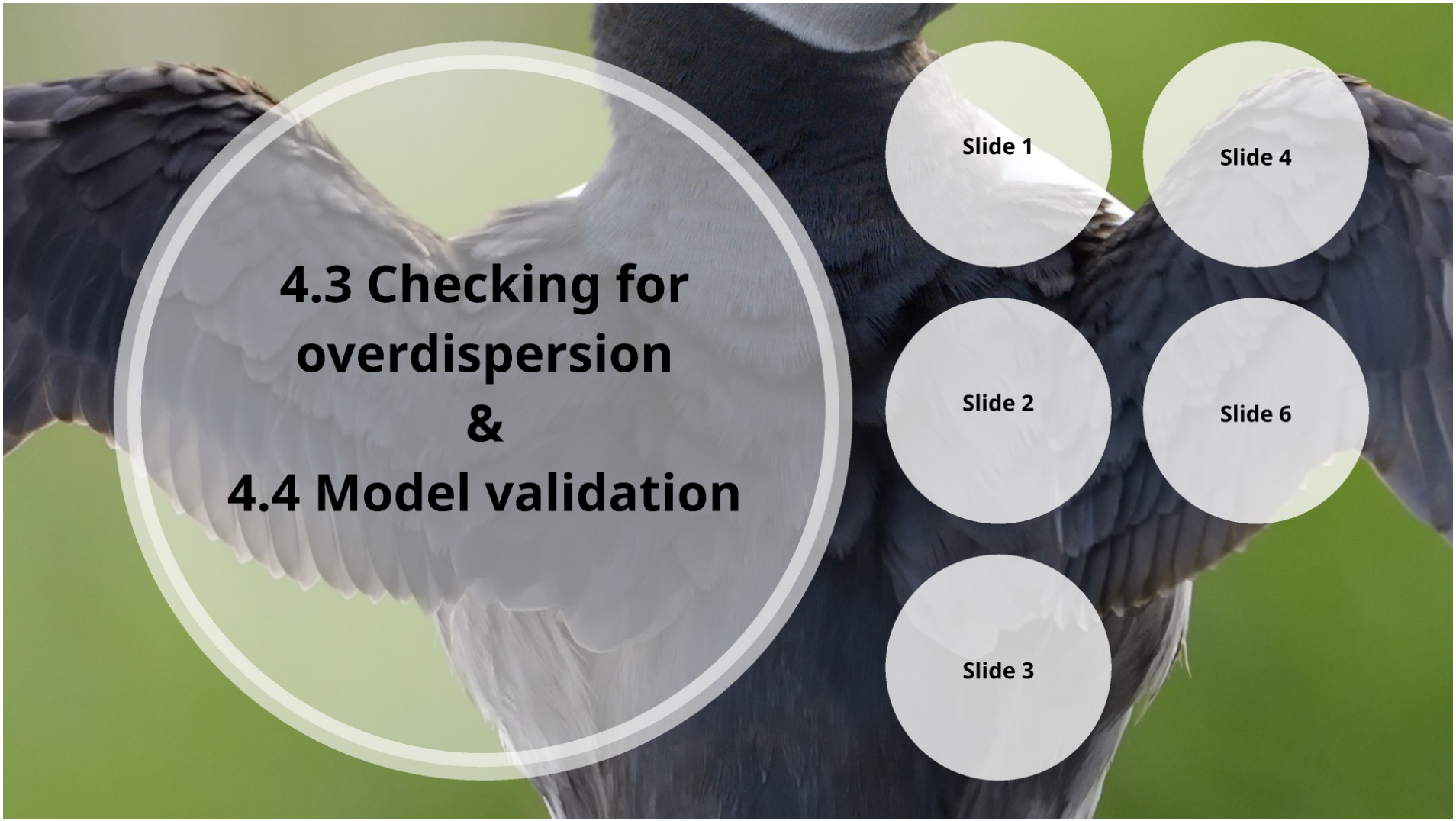
Numerical output:

```
summary(M2)
```

```
##  Family: poisson  ( log )
## Formula:          Nests ~ 1 + Distance
## Zero inflation:         ~1
## Data: ZipData
##
##      AIC      BIC   logLik deviance df.resid
##   2843.8   2858.5  -1418.9   2837.8      997
##
##
## Conditional model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.036953   0.073295  -0.504    0.614
## Distance     0.048700   0.001644  29.616   <2e-16
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.40895    0.06847   5.973 2.33e-09
```

$$\begin{aligned} \log(\mu_i) &= -0.036 + 0.048 \times \text{Distance}_i \\ \text{logit}(\pi) &= 0.408 \end{aligned} \tag{4.4}$$


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.3 Checking for overdispersion
&
4.4 Model validation
Slide 1
Slide 4
Slide 2
Slide 6
Slide 3

4.3 Checking for overdispersion

We need to calculate the Pearson residuals:

```
> resid(M2, type="pearson")
```

‘Pearson residuals are not implemented for models with zero-inflation or variable dispersion’.

We need to calculate these ourselves!

```r
X <- model.matrix(M2)
head(X, 4)
```

```
##   (Intercept) Distance
## 1           1 3.000000
## 2           1 3.057057
## 3           1 3.114114
## 4           1 3.171171
```

```r
betas  <- fixef(M2)$cond
muPois <- exp(X %*% betas)
```

```r
gamma1 <- fixef(M2)$zi
Pi <- exp(gamma1) / (1 + exp(gamma1))
```

```r
ExpY <- (1 - Pi) * muPois        #Expected ZIP values
VarY <- (1 - Pi) * (muPois + Pi * muPois^2) #Variance
E2   <- (ZipData$Nests - ExpY) / sqrt(VarY) #Pearson residuals
```

```r
N  <- nrow(ZipData)
k  <- length(fixef(M2)$cond) + length(fixef(M2)$zi)
Dispersion <- sum(E2^2) / (N - k)
Dispersion
```

```
## [1] 1.01293
```

Also assess overdispersion via a simulation study

- Similar as we did for the Poisson GLM.

