# Project 3
# Computational Numeric Statistics

João Camacho, N. 56861, Analysis and Engineering of Big Data
Ana Mendes, N. 57144, Analysis and Engineering of Big Data
Lia Schmid , N. 57629, Cienceas Cognitivas
Emanuele Vivoli, N. 57284, Engineering Informatics
Simão Gonçalves, N. 54896, Analysis and Engineering of Big Data

CONTENTS



## I. GENERALIZED LINEAR MODELS

### A. Data

Here we will describe the data set that was used for the project.

This data refers to the horseshoe crabs and satellites data set from J. Brockmann's work on nesting horseshoe crabs in 1996 [2]. We will use Poisson regression to study the relationship between the number of satellites (male crabs residing near a female crab), for a female horseshoe crab and the different features of the female horseshoe crab, such as its color, the width of its back, etc. More information about this data set and the uses of it can be found here: [1] Horseshoe Crabs and Satellites

The data set has 173 examples with 5 features, here is a summary of the features.

```
# Load data
# observations(Obs), crab's color(C),
# spine condition(S),
# carapace width(W), weight(Wt),
# and number of satellites for female
# crabs(Sa)
crab=read.csv("crabs_preprocessed.csv")
summary(crab)
#Output
```

| | Weight | CW | Satellites |
|---|---|---|---|
| Min | 21 | 1.2 | 0 |
| 1st qua | 24.9 | 2 | 0 |
| Median | 26.1 | 2.35 | 2 |
| Mean | 26.3 | 2.437 | 2.919 |
| 3rd qua | 27.7 | 2.85 | 5 |
| Max | 33.5 | 5.2 | 15 |

To better visualize we represent the table in a Boxplot way. Are shown the 50% of the data inside of the box, the first and the third quartile and any data not included between the whiskers is plotted as an outlier with a dot.

The Color and SC features are both categorical and the Weight, CW and Satellites are all numerical.

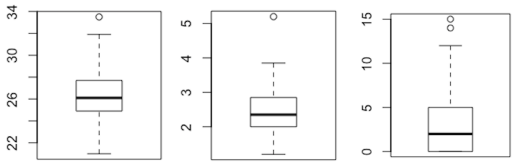



Fig. 1: (left) weight; (central) width; (right) satellites.

The Color values are the female crab's color and have the following labels 1: light; 2: medium light; 3: medium; 4: medium dark; 5: dark.

The SC values are the female crab's spine condition, labeled as 1: both good; 2: one worn or broken; 3: both worn or broken.

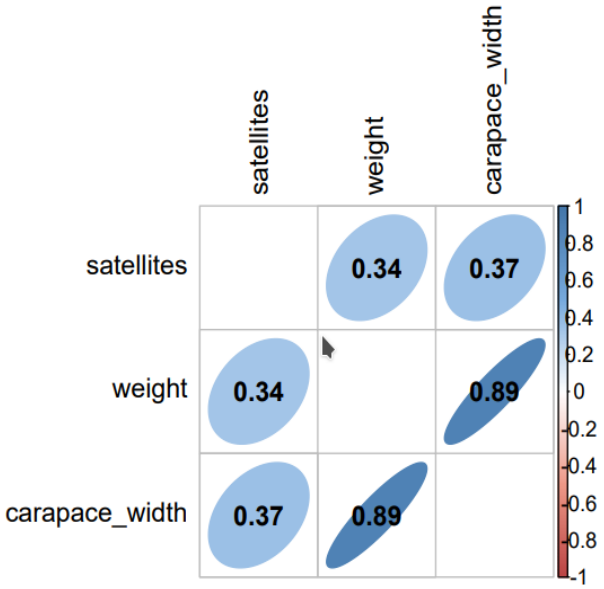
The Weight is the female crab's weight (kg).

The CW is the female crab's carapace width (cm).

And finally the Satellites that are the number of males living close to the female. This is the feature that we are going to try to predict since it is a numerical count value.

Now we will look at the correlation between features:

Fig. 2: Correlation Matrix




As we can see from the correlation matrix the features that are more correlated are the carapace_width and the weight of the crab,probably due to the crab being big in the first place. The most interesting part is that there is some correlation with the carapace width and the number of satellites, this is probably due to a survival instinct in crabs to choose a bigger mate for survival purposes, there is a similar value in correlation between weight and satellites but this is due to the features carapace width and weight being so correlated with each other that they both are similar in nature so they tend to have similar correlation with other variables.

Because of the firsts two features are categorical, we need a way to visualize them in order to see some those frequencies information.

Fig. 3: Colors

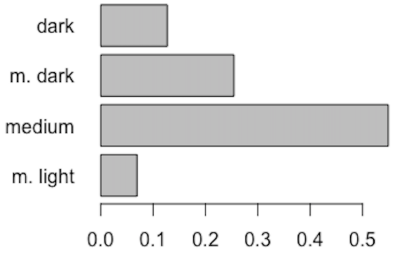
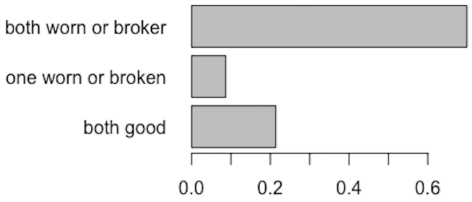



Fig. 4: Spine Condition




The Figures 3 and 4 show that the frequencies of medium color and the frequencies of both spine condition as broken are the highest. We can also see that, even if in the documentation of the data set there is the "light" as color, in the real data set values the values start from "medium light", so there is an absence of one class for the color variable.

### *B. Exponential Family*

Let $K \sim P(\lambda) : \forall k \in \mathbf{N}$: which has probability mass function (p.m.f.) given by:

$$\mathbf{P}[Y = k] = e^{-\lambda}\frac{\lambda^k}{k!}$$

We show that this belongs to the exponential family:

$$\begin{aligned} f(k;\lambda,\phi) &= e^{-\lambda}e^{k\times log(\lambda)}e^{-log(k!)} \\ &= exp\left\{-\lambda + k \times log(\lambda) - log(k!)\right\} \\ &= exp\left\{\frac{k \times log(\lambda) - \lambda}{1} + (-log(k!))\right\} \end{aligned}$$

with this assignment of the following function:

$T(k) = k$
$\eta(\lambda) = log(\lambda)$
$b(\phi) = 1$
$A(\lambda) = \lambda$
$c(k,\phi) = -log(k!)$

we obtain the representation of an Exponential Family distribution.

$$f(k;\lambda,\phi) = exp\left\{\frac{T(k)\times\eta(\lambda) - A(\lambda)}{b(\phi)} + c(k,\phi)\right\}$$

Exists a standard way of presenting this object that is called Canonical form. Often, it is one which provides the simplest representation of an object and which allows it to be identified in a unique way. In case of Exponential family this representation is not unique, but still is important to generalize the form. The Canonical form of the previous equation is when $\eta = log(\lambda)$, and this function is called canonical link.

