IMS Inleveropdracht 2

2023-10-19

Introduction

The goal of this statistical experiment is to determine whether the gas extraction from the Groningen gas field has some influence on the number of earthquakes in the Netherlands. This is accomplished by finding two model for a data set containing all earthquakes with a magnitude of 3.0 or larger on the Richter scale, where earthquakes that happen within 3 days of each other count as one. One model will contain all the data from the years before the gas extraction (1900-1962), the other model will contain all the data from the years after (1963-2022). The goal is to find a ML-estimator for both models, where the same family of distribution is used for both models. After that a mean squared error will be computed, and finally a confidence interval will be given.

Organizing the data

The available data is first organised so it can be used to calculate with.

```
#organising the data
load("Earthquakes2.Rdata") #loading the data
sep_data <- stack(Data) #seperates the data from our data file
all_values <- sep_data$values #seperates just the values into a vector
prior_gas_values <- all_values[1:63] #data from the years prior to gas extraction
post_gas_values <- all_values[64:123] #data from the years after gas extraction began
```

The point graphs

A point graph for both data sets with the x-axis being all the possible values (in our case 0-7) and the y-axis being the density of every point is created to visually establish what family of distributions could be used.

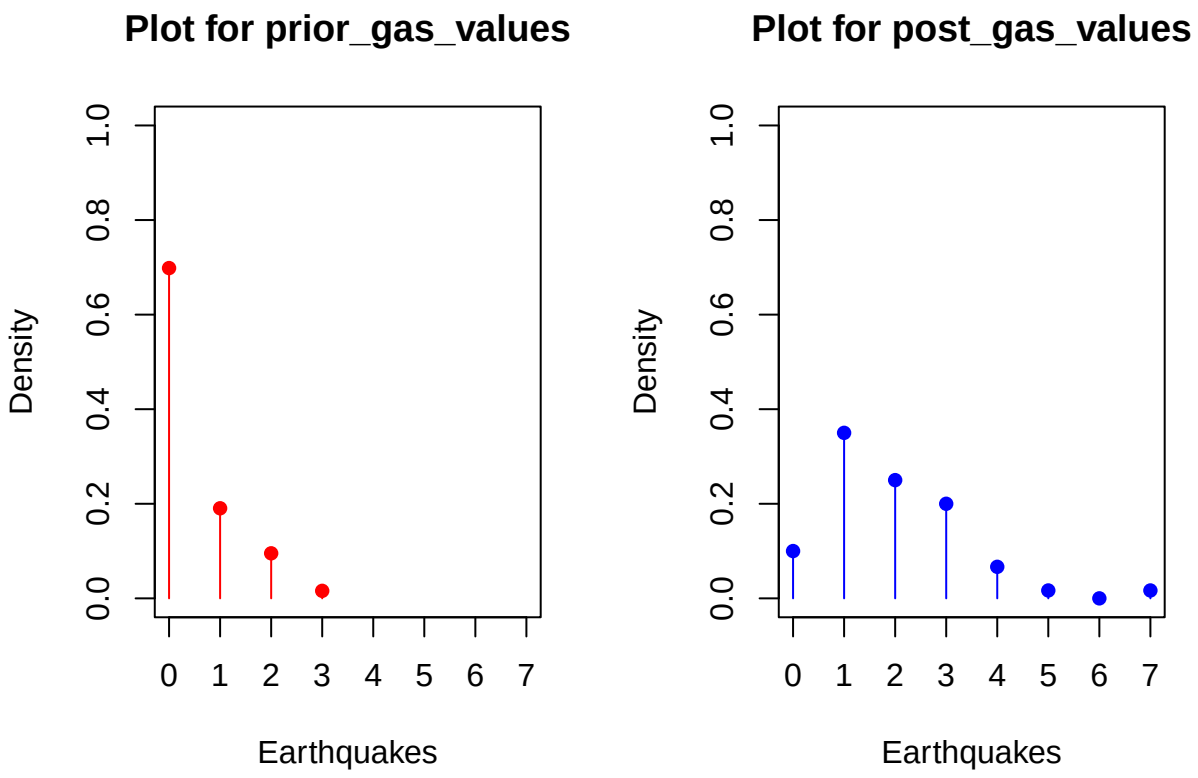
```
#plotting the data
par(mfrow=c(1, 2))
#First we need to know some information, like the maximum amount of earthquakes to plot
max_amount = max(prior_gas_values, post_gas_values)

#we generate a table from the prior_gas_values vector
prior_freqs_table <-
  table(factor(prior_gas_values,
               levels = 0:max(prior_gas_values)))/length(prior_gas_values)
prior_freqs <- stack(prior_freqs_table)$values #this will be our y-axis vector
prior_x <- c(0:(length(prior_freqs)-1)) #this will be our x-axis vector

#we generate a table from the post_gas_values vector
post_freqs_table <-
  table(factor(post_gas_values,
               levels = 0:max(post_gas_values)))/length(post_gas_values)
post_freqs <- stack(post_freqs_table)$values #this will be our y-axis vector
post_x <- c(0:(length(post_freqs)-1)) #this will be our x-axis vector
```

```r
# We create a blank plot
plot(x = 1,
     type = "n",
     xlim = c(0, max_amount),
     ylim = c(0, 1),
     pch = 16,
     xlab = "Earthquakes",
     ylab = "Density",
     main = "Plot for prior_gas_values")
#Now we plot the points over it
points(prior_x, prior_freqs, pch = 16, lwd = 1, col = "red")
segments(x0 = prior_x, y0 = 0, x1 = prior_x, y1 = prior_freqs, col = "red")

#We create another blank plot
plot(x = 1,
     type = "n",
     xlim = c(0, max_amount),
     ylim = c(0, 1),
     pch = 16,
     xlab = "Earthquakes",
     ylab = "Density",
     main = "Plot for post_gas_values")
#And we plot the points over it
points(post_x, post_freqs, pch = 16, lwd = 1, col = "blue")
segments(x0 = post_x, y0 = 0, x1 = post_x, y1 = post_freqs, col = "blue")
```



The chosen distribution

The Poisson distribution is chosen. First of all, the data has values ranging from the integers ranging from 0 to 3 and 0 to 7 for pre- and post gas extraction respectively. Because no values in between these integers can be taken, it follows that a discrete distribution is an accurate choice. The Poisson distribution is chosen because it can take the shape of an exponential declining distribution, like our prior_gas_values graph, but it can also take the shape of a fast incline with a slower decline, like the graph of the prior_gas_values data set.

Determining the ML-estimators

For the computation of the ML-estimators