Sim2 <- simulate(M2, 1000, seed = 12345)

To do this manually:

```
N <- nrow(ZipData)
Sim2Self<- matrix(nrow = N, ncol = 1000)
for (i in 1:N){
  Sim2Self[,i] <- VGAM::rzipois(N, lambda = muPois, pstr0 = Pi)
}
```

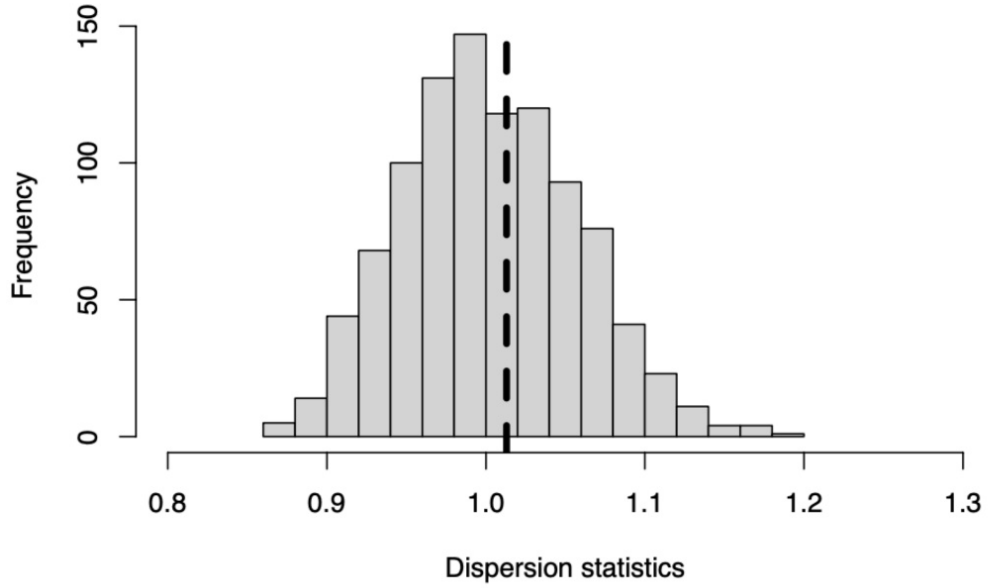


FIGURE 4.2: Results of the simulation study for the ZIP model to investigate the range of dispersion statistics. The vertical dotted line represents the dispersion statistic obtained by applying the ZIP model on the data.

4.4. Model validation

- Plot Pearson residuals versus the fitted values.
- Plot Pearson residuals versus all covariates in the model
- Check Pearson residuals for spatial and/or temporal dependency (if relevant).

We can also use DHARMa.

Doing a simulation exercise (simulating data and applying the same model) gives a good impression of what patterns can be expected from a specific model.

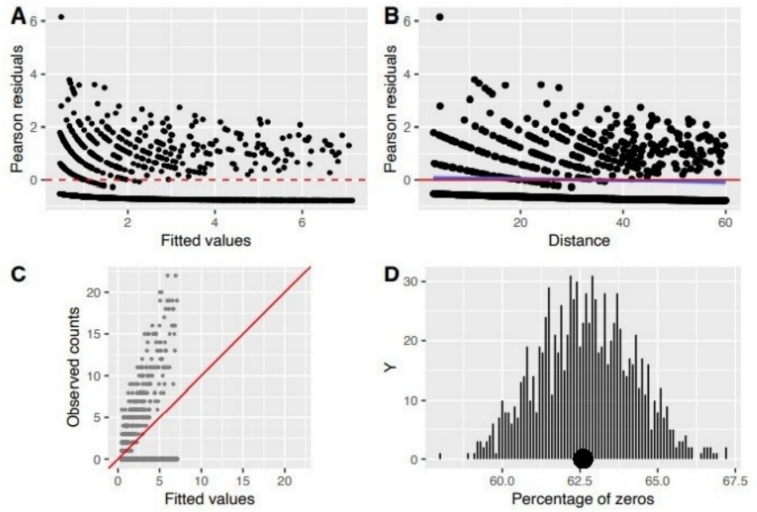


FIGURE 4.3: A: Pearson residuals plotted versus fitted values obtained by the ZIP model applied on the simulated data. B: Pearson residuals versus the covariate Distance. C: The nest data plotted versus the fitted values of the ZIP model. D: Frequency plot of the percentage of zeros in 1000 simulated data sets from the model, and the percentage of zeros in the nest data (dot).


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.5 Zero-inflated Poisson model visualisation
Slide 1
Slide 2

Zero-inflated Poisson model visualisation

```
MyData <- data.frame(Distance = seq(from = min(ZipData$Distance),
                                    to   = max(ZipData$Distance),
                                    length = 50))
```

```
#Poisson part
Xpois   <- model.matrix(~1 + Distance, data = MyData)
betas   <- fixef(M2)$cond
mu.pois <- exp(Xpois %*% betas)
#Bernoulli part
gammas <- fixef(M2)$zi
Pi <- exp(gammas) / (1 + exp(gammas))
```

```
MyData$ExpY.zip <- (1 - Pi) * mu.pois
```