The Canonical form of our function is so:

$$f(k;\eta) = exp\left\{\frac{k\eta - e^{\eta}}{1} + (-log(k!))\right\}$$

where now the cumulant function is given by $A(\eta) = e^{\lambda}$.

Let X be a random variable with distribution belonging to the exponential family. Then

$$\mathbf{E}(X) = A'(\theta) \tag{1}$$

$$\mathbf{V}(X) = A''(\eta)b(\phi) \tag{2}$$

Proof of (1):
Let us assume T(X) is continuous (discrete analogous). Then we know

$$1 = \int_{\mathbf{R}} f_X(T(x);\theta,\phi)dx \tag{3}$$

Deriving in order $\theta$ one has

$$0 = \frac{d}{d\theta}\int_{\mathbf{R}} f_X(x;\theta,\phi)dx = \int_{\mathbf{R}} \underbrace{\left(\frac{d}{d\theta}f_X(x;\theta,\phi)\right)}_{\Delta} dx \tag{4}$$

Using the exponential family the derivative of the function in the exponential form $\Delta$ is

$$\begin{aligned}\frac{d}{d\theta}\left(f_X(x;\theta,\phi)\right) &= \frac{d}{d\theta}exp\left\{\frac{T(x)\theta - A(\theta)}{b(\phi)} + c(x,\phi)\right\}\\ &= \frac{1}{b(\phi)}\left(x - \frac{dA(\theta)}{d\theta}\right)exp\left\{\frac{T(x)\theta - A(\theta)}{b(\phi)} + c(x,\phi)\right\}\\ &= \frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\end{aligned} \tag{5}$$

So using this and (4):

$$\begin{aligned}0 &= \int_{\mathbf{R}} \Delta dx\\ &= \int_{\mathbf{R}}\left(\frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right)dx\end{aligned}$$

With $b(\phi) \neq 0$ and the above form, we finally obtain:

$$\begin{aligned}\underbrace{\int_{\mathbf{R}} x f_X(x;\theta,\phi)dx}_{E(X)} - \underbrace{\frac{dA(\theta)}{d\theta}}_{\star}\underbrace{\int_{\mathbf{R}} f_X(x;\theta,\phi)dx}_{=1(see(3))} &= 0\\ \Rightarrow \mathbf{E}(X) &= \frac{dA(\theta)}{d\theta}\end{aligned} \tag{6}$$

Proof of (2):
The procedure for the variance is equivalent. Using the exponential family and the result of the derivative in the Expectation (6), one has

$$\begin{aligned}&\frac{d^2}{d\theta^2}f_X(x;\theta,\phi)\\ &= \frac{d}{d\theta}\left(\frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right)\\ &= \frac{x}{b(\phi)}\underbrace{\frac{d}{d\theta}\left(f_X(x;\theta,\phi)\right)}_{(5)} - \frac{1}{b(\phi)}\underbrace{\frac{d}{d\theta}\left(\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right)}_{\star}\end{aligned} \tag{7}$$

Solving * :

$$\begin{aligned}&\frac{d}{d\theta}\left(\underbrace{\frac{dA(\theta)}{d\theta}}_{u}\underbrace{f_X(x;\theta,\phi)}_{v}\right)\\ &= \underbrace{\frac{d^2A(\theta)}{d\theta^2}}_{u'}\underbrace{f_X(x;\theta,\phi)}_{v} + \underbrace{\frac{dA(\theta)}{d\theta}}_{u}\underbrace{\frac{d}{d\theta}f_X(x;\theta,\phi)}_{v'=(5)}\\ &= \frac{d^2A(\theta)}{d\theta^2}f_X(x;\theta,\phi)\\ &+ \frac{dA(\theta)}{d\theta}\left(\frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right)\end{aligned} \tag{8}$$

Continuing (7) using (5) and * which is (8) one has

$$\begin{aligned}
&\frac{d^2}{d\theta^2}(f_X(x;\theta,\phi)) \\
&= \frac{x}{b(\phi)}\left(\frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right) \\
&- \frac{1}{b(\phi)}\left(\frac{d^2A(\theta)}{d\theta^2}f_X(x;\theta,\phi)\right) \\
&- \frac{1}{b(\phi)}\left(\frac{dA(\theta)}{d\theta}\left(\frac{x}{b(\phi)}f_X(x;\theta,\phi) - \frac{1}{b(\phi)}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi)\right)\right) \\
&= \frac{x^2}{b(\phi)^2}f_X(x;\theta,\phi) - \frac{x}{b(\phi)^2}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi) \\
&- \frac{1}{b(\phi)}\frac{d^2A(\theta)}{d\theta^2}f_X(x;\theta,\phi) - \frac{x}{b(\phi)^2}\frac{dA(\theta)}{d\theta}f_X(x;\theta,\phi) \\
&+ \frac{1}{b(\phi)^2}\Big(\frac{dA(\theta)}{d\theta}\Big)^2 f_X(x;\theta,\phi) \\
&= \frac{x^2 f_X(x;\theta,\phi)}{b(\phi)^2} - 2\frac{x f_X(x;\theta,\phi)}{b(\phi)^2}\frac{dA(\theta)}{d\theta} \\
&- \frac{f_X(x;\theta,\phi)}{b(\phi)}\frac{d^2A(\theta)}{d\theta^2} + \frac{f_X(x;\theta,\phi)}{b(\phi)^2}\Big(\frac{dA(\theta)}{d\theta}\Big)^2
\end{aligned}$$

Again, we make use of

$$0 = \int_{\mathbf{R}}\left(\frac{d^2}{d\theta^2}f_X(x;\theta,\phi)\right)dx \tag{9}$$