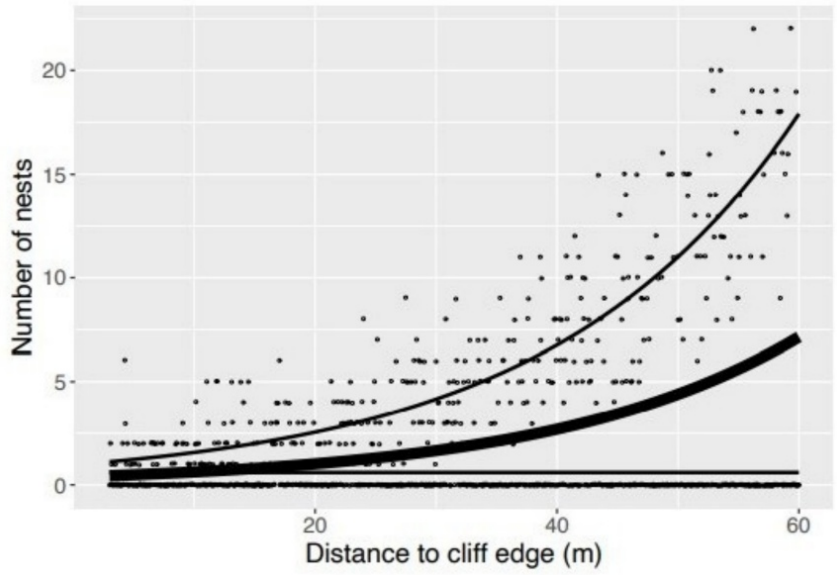


FIGURE 4.4: Fitted values of the ZIP model (thick solid line in the middle), fit of the Poisson part (upper thinner line), and the fit of the Bernoulli part (lower thinner line). The thick line at 0 indicates the observations equal to 0.

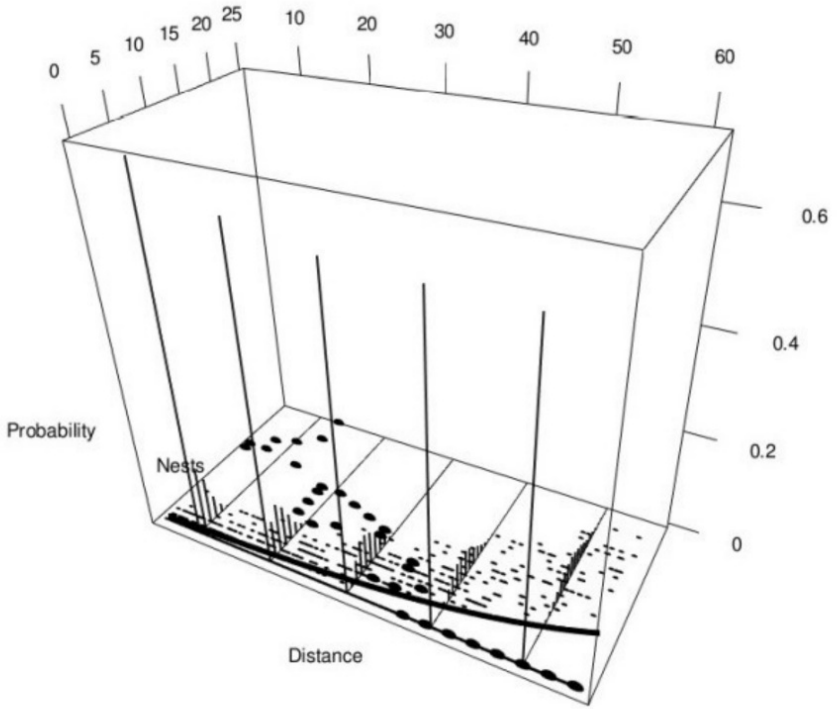


FIGURE 4.5: 3-D presentation of the ZIP model applied on simulated puffin data.

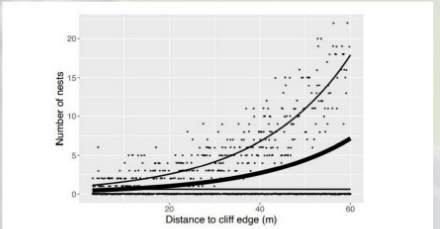


FIGURE 4.4: Fitted values of the ZIP model (thick solid line in the middle), fit of the Poisson part (upper thinner line), and the fit of the Bernoulli part (lower thinner line). The thick line at 0 indicates the observations equal to 0.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.6 Applying a Poisson GLM on zero-inflated data
Slide 1
Slide 2

Suppose that:

- We do not know that the data were simulated from a ZIP model.
- And we apply a Poisson GLM.

```
M3 <- glm(Nests ~ 1 + Distance,
          family = "poisson",
          data = ZipData)
print(summary(M3)$coefficients, digits = 3)
```

```
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.6685    0.06387   -10.5  1.23e-25
## Distance      0.0401    0.00145    27.6 6.88e-168
```

This model is overdispersed.

```
E3 <- resid(M3, type = "pearson")
sum(E3^2) / (nrow(ZipData) - length(coef(M3)))
```

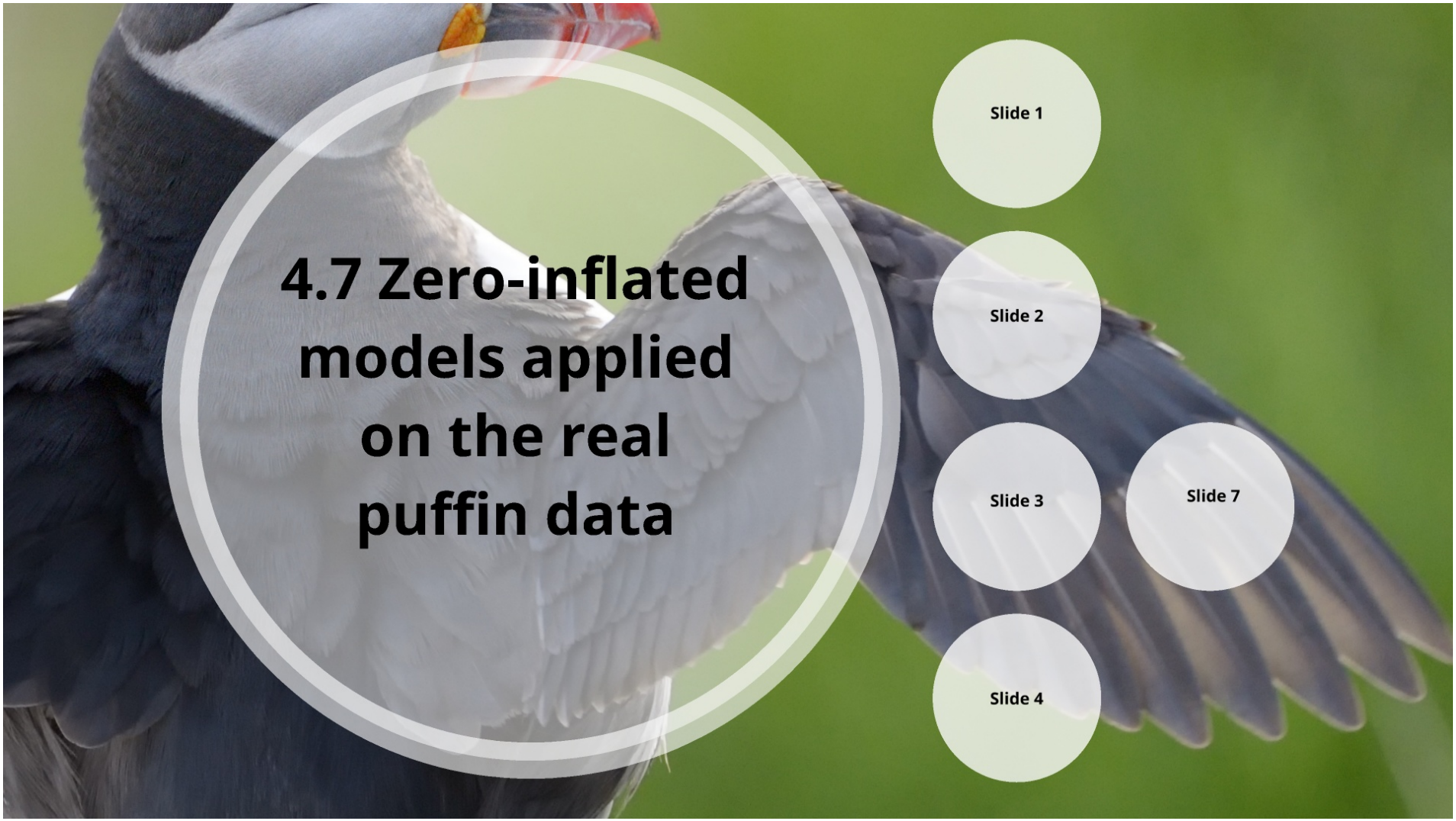
```
## [1] 4.768542
```

We can correct for the overdispersion by applying a quasi-Poisson GLM:

- But this would result in much wider confidence intervals than the ZIP model.

For this specific simulated data set, the Poisson GLM gives similar parameter estimates as the ZIP model. However, the Poisson GLM is overdispersed as both the zeros and the non-zero data are outside the range of the Poisson density curves. It is not too difficult to simulate data for which the Poisson and ZIP models give different estimates. The moral of this section is that selecting the wrong model (e.g. a quasi-Poisson GLM) may (or may not) lead to biased parameter estimates and for sure to standard errors that are unnecessarily wide.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.7 Zero-inflated models applied on the real puffin data
Slide 1
Slide 2
Slide 3
Slide 7
Slide 4

We will now return to the puffin data from Nettleship (1972).

We are going to do something that we normally recommend not to do:

- Applying a series of models on the puffin data.
- Compare them via their AICs.

The reason that we recommend not to do this is as follows.

- If the Poisson GLM is overdispersed, then we need to figure out why it is overdispersed.
- A ZIP model should be applied next only if the excessive number of zeros is the cause.
- If the overdispersion is due to a missing covariate, or a covariate effect that is non-linear, then alternative models should be applied.

Hence, the choice of which model should be applied after the Poisson GLM is a very delicate process.

- If the wrong model is selected, then we may end up with the wrong ecological conclusions.

Blindly applying the Poisson, negative binomial (NB), generalised Poisson (GP), Conway–Maxwell–Poisson (CMP) models and also their zero-inflated cousins does not comply with this approach.

However, the motivation for doing this at this stage of the book is that we just want to show how to implement these models.

Once we are familiar with the coding and numerical output, we will present a series of case studies where we apply the correct analysis strategy in later chapter

This section is about the tools, not how to apply them in an appropriate way.