With $b(\phi) \neq 0$, multiplying the equation by $b(\phi)^2$ and using the fact that $f_X$ is a density

$$\begin{aligned}
0 &= \int_{\mathbf{R}}\Big(\frac{x^2 f_X(x;\theta,\phi)}{b(\phi)^2} - 2\frac{x f_X(x;\theta,\phi)}{b(\phi)^2}\frac{dA(\theta)}{d\theta} \\
&- \frac{f_X(x;\theta,\phi)}{b(\phi)}\frac{d^2A(\theta)}{d\theta^2} + \frac{f_X(x;\theta,\phi)}{b(\phi)^2}\Big(\frac{dA(\theta)}{d\theta}\Big)^2\Big)dx \\
&= \int_{\mathbf{R}} x^2 f_X(x;\theta,\phi)dx - 2\frac{dA(\theta)}{d\theta}\underbrace{\int_{\mathbf{R}} x f_X(x;\theta,\phi)dx}_{=\mathbf{E}(X)} \\
&- b(\phi)\frac{d^2A(\theta)}{d\theta^2}\underbrace{\int_{\mathbf{R}} f_X(x;\theta,\phi)dx}_{=1} \\
&+ \Big(\frac{dA(\theta)}{d\theta}\Big)^2\underbrace{\int_{\mathbf{R}} f_X(x;\theta,\phi)dx}_{=1} \\
&= \underbrace{\int_{\mathbf{R}} x^2 f_X(x;\theta,\phi)dx}_{=\mathbf{E}(X^2)} - 2\Big(\frac{dA(\theta)}{d\theta}\Big)^2 \\
&- b(\phi)\frac{d^2A(\theta)}{d\theta^2} + \Big(\frac{dA(\theta)}{d\theta}\Big)^2
\end{aligned} \tag{10}$$

Solving this for $\mathbf{E}(X^2)$ one gets

$$\begin{aligned}
\mathbf{E}(X^2) &= 2\Big(\frac{dA(\theta)}{d\theta}\Big)^2 + b(\phi)\frac{d^2A(\theta)}{d\theta^2} - \Big(\frac{dA(\theta)}{d\theta}\Big)^2 \\
&= \Big(\frac{dA(\theta)}{d\theta}\Big)^2 + b(\phi)\frac{d^2A(\theta)}{d\theta^2}
\end{aligned} \tag{11}$$

We know that

$$\begin{aligned}
Var(X) &= \mathbf{E}(X^2) - \mathbf{E}(X) \\
&= \Big(\frac{dA(\theta)}{d\theta}\Big)^2 + b(\phi)\frac{d^2A(\theta)}{d\theta^2} - \Big(\frac{dA(\theta)}{d\theta}\Big)^2 \\
&= b(\phi)\frac{d^2A(\theta)}{d\theta^2}
\end{aligned}$$

So, transferring this to the discrete case one has:

$$\mathbf{E}(T(K)) = \mathbf{E}(K) = A'(\eta) = \frac{d}{d\eta}A(\eta) = e^\eta = \lambda \tag{12}$$

$$\mathbf{V}(T(K)) = \mathbf{V}(K) = A''(\eta) = \frac{d^2}{d\eta^2}A(\eta) = e^\eta = \lambda \tag{13}$$

$$\delta(\eta) = \frac{k - A'(\eta)}{b(\phi)} = k - e^\eta \tag{14}$$

$$\delta'(\eta) = -\frac{A'(\eta)}{b(\phi)} = -e^\eta \tag{15}$$

### C. Score function and MLE

Let $k_1, ..., k_n$ be a realization of the random sample $K_1, ..., K_n$ from the population $\mathbf{K}$ of the model $K \sim P(\lambda)$. Rewriting the score function one has

$$\delta(\eta) = \sum_{i=1}^{n} k_i - ne^\eta \tag{16}$$

Solving the score equation for $\eta$ we get

$$\begin{aligned}
\delta(\eta) = 0 &\Leftrightarrow \sum_{i=1}^{n} k_i - ne^\eta = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} k_i = ne^\eta \\
&\Leftrightarrow \bar{k} = e^\eta \\
&\Leftrightarrow log(\bar{k}) = \eta
\end{aligned} \tag{17}$$

As

$$\delta'(\eta) = -ne^\eta < 0 \forall \eta \tag{18}$$

we can conclude that the maximum likelihood estimator of $\eta$ is $\hat{\eta}_{MLE} = log(\bar{k})$. Back transforming, using $n^{-1}(\lambda) = e^\lambda$ we conclude that the MLE of $\lambda$ is $\hat{\lambda}_{MLE} = \bar{k}$

### D. Properties of the Fisher Information

Regarding Fisher information, because we are in the exponential family, one has the following properties:

*Prop.*:

$$\begin{aligned} I(\theta, \mathbf{X}_n) &= -\mathbf{E}_\theta \left[ (\frac{d}{d\theta} l_{\mathbf{X}n,\theta}(\mathbf{X}_n))^2 \right] \\ &= -n\mathbf{E}_\theta \left[ (\frac{d^2}{d\theta^2} l_{\mathbf{X}n,\theta}(\mathbf{X}_1)) \right] \\ &= nI(\theta, \mathbf{X}_1) \end{aligned} \tag{19}$$

*Proof*:

$$\begin{aligned} I(\theta, \mathbf{X}_n) &= -\mathbf{E}_\theta \left[ (\frac{d^2}{d\theta^2} l_{\mathbf{X}n,\theta}(X_1, X_2, ..., X_n)) \right] \\ &= -\mathbf{E}_\theta \left[ (\frac{d^2}{d\theta^2} \sum_{i=1}^{n} log(f_{\mathbf{X},\theta}(X_k)) \right] \\ &= \sum_{i=1}^{n} \left( -\mathbf{E}_\theta \left[ (\frac{d^2}{d\theta^2} log(f_{\mathbf{X},\theta}(X_k)) \right] \right) \\ &= \sum_{i=1}^{n} \underbrace{\left( -\mathbf{E}_\theta \left[ (\frac{d^2}{d\theta^2} log(f_{\mathbf{X},\theta}(X_1)) \right] \right)}_{=I(\theta, \mathbf{X}_1)} \\ &= nI(\theta, \mathbf{X}_1) \end{aligned} \tag{20}$$

Using this insight, one has

$$I(\eta, \mathbf{X}_n) = n\mathbf{E}(A''(\eta)) = n\mathbf{E}(e^\eta) = ne^\eta \tag{21}$$