Poisson and zero-inflated Poisson models

$$\begin{aligned} \text{Nests}_i &\sim \text{Poisson}\,(\mu_i) \\ \text{E}\,[\text{Nests}_i] &= \mu_i \\ \text{var}\,[\text{Nests}_i] &= \mu_i \end{aligned} \tag{4.5}$$

The expected values μ_i are modelled with a log-link function.

$$\mu_i = \exp(\text{Intercept} + \beta \times \text{Distance}_i) \tag{4.6}$$

```
M.pois <- glmmTMB(Nesting ~ Distance,
                  family = "poisson",
                  data = PF)
```

The ZIP model is defined in Equation (4.7).

$$\begin{aligned} \text{Nest}_i &\sim \text{ZIP}\,(\mu_i, \pi_i) \\ \text{E}\,[\text{Nests}_i] &= (1-\pi_i) \times \mu_i \\ \text{var}\,[\text{Nests}_i] &= (1-\pi_i) \times (\mu_i + \pi_i \times \mu_i^2) \end{aligned} \tag{4.7}$$

$$\pi_i = \frac{\exp(\text{Intercept})}{1 + \exp(\text{Intercept})}$$

```
M.zip <- glmmTMB(Nesting ~ Distance,
                 family = "poisson",
                 ziformula =~ 1,
                 data = PF)
```

NB and zero-inflated NB models

The mean and the variance of an NB GLM are given in Equation (4.9).

$$\begin{aligned} \text{Nests}_i &\sim \text{NB}(\mu_i, \theta) \\ \text{E}[\text{Nest}_i] &= \mu_i \\ \text{var}[\text{Nest}_i] &= \mu_i + \frac{\mu_i^2}{\theta} \end{aligned} \tag{4.9}$$

```
M.nb <- glmmTMB(Nesting ~ Distance,
                family = "nbinom2",
                data = PF)
M.zinb <- glmmTMB(Nesting ~ Distance,
                  family = "nbinom2",
                  ziformula =~ 1,
                  data = PF)
```

$$\begin{aligned} \text{Nests}_i &\sim \text{ZINB}(\mu_i, \pi_i, \theta) \\ \text{E}[\text{Nest}_i] &= (1-\pi_i) \times \mu_i \\ \text{var}[\text{Nest}_i] &= (1-\pi_i) \times \mu_i \times (1 + \pi_i \times \mu_i + \frac{\mu_i}{\theta}) \end{aligned} \tag{4.10}$$

Compare all models via AIC

```
> AIC(M.pois, M.zip, M.nb, M.zinb)
       df      AIC
M.pois  2 196.8612
M.zip   3 181.6142
M.nb    3 187.9263
M.zinb  4 183.5141
```

ZIP model has the lowest AIC.
Its interpretation is like the model for the simulated data.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4


4.8 True and false zeros
Slide 1
Slide 5
Slide 2
Slide 6
Slide 3
Slide 7
Slide 4

We explained the ZIP as a combination of:

- A Bernoulli flip of a coin that creates zeros and ones.
- A Poisson process that creates the counts.

Kuhnert et al. (2005) and Martin et al. (2005):

- Introduced the concept of true and false zeros.
- Their ideas were applied in Zuur et al. (2009a, 2012a, 2016a, 2018).
- We will explain them here as well.

Suppose that the observed number of puffin nests contains a large number of zeros.

Why do we have zeros?

- Sampling took place outside the range of the species distribution.
- The observer made an error.
- The environmental conditions are not favourable for puffins.

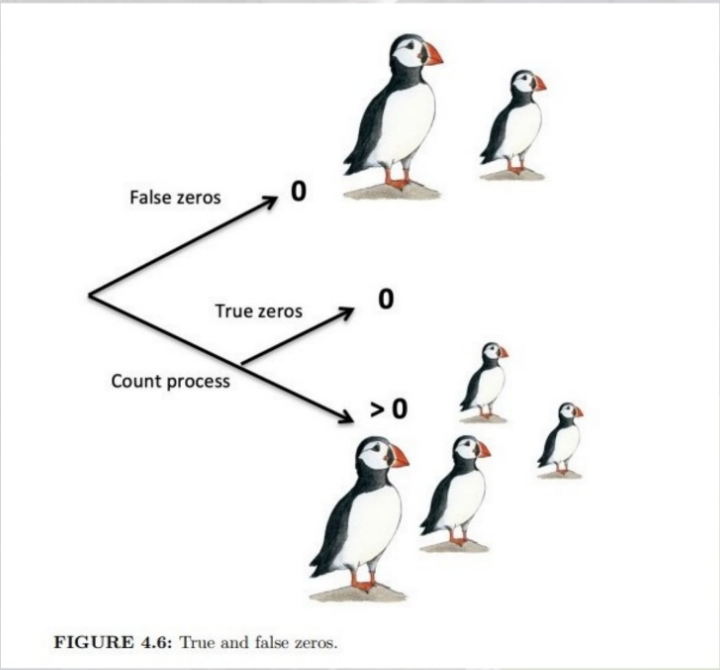


FIGURE 4.6: True and false zeros.

Lower branch shows the phrase ‘count process’.

- Is going to be a Poisson GLM.
- Can also use other distributions:
  - Negative binomial.
  - Binomial.
  - Beta.
  - Gamma.

The Poisson GLM part:

- Contains covariates.
- May (or may not) produce zeros.
  - These are called true zeros;
  - They comply with the covariates in the Poisson model.

False zeros:

- These are zeros that do not comply with the covariates in the Poisson part of the model.

The label 'false' implies:

- We don't like them.
- They are not good.

Zuur et al. (2009a, 2012a) used arguments like:

- Sampling at the wrong time or place,
- The sampling period is too short,
- Observer error, etc.

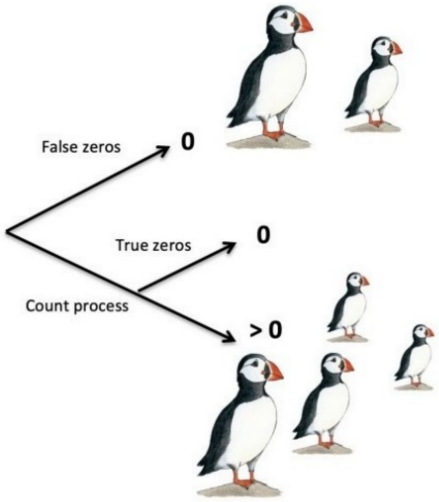


FIGURE 4.6: True and false zeros.

True and false zeros:

- Is only an excuse for including a component in the model that captures the zeros that do not comply with the Poisson part of the model.

If a covariate in the count part explains the zeros, and we drop this covariate:

- We may end up with a lot of false zeros that were originally true zeros!

'False zeros':

- We pretend that they could have been non-zeros.
- A zero in the spreadsheet, but a nonzero in the field.

But this is just imagination, an ecological excuse to apply a ZIP model.

When we fit the model, we do not know which zeros are true and false.

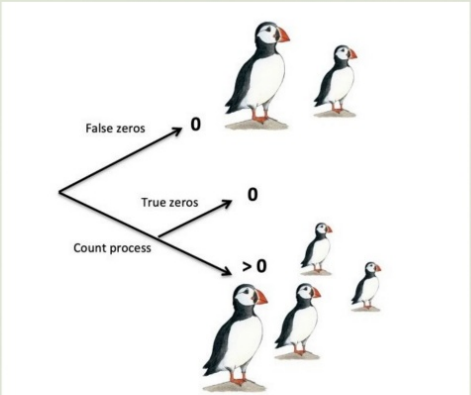


FIGURE 4.6: True and false zeros.

The zeros that:

- Comply with the count part of a ZIP model will have a low probability of being false zeros.
- Do not comply with the count part will have a higher probability of being false zeros.

In Subsection 4.1 we created the W vector.

- It sets certain values to 0, even if there was a positive count originally.

As a result:

- The "Nests" vector had an excessive number of zeros.

The nonzero values in Nests that were set to 0 due to W are the false zeros.

- We sampled a zero but there was a nest, and we can't explain it with a covariate in the count part.

True zeros:

- The zeros in Nests that were already 0 due to the covariate effect.

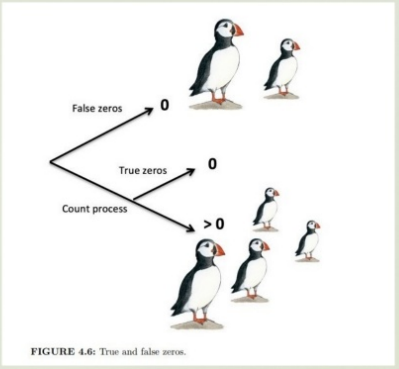


FIGURE 4.6: True and false zeros.

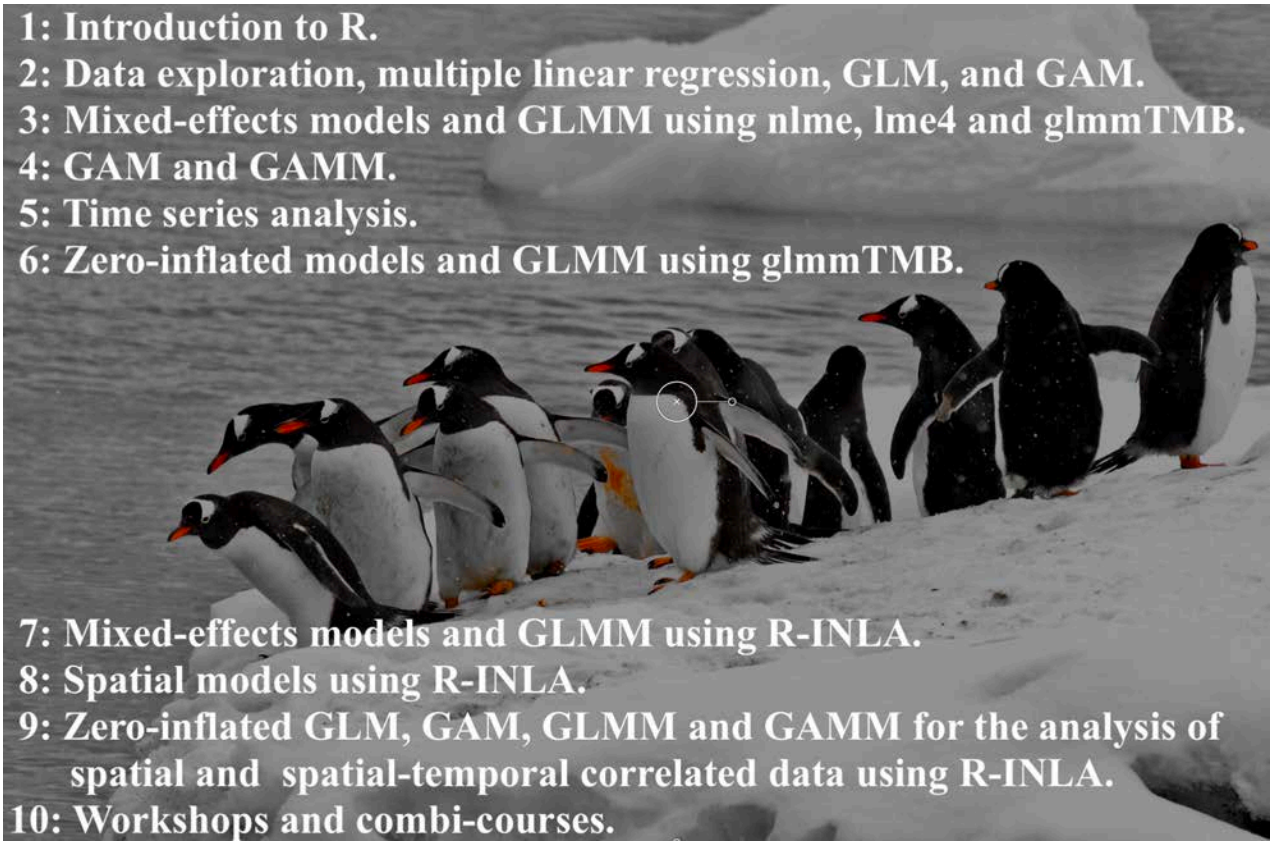
```
Fantasy <- data.frame(PlotID   = 1:1000,
                      Distance = Distance,
                      W        = W,
                      mu       = mu,
                      W.mu     = W * mu,
                      Nests    = Nests)
tail(Fantasy, 10)
```

```
##       PlotID Distance W       mu     W.mu Nests
## 991      991 59.48649 0 17.71345  0.00000     0
## 992      992 59.54354 0 17.76406  0.00000     0
## 993      993 59.60060 0 17.81481  0.00000     0
## 994      994 59.65766 0 17.86570  0.00000     0
## 995      995 59.71471 0 17.91674  0.00000     0
## 996      996 59.77177 1 17.96793 17.96793    19
## 997      997 59.82883 0 18.01926  0.00000     0
## 998      998 59.88589 0 18.07074  0.00000     0
## 999      999 59.94294 0 18.12237  0.00000     0
## 1000    1000 60.00000 0 18.17415  0.00000     0
```