The back transformation gives us

$$I_n(k) = \frac{n}{A''(\eta)} = \frac{n}{e^\eta} \tag{22}$$

### *E. Variable selection*

- **p-value** The p-value relates to the probability of observing any value equal or larger than z. A small p-value indicates that it is unlikely to observe a relationship between the predictor and response (satellites) variable due to chance. Typically, a p-value of 5% or less is considered. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between the predictor and response variables.
- **Deviance** Deviance is a measure of goodness of fit. The Saturated Model is a model that considers each observation to have its own parameter, which translates into "n" parameters to estimate. The Null Model considers that there is one parameter for all observation. The Proposed Model considers that the observations can be estimated by parameters equal to the number of variables plus one for the intercept. We use the log likelihoods of those models to calculate Null and Residual Deviance. The Null deviance gives us information on how well the dependent variable is predicted by the model with the intercept only. The Residual Deviance gives information on how well the dependent variable is predicted when all independent variables are included in the model [3].

$$\begin{aligned} NullDeviance &= 2 \times (l_{sat} - l_{null}) \\ ResidualDeviance &= 2 \times (l_{sat} - l_{prop}) \end{aligned} \tag{23}$$

- **AIC and BIC**: AIC or Akaike Information Criterion and The Bayesian Information Criterion are measures of goodness of fit of an estimated statistical model and can be used for both model selection and evaluation. They are both based on the Deviance but penalize the model for the number of parameters and respectively for the sample size. So, in general we wish to achieve a relatively small AIC and BIC value and consider a model with a lower value as the better one. Let k be the number of parameters and n the sample size, then

$$\begin{aligned} AIC &= (-2) \times l + (2 \times k) \\ BIC &= (-2) \times l + (k \times ln(n)) \end{aligned} \tag{24}$$

- **Pearson Residuals**: $\forall j \in 1, ..., m : r_j^P = \frac{y-\mu}{\sqrt{Var(\mu}}$. As the distribution of the Pearson residual is usually skewed for non-Normal distributions as in our case of the Poisson distribution it may fail to have properties similar to those of a Normal-theory residual. Therefore, we do not consider it in our analysis.

We decided to examine the AIC value and the p-value for our variable selection. As our sample size is not so big, the p-value is still a good measure, as it tends to be very small when the data set is very large.

### *F. Poisson Regression Model*

The Poisson regression model modeling a counting variable Y, counting the number of times that a certain event occurs during a given time period. The model wants to explain this variable Y using explicative variables, $X = (X_1, X_2, ..., X_k)$.Where the model is given by:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = x_i^T \beta \tag{25}$$

Y has a Poisson distribution, $y_i \sim Poisson(\mu_i)$, i=1,...,N, and $\mu_i$ is the expected count of $y_i$, $E(Y) = \mu$.

In our case, the response variable is *Satellites*, which is the term used for crab males living close to the female.

Now we will use the R routine glm() to fit the data.

```
# Factorize color, and spine condition
crabs$color = factor(crabs$color,
                     levels =c(1,2,3,4,5))
crabs$spine_condition =
factor(crabs$spine_condition,
       levels = c(1,2,3))

fit <- glm(satellites ~ ., data = crabs,
family=poisson(link=log))
summary(fit)
stepAIC(fit)
```

The results were the following:

We will use the p-value (the column Pr(>|z|)) to determine the association between the response and each term in the model is statistically significant. The null hypothesis is that

TABLE I: Fit Poisson Regression Model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.35722 | 0.96700 | -0.369 | 0.71182 |
| color2 | -0.26491 | 0.16811 | -1.576 | 0.11507 |
| **color3** | **-0.51374** | **0.19536** | **-2.630** | **0.00855** |
| **color4** | **-0.53126** | **0.22692** | **-2.341** | **0.01922** |
| SC 2 | -0.15044 | 0.21358 | -0.704 | 0.48119 |
| SC 3 | 0.08742 | 0.11993 | 0.729 | 0.46604 |
| weight | 0.01651 | 0.04894 | 0.337 | 0.73582 |
| **CW** | **0.49712** | **0.16628** | **2.990** | **0.00279** |

TABLE II: AIC

| | Df | AIC |
|---|---|---|
| SC | 2 | 924.11 |
| weight | 1 | 927.93 |
| color | 3 | 918.66 |
| CW | 1 | 918.98 |

there is no association between the term and the variable and if the $p\_value < 0.05$ we reject the null hypothesis and we can conclude that there is a statistically significant association between the response variable and the term.

In our case, the variables that have a $p\_value < 0.05$ are *color3*, *color4* and *CW*. For the other variables we not reject $H_0$, we cannot conclude that there are a statistically significant association. Analyzing the AIC we can see that the *CW* and *color* variables are the ones that have a lower value, therefore we want to refit the model with these terms.

```
fit.2<- step(glm(satellites~., data= crabs,
                 family=poisson(link=log)))
summary(fit.2)
```

TABLE III: Fit Final Poisson Regression Model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -0.04961 | 0.23311 | -0.213 | 0.8315 |
| color2 | -0.20508 | 0.15371 | -1.334 | 0.1821 |
| color3 | -0.44966 | 0.17574 | -2.559 | 0.0105 |
| color4 | -0.45228 | 0.20843 | -2.170 | 0.0300 |
| CW | 0.54608 | 0.06809 | 8.020 | 1.06e-15 |

Confidence interval for coefficient (95%):

```
round(confint(fit.2),3)
```

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -0.509 | 0.406 |
| color2 | -0.496 | 0.108 |
| color3 | -0.788 | -0.098 |
| color4 | -0.863 | -0.044 |
| CW | 0.410 | 0.677 |

The prediction model is:

$$\begin{aligned} log(satellites) = &-0.205 \cdot color2 - 0.45 \cdot color3 \\ &- 0.452 \cdot color4 + 0.546 \cdot CW - 0.05 \end{aligned} \tag{26}$$

The standard error of the coefficient to measure the precision of the estimate of the coefficient, if the standard error is smaller, more precise is the estimate. We can observe that the variable *CW* was a low standard error, so is a good estimate. The *color* was a standard error higher than *CW*, but not have values very high, therefore these variables are good estimators for the model.

The estimated coefficient for *CW* is 0.546, this means a female crab with wider carapace will have more male satellites on an order of $e^{(0.546)} = 1.726$. For the *color* we have $e^{(-0.205)} = 0.815$, $e^{(0.45)} = 1.56$ and $e^{(0.452)} = 1.571$, to *medium light* , *medium*, *medium dark*, respectively, with this we can see there is an increase in the number of males when the color of the carapace is medium or medium dark.