Hey observer... you made an error. We were just at a different cliff. We were counted as a '*false*' zero.


Introduction
4.1 Conceptual explanation of the zero-inflated Poisson model
4.2 Fitting the zero-inflated Poisson model in R
4.3 Checking for overdispersion
4.5 Zero-inflated Poisson model visualisation
4.6 Applying a Poisson GLM
4.7 Zero-inflated models on real puffin data
4.8 True and false zeros
The World of Zero-Inflated Models
Chapter 4

HIGHLAND STATISTICS COURSES



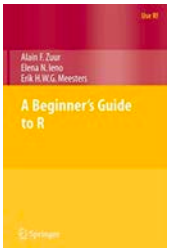
Highland Statistics Ltd. provides 10 different statistics courses. The course instructors are Dr. Alain Zuur (statistician) and Dr. Elena Ieno (biologist). Our statistical and biological backgrounds ensure a lively and enjoyable interaction with our participants. One of our strong points is explaining statistics in a non-technical and understandable language.

Some of our (online or onsite) courses are run as in-house courses whereas other courses are open. For a list of upcoming courses, see: http://highstat.com.

If you have various colleagues who are interested in one of our courses it may be more cost-effective to organise a course at your institute. This can be done as an in-house course or as an open course. With an in-house course, you decide who participates and we charge a fixed fee. Our fee will depend on the country. For an open course, we will require a conference room (plus projector) for about 30 people, and an additional 10 local participants.

SHORT COURSE DESCRIPTION

Course 1: Introduction to R using a protocol for conducting and presenting results of regression-type analyses



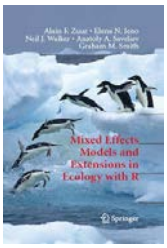
- In this course, we provide an introduction to R and at the same time explain how to conduct data exploration, apply (simple) linear regression models, communicate results, and also determine optimal sample size (using power analysis) in case you want to set up a new field study or experiment.

Course 2: Data exploration, regression, GLM and GAM: with an intro to R.

- We begin with an introduction to R and provide a protocol for data exploration to avoid common statistical problems. We will discuss how to detect outliers, deal with collinearity and transformations. An important statistical tool is multiple linear regression. Various basic linear regression topics will be explained from a biological point of view. We will discuss potential problems and show how generalised linear models (GLM) can be used to analyse count data, presence-absence data and proportional data. Sometimes, parametric models (linear regression, GLM) do not quite fit the data and in such cases generalised additive models (GAM; a smoothing technique) can be used.

Course 3: Introduction to mixed effects models and GLMM

- The course starts with a short revision of multiple linear regression and generalised linear models, followed by an introduction to linear mixed-effects models and generalised linear mixed-effects models (GLMM) to analyse hierarchical or clustered data, e.g. multiple observations from the same animal, site, area, nest, patient, hospital, vessel, lake, hive, transect, etc. In the second part of the course GLMMs are applied on continuous (e.g. biomass), binary (e.g. absence/presence of a disease), proportional (e.g. % coverage) and count data using the Gaussian, Poisson, negative binomial, Bernoulli, binomial, beta, and gamma distributions.



Course 4: Introduction to GAM and GAMM

- We start with a short revision of data exploration and linear regression. We then introduce generalised additive models (GAM) to model non-linear relationships. We will execute these models in mgcv. In Module 2, we will revise linear mixed-effects models and show how to implement a generalised additive mixed-effects model (GAMM). We also show how to include an interaction between a smoother and a categorical covariate. In Module 3, we will revise basic GLMs and extend these towards GAMs. In Modules 4 and 5 we will discuss GAMMs for the analysis of count data,

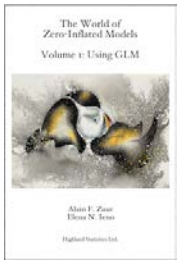
absence-presence data, proportional data, and continuous data. We also discuss 2-dimensional smoothers (including the soap-film smoother for study areas with barriers; e.g. an island in the sea).

Course 5: Time series analysis using regression techniques

- The course starts with a short revision of data exploration and multiple linear regression models. A non-technical introduction of generalised additive models (GAM) is provided. GAMs will be used to estimate long-term trends, seasonal patterns, covariate effects and auto-regressive correlation. We also provide a short introduction to linear mixed-effects models and generalised linear mixed-effects models (GLMM) to analyse hierarchical data (e.g. short time series from the same core or site). GLMMs and GAMMs are used to estimated trends, seasonality and covariate effects in multivariate time series.

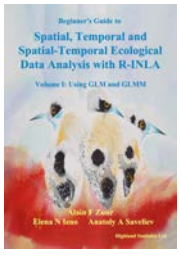
Course 6: Introduction to zero-inflated models

- The course starts with a short revision of data exploration, multiple linear regression and Poisson GLM. We then discuss 3 more models for the analysis of count data, namely the negative binomial, generalised Poisson and Conway-Maxwell-Poisson GLMs. After a short theory presentation in which we explain how to extend these models towards zero-inflated models, we apply them to various data sets. We also use the Tweedie GLM and the zero-altered Gamma GLM for the analysis of zero-inflated continuous data. In the second part of the course, we start with a short revision of linear mixed-effects models. This is followed by a series of exercises in which we analyse zero-inflated count data, continuous data, and proportional data using zero-inflated GLMMs. Throughout the course we will use the glmmTMB package in R.



Course 7: Mixed-effects models and GLMM using R-INLA

- The course begins with a brief revision of multiple linear regression, followed by an introduction to Bayesian analysis and how to execute regression models in R-INLA. We then explain linear mixed-effects models to analyse nested data, followed by a series of mixed modelling exercises in R-INLA. Nested data means multiple observations from the same animal, site, area, nest, patient, hospital, vessel, lake, hive, transect, etc. In the second part of the course GLMMs are applied on count data, binary data (e.g. absence/presence of a disease), proportional data (e.g. % coverage) and continuous data (e.g. biomass or distance) using the Poisson, negative binomial, Bernoulli, binomial, beta and gamma distributions.



Course 8: Intro to Regression Models with Spatial Correlation using R-INLA

- We begin with an introduction how to add dependency to regression models using frequentist tools. After discussing the limitations of this approach we switch to Bayesian

techniques. R-INLA is used to implement regression models, generalised linear models (GLM) and generalised linear mixed-effects models (GLMM) with spatial dependency

Course 9: Zero-inflated GLM, GAM, GLMM and GAMM for the analysis of spatial and spatial-temporal correlated data using R-INLA

- We will start with a short revision of multiple linear regression, followed by a basic introduction to Bayesian analysis, and we show how to execute a linear regression model in R-INLA. In the second module, we will explain how to deal with zero-inflated count data and zero-inflated continuous data using zero-inflated Poisson, zero-inflated negative binomial, and zero-inflated Gamma GLMs. We also explain hurdle models. In the third module, we will introduce generalised additive models (GAM) to model non-linear relationships. We show how to execute these in mgcv and also in R-INLA. In the fourth part of the course, we will revise linear mixed-effects models and implement these in R-INLA. We also apply generalised linear mixed-effects models (GLMM) and generalised additive mixed-effects model (GAMM) in R-INLA. In the fifth part of the course, we will apply zero-inflated GAMs and GAMMs (and GLMMs) on various spatial correlated data sets. In module 6, we apply GAM, GAMM, and GLMM on spatial-temporal correlated data. We also deal with natural barriers for the spatial correlation (e.g. benthic species that live on a coral reef around an island). We will use barrier models; these ensure that spatial correlation seeps around a barrier (in this case an island). All exercises are executed in R-INLA.

Course 7: Workshop and combi-course

- Combine the appropriate modules and use your own data sets during the course.

RECOMMENDED ORDER OF COURSES

If you do not have spatial or temporal data, then we recommend to attend the following courses within a time span of 3 years.

1. Introduction to data exploration, regression, GLM and GAM. With introduction to R (course 2).
2. Introduction to mixed effects models and GLMM (course 3).
3. Depending on whether you have zero-inflation and/or non-linear relationships and/or time series, you can then attend the 'Introduction to GAM and GAMM', 'Introduction to zero-inflated models', or the 'Time series' course (courses 4, 5 or 6).

If you have spatial (or spatial-temporal) data, then we recommend the following courses.

1. Intro to data exploration, regression, GLM and GAM. With introduction to R (course 2).
2. Introduction to mixed effects models and GLMM using R-INLA (course 7).
3. Introduction to Regression Models with Spatial Correlation using R-INLA (course 8).
4. If you have zero-inflation and/or non-linear relationships and spatial (-temporal) data, you can then attend the 'Zero-inflated GLM, GAM, GLMM and GAMM for the analysis of spatial and spatial-temporal correlated data using R-INLA' course.

It is possible to skip the 'Introduction to mixed effects models and GLMM using R-INLA' course before taking the spatial INLA courses, but this means a steeper learning curve.

GENERAL INFORMATION

All courses are non-technical, are taught in R, and use a course website that contains:

- 5 - 10 theory presentations (on-demand video and also downloadable pdf files).
- 15 - 25 exercises (on-demand video, with downloadable data sets and documented R code).
- A Discussion Board where you can ask course-related questions.
- Live chat facilities for short questions.
- Recorded Zoom sessions from previous 'Live' online courses are available to watch.

Access to the course website is for 12 months.

Live interaction.

- All courses include a 1-hour face-to-face video chat with the instructors. You can ask questions about your own data.
- For the self-study courses, you can request 4 times a 30 minutes video chat with the instructors and ask questions related to the course material.
- Live interaction needs to take place within 12 months after being given access to the course website.

- You can start an online course at any time.
- The workload of each course is about 40 hours.
- A certificate (pdf file) will be provided upon completing the course or after completing a short online assessment.

For further information go to: http://highstat.com

Email: Dr. Alain F. Zuur at highstat@highstat.com