### G. IRWLS

As outlines in the beginning for our model one has:

*1) link function:*

$$\eta(\lambda) = log(\lambda)$$

*2) Derivative of the link function:*

$$\eta'(\lambda) = \frac{d\eta(\lambda)}{d\lambda} = \frac{dlog(\lambda)}{d\lambda} = \frac{1}{\lambda}$$

*3) Inverse Link:*

$$\eta^{-1}(\lambda) = e^{\lambda}$$

*4) Weight and Inverse Weight:*

$$\begin{aligned} W^{-1} &= V\Big(\frac{d\eta(\lambda)}{d\lambda}\Big)^2 = V(\frac{dlog(\lambda)}{d\lambda})^2 \\ &= V(\frac{1}{\lambda})^2 = V(\frac{1}{\lambda^2}) = \lambda\frac{1}{\lambda^{\not 2}} = \frac{1}{\lambda} \end{aligned}$$

This makes sense, as we are using the canonical link and then the derivative of the link is the inverse of the variance, resulting in W = V = $\lambda$ .

*5) Working response Z:*

$$\mathcal{Z} = \eta + (y - \lambda)\frac{d}{d\lambda}g(\lambda) = \log(\lambda) + \frac{(y - \lambda)}{\lambda}$$

*6) Likelihood:*

$$L(\lambda;k) = \prod_{i=1}^{n} f(k_i|\lambda) = \prod_{i=1}^{n} exp\Big\{\frac{k_i\eta - e^{\eta}}{1} + (-log(k_i!))\Big\}$$
$$= \prod_{i=1}^{n} exp\Big\{k_i log(\lambda) - \lambda - log(k!)\Big\}$$
$$= exp\Big\{log(\lambda)\sum_{i=1}^{n} k_i - n\lambda - \sum_{i=1}^{n} log(k_i!)\Big\}$$

*7) Log-Likelihood:*

$$l(\lambda;k) = log(\lambda)\sum_{i=1}^{n} k_i - n\lambda - \sum_{i=1}^{n} log(k_i!)$$

*8) Deviance:* In the following the saturated model deviance is the one of a model where each observation has its own $\lambda_i$ and a MLE of $\hat{\lambda}_i = y_i$. In the estimated model the corresponding MLE is used. The change in deviance follows a $\chi^2$- distribution. The degrees of freedom equal the change in number of parameters in the models.

$$D = -2\sum_{i=1}^{n}\Big(k_i log(\lambda_i) - \lambda_i - log(k_i!)\Big)$$
$$-2\sum_{i=1}^{n}\Big(k_i log(k_i) - k_i - log(k_i!)\Big)$$
$$= 2\sum_{i=1}^{n}\Big(k_i(log(k_i) - log(\lambda_i)) - (k_i - \lambda_i))\Big)$$
$$= 2\sum_{i=1}^{n}\Big(k_i(log(\frac{k_i}{\lambda_i}) - (k_i - \lambda_i))\Big)$$

With this parameters calculated we can compute the Iteratively reweighted least squares estimation algorithm:

```
IRWLS <- function(x,y,tolerance,lev){
  # x         : predictor
  # y         : binary responce
  # tolerance: stopping criterion
  # lev       : level for the confidence
  #intervals
  Dev       = 0
  delta.Dev = 2*tolerance
  n         = length(x)
  mu <- rep(mean(y), n)
# initialize mu
  eta <- log(mu)
# initialize eta
  while ( abs(delta.Dev) > tolerance) {
    w <- mu
# weight = variance
    z <- eta + (y - mu)/(mu)
# working response
    mod <- lm(z ~ x, weights = w)
# weighted regression
    eta <- mod$fit
# linear predictor
    mu <- exp(eta)
# fitted value
    Dev.old    = Dev
    print(mu)
    Dev        = 2*sum(y*log(y/mu)-(y-mu))
    print(Dev)
    delta.Dev = Dev- Dev.old
    print(abs(delta.Dev))
  }
  model.coef= mod$coefficients
  model.se=sqrt(diag(summary(mod)
  $cov.unscaled))
  lower= mod$coefficients-
  qnorm(1-(1-lev)/2)*model.se
  upper= mod$coefficients+
  qnorm(1-(1-lev)/2)*model.se
  CI= cbind(lower,upper)
  Z= mod$coefficients/model.se
  pvalues=2*pnorm(abs(Z),lower.tail=FALSE)

  list(coeff=model.coef,se=model.se,
  default.glm.ConfInt=CI,
  z.stat=Z,p.values=pvalues)
}
```

Using this function and implement it to fit the best model from exercise 4. We are going to use it to predict the count values of the satellites based on the Color. We select the y as the variable with the least AIC from the exercice 4 and x as the satellites we want to predict (switching the x and y causes errors due to the y having zeros due to zero counts).

```
y=as.numeric(crabs$color)
x=as.numeric(crabs$satellites)
mymodel <- IRWLS(x,y,0.000001,0.95)
mymodel
#Output
$coeff
(Intercept)           x
 0.94998553 -0.02066593

$se
(Intercept)           x
 0.06548940  0.01608976

$default.glm.ConfInt
                  lower       upper
(Intercept)  0.82162866  1.07834240
x           -0.05220127  0.01086941

$z.stat
(Intercept)           x
  14.505943   -1.284415

$p.values
```

```
(Intercept)             x
1.111005e-47  1.989967e-01
```

Looking at the p-values we can assume that the model is well fitted to the satellites since $1.989967e-01 > 0.05 = \alpha$.

## II. Bayesian Inference

This section intends to apply Bayesian inference on the following data.

### A. Data

In a 2006 study published in The New England Journal of Medicine, 78 pairs of patients with Parkinson's disease were randomly assigned to receive treatment (which consisted of deep brain stimulation of a region of the brain affected by the disease) or control (which consisted of taking a prescription drug). The researchers found that in 50 of 78 pairs, the patients who received deep-brain stimulation had improved more than their partner in the control group. The parameter of interest is $\theta$, the probability of doing better on treatment than control.

### B. Modelling the data

Given that the parameter of interest, $\theta$, is the probability of doing better on treatment than on control, a natural distribution that arises is the Binomial, $Y \sim B(n, \theta)$. Where Y counts the number of pairs where the patient who received deep-brain stimulation improved their condition more than their partner's.

From the Binomial, we define the likelihood function $L(\theta \mid y)$ as:

$$\begin{aligned} L(\theta \mid y) &= \binom{n}{y} \theta^y (1-\theta)^{(n-y)} \\ &\propto \theta^y (1-\theta)^{(n-y)} \end{aligned} \tag{27}$$

Equation 27 can also be written in the form of a Beta distribution. If we take the following change of variables:

$$\begin{cases} \alpha = y + 1 \\ \beta = n - y + 1 \end{cases} \tag{28}$$

We get:

$$\begin{aligned} L(\theta \mid y) &\propto \theta^{(\alpha-1)} (1-\theta)^{(\beta-1)} \\ &\propto Beta(\alpha, \beta) \end{aligned} \tag{29}$$

The exercise asks us to first consider the prior $Beta(1, 1)$. This prior has two key advantages:

- It is equivalent to a standard uniform distribution, $U(0, 1)$, whose domain sits in the interval [0,1]. This makes sense because the parameter $\theta$ is a probability, and also because this prior gives equal weight to all values between [0,1];

- We derived above that the likelihood function is proportional to a Beta distribution. By choosing a Beta distribution as a prior, we know already that the posterior will be a Beta distribution as well, since Beta-Binomial is a known prior conjugate that generates a Beta as posterior.

The data is that in 50 out of the 78 pairs had the patient who received the treatment get better than their partner. To Formulate this into the Binomial parameters, this means that:

$$\begin{cases} y = 50 \\ n = 78 \end{cases} \tag{30}$$

In terms of the our likelihood expression which contains the new variables $\alpha$ and $\beta$, their values are:

$$\begin{cases} \alpha = 51 \\ \beta = 29 \end{cases} \tag{31}$$

Our likelihood is then:

$$\begin{aligned} L(\theta \mid y) &\propto p(y \mid \theta) \\ &\propto Beta(51, 29) \end{aligned} \tag{32}$$

The posterior is, from Bayes formula:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \cdot p(\theta) \\ &\propto Beta(51, 29) \cdot Beta(1, 1) \\ &\propto Beta(51, 29) \end{aligned} \tag{33}$$

Because the posterior is proportional to a Beta(51,29) which is itself a known distribution (integrates to 1 over the domain, which is $\theta \in [0, 1]$ ), we can say (instead of proportional to) that it is exactly equal to $Beta(51, 29)$. Therefore:

$$p(\theta \mid y) = Beta(51, 29) \tag{34}$$

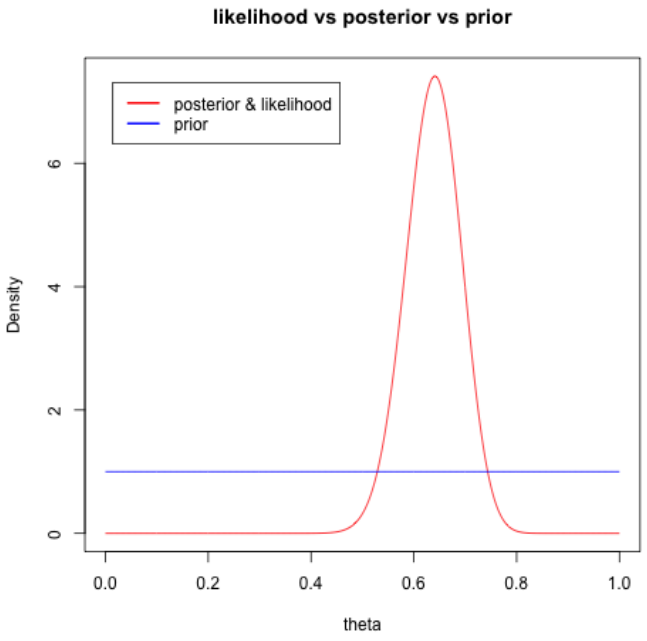



Fig. 5: prior & likelihood & posterior probability density functions

The posterior mean and standard deviations can be calculated by definition from the Beta distribution:

$$\begin{cases} E[\theta \mid y] = \int_0^1 \theta \cdot Beta(51,29)d\theta \underset{def.}{=} \frac{\alpha}{\alpha+\beta} = 0.6375 \\ Var[p(\theta \mid y)] \underset{def.}{=} \frac{\alpha \cdot \beta}{(\alpha+\beta)^2 \cdot (\alpha+\beta+1)} = 0.00285 \\ SD[p(\theta \mid y)] = 0.0534 \end{cases} \tag{35}$$

Fig. 5 contains the plots for the prior, likelihood and posterior distributions. From there we can visually see that the value obtained for the posterior mean makes sense.

### C. Credible intervals

Now that we have the posterior we can compute credible intervals. We will compute two of these intervals: the symmetric (or central) and the highest posterior density (HPD).

For the centered credible interval, a 95% credible interval (in other words $(1-\alpha)\%$ with $\alpha = 5\%$) for the value of $\theta$ involves finding the 0.025 and 0.975 quantiles of the $Beta(51, 29)$ distribution. It's called central because it consists in leaving out the remaining $\alpha\%$ area equally to both sides of the interval $(\alpha/2)\%$ to the left and $(\alpha/2)\%$ to the right. Using R's *qbeta* these quantiles can easily be found and we end up with the following 95% credible interval for $\theta$:

$$CCI_{95\%}(\theta) \sim [0.530, 0.739] \tag{36}$$

In an HPD credible interval, however, any point within the interval has to have a higher density than any other point outside. This is the main difference with the central interval. Both these intervals will match whenever the distribution is symmetric and unimodal, but if the posterior is skewed then the central interval will contain points within it which will have a lower density than some points outside of it.
An HPD credible interval is, by definition, the region defined by:

$$H_\tau = \{\theta; p(\theta \mid x) > \tau\} \tag{37}$$

In order to find such interval we need to solve the equation:

$$p(\theta \mid y) = \tau \tag{38}$$

This only works for posteriors which are uni-modal and the parameter is uni-dimensional. Both conditions are satisfied in our scenario.

However we usually want to satisfy a coverage, which means we would need to solve the following equation:

$$\int_{\{\theta; p(\theta|y) > \tau\}} p(\theta \mid y)d\theta = \alpha \tag{39}$$

The first step to compute the interval (starting from eq. 37) is then solving the following equation:

$$\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)} = \tau \tag{40}$$

But this equation doesn't have an analytical solution, therefore we have to resort to a numerical solution. We will use the R package ***HDInterval*** to compute this HDP credible interval. We now present the HPD interval obtained:

$$HPD_{95\%}(\theta) \sim [0.532, 0.741] \tag{41}$$

Comparing both intervals, we notice that the HPD has a slight "shift" to the right. The Beta(51,29) distribution (which can be visualized in fig. 5) is not symmetric and its mass is also to the right side, which explains the interval shift compared to the central interval.

Besides computing credible intervals, with the posterior calculated we can also answer questions of the type: "What's the probability that $\theta$ is in the interval C?". To compute these probabilities we have to integrate the posterior over the interval. Therefore, if we want to know the probability of $\theta$ being over 0.5 for example:

$$P(\theta > 0.5) = \int_{0.5}^{1} beta(51, 29) \sim 0.994 \tag{42}$$

This result makes sense because 50 out of the 78 pairs had the positive characteristic, therefore it should be highly probable that the $\theta$ parameter of the Binomial(n,$\theta$) is over 0.5.

### D. Jeffreys prior

We will now recompute the results so far but starting with the Jeffrey's prior. By definition, the Jeffrey's prior, $\pi_J(\theta)$, is computed with Eq. 43:

$$\pi_J(\theta) \propto I(\theta)^{-\frac{1}{2}} \tag{43}$$

where $I(\theta)$ is the Fisher Information given by:

$$I(\theta) = -E_\theta\Big[\frac{d^2 logp(X \mid \theta)}{d\theta^2}\Big] \tag{44}$$

Given that we are modelling our problem with a Binomial distribution, the likelihood function, $p(x|\theta)$, is:

$$p(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{45}$$

The log likelihood is then:

$$log\binom{n}{x} \cdot xlog(\theta) + (n-x)log(1-\theta) \tag{46}$$

The first derivative of the log likelihood (ignoring terms that don't depend on theta):

$$\frac{d}{d\theta} logp(x \mid \theta) \propto \frac{x}{\theta} - \frac{n-x}{1-\theta} \tag{47}$$

And the second derivative:

$$\frac{d^2 2}{d\theta^2} logp(x \mid \theta) \propto -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \tag{48}$$

Now to compute the final step which is the expectation of eq. 48 we note two things:

- What's random is the data
- If X $\sim Bin(n,\theta)$, then: $E[X] = n\theta$

Therefore, we get:

$$\begin{aligned} I(\theta) &= -E_\theta\Big[\frac{d^2 logp(X \mid \theta)}{d\theta^2}\Big] \\ &= -\frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} \\ &= \frac{n}{\theta} + \frac{n}{1-\theta} \\ &= \frac{n}{\theta(1-\theta)} \end{aligned} \tag{49}$$

With this, the jefreeys prior, $\pi_J(\theta)$, is:

$$\pi_J(\theta) \propto I(\theta)^{\frac{1}{2}} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \tag{50}$$

Eq. 50 is the form of a $Beta(\frac{1}{2}, \frac{1}{2})$ density. In fig. 6 we have the plot of the previous prior and the current Jeffrey's prior.

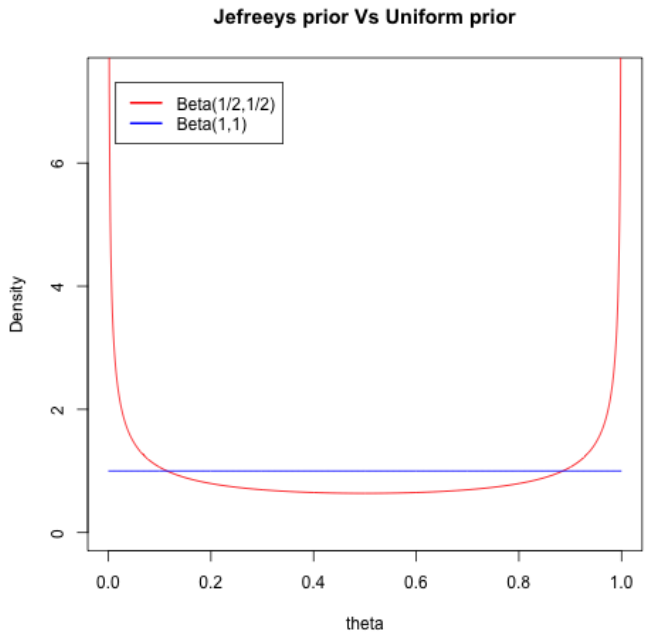



Fig. 6: Jeffrey's prior - Beta(1/2,1/2), and Uniform prior comparison

From fig. 6 we can see that the jeffrey's prior gives a lot more weight to the extreme values and less weight to the middle of the domain. We can further interpret this as the following: if new data comes up and suggests that $\theta$ is somewhere around 0.5, it can very likely come from a variety of $\theta$'s around 0.5. The information from that wouldn't help us as much distinguish where $\theta$ is actually. However, if the data we get suggests $\theta$ is close to the extremes, then it provides more information because is more likely that $\theta$ is very near the extremes ( because it couldn't have likely come from many places further from the extreme).

We are now set to compute the previous results:

1) The posterior and its basic statistics (mean and standard deviation);
2) HPD and central intervals;
3) plot 3 distributions: prior, posterior and likelihood;
4) Compute p($\theta > 0.5$).

Starting with the posterior, and taking from eq. 33 we have the new posterior as:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \cdot p(\theta) \\ &\propto Beta(51, 29) \cdot Beta\Big(\frac{1}{2}, \frac{1}{2}\Big) \\ &\propto \frac{x^{50} \cdot (1-x)^{28}}{\frac{\Gamma(51)\cdot\Gamma(29)}{\Gamma(80)}} \cdot \frac{x^{-0.5} \cdot (1-x)^{-0.5}}{\frac{\Gamma(0.5)\cdot\Gamma(0.5)}{\Gamma(1)}} \\ &\propto x^{49.5} \cdot (1-x)^{27.5} \end{aligned} \tag{51}$$

From eq. 51 we conclude that the jeffrey's posteriori is the Beta(50.5,28.5) distribution. And because the beta distribution integrates to 1 we can state that:

$$p(\theta \mid y) = Beta(50.5, 28.5) \tag{52}$$

The posterior mean and standard deviations can be calculated by definition from the Beta distribution:

$$\begin{cases} E[\theta \mid y] \underset{def.}{=} \frac{\alpha}{\alpha+\beta} = \frac{50.5}{50.5+28.5} = 0.6392 \\ Var[p(\theta \mid y)] \underset{def.}{=} \frac{\alpha\cdot\beta}{(\alpha+\beta)^2\cdot(\alpha+\beta+1} = 0.00288 \\ SD[p(\theta \mid y)] = 0.0537 \end{cases} \tag{53}$$

Now to compute the central interval, we compute again the 0.025 and 0.0975 quantiles of the Beta(50.5,28.5) distribution:

$$CCI_{95\%} \sim [0.531, 0.741] \tag{54}$$

And the HPD interval is:

$$HPD_{95\%}(\theta) \sim [0.533, 0.743] \tag{55}$$

Computing $P(\theta > 0.5)$ on the jeffrey's posterior this time:

$$P(\theta > 0.5) = \int_{0.5}^{1} beta(50.5, 28.5) \sim 0.9936 \tag{56}$$

Fig. also shows us the prior, likelihood and posterior plots. We can see that the posterior is practically on top of the likelihood with a very slight shift to the right.

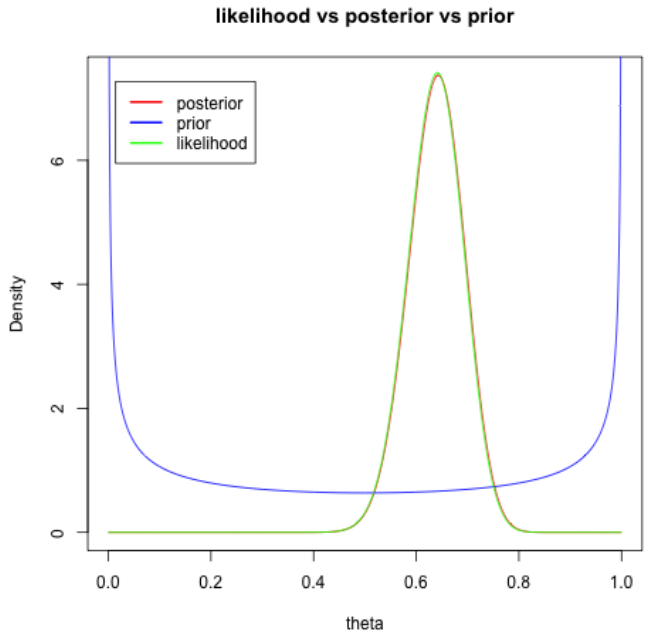



Fig. 7: Jefreeys prior & likelihood & posterior probability density functions

To compare both priors we present two tables below. We can see that the results practically haven't changed from using the Beta(1,1) as prior.

| Prior\Post stats | Expect. | Variance | SD | P($\theta$¿0.5) |
|---|---|---|---|---|
| Uniform prior | 0.6375 | 0.00285 | 0.0534 | 0.994 |
| Jeffreys prior | 0.6392 | 0.00288 | 0.0537 | 0.994 |

| Prior\Post stats | Centered CI | HPD CI |
|---|---|---|
| Uniform prior | [0.530, 0.739] | [0.532, 0.741] |
| Jeffreys prior | [0.531, 0.741] | [0.533, 0.743] |

### *E. Predictive distribution*

Now we are considering that 10 more pairs of patients have undergone the same treatment. Two questions arise for this situation:

1) Based on our posterior, what is the probability that 6 or more pairs will do better on treatment than control?
2) What is the expected number of patients that will do better on treatment than control?

To answer the first question we first have to define the posterior predictive distribution. This distribution is defined in eq. 57:

$$P(y^* \mid y) = \int_\Theta f(y^*, \theta|y) d\theta \tag{57}$$

However, from the conditional probability formula we can also express eq. 57 as:

$$\begin{aligned} P(y^* \mid y) &= \int_\Theta f(y^*, \theta \mid y) d\theta \\ &= p(y^* \mid \theta, y) \cdot p(\theta \mid y) \end{aligned} \tag{58}$$

And if we consider the new data to be independent of the observed data:

$$\begin{aligned} P(y^* \mid y) &= \int_\Theta p(y^* \mid \theta, y) \cdot p(\theta \mid y) \\ &= \int_\Theta p(y^* \mid \theta) \cdot p(\theta \mid y) \end{aligned} \tag{59}$$

The second term is the posterior on the initial data (which we have already), and the first term is the likelihood on the new data. This is a known integral which results in a Beta-Binomial$(n, \alpha^*, \beta^*)$ distribution. This distribution is essentially the binomial distribution on a fixed sample size of n and the probability of success of each trial is also fixed but drawn from a Beta$(\alpha^*, \beta^*)$ distribution. Using the posterior obtained from the uniform prior, the posterior predictive distribution then becomes:

$$\begin{aligned} P(y^* \mid y) &= BetaBinomial(n, \alpha^*, \beta^*) \\ &= BetaBinomial(10, 51, 29) \end{aligned} \tag{60}$$

We now want to compute the probability that 6 or more pairs of patients experience improvement with the treatment. Considering the BetaBinomial is a discrete distribution, we will need to compute 1 - P(X ¡ 6). This can be achieved with R's *extraDistr* package. The result is presented below:

$$p(\text{'At least 6 pairs improve'} \mid y) = 0.714 \tag{61}$$

We can now address the second question of this exercise. The expected value of a BetaBinomial distribution is known and given by eq. 62. So, the expected number of patients that will do better on treatment than in control is 6.375.

$$E[X \sim BetaBinom(n, \alpha, \beta)] = \frac{n \cdot \alpha}{\alpha + \beta} \tag{62}$$

### *F. Treatment effectiveness conclusion*

The 95% HPD interval contains $\theta$ within the interval [0.532,0.741]. This means that theta is very likely between those values. In other words, we expect around 63% of the patients to improve more under the treatment than in the control group. Dismissing the fact that we have no information about side effects of the treatment, we can conclude that the treatment indeed works with expected positive results on 6 out of every 10 patients.

## REFERENCES

[1] Ye Guan, Xinyue Yang, Jieni Wan horseshoe crabs and satellites. https://jbhender.github.io/Stats506/F17/Projects/Poisson_Regression.html. Accessed: 2019-12-17.

[2] H Jane Brockmann. Satellite male groups in horseshoe crabs, limulus polyphemus. *Ethology*, 102(1):1–21, 1996.

[3] Peter McCullagh. *Generalized linear models*. Routledge, 2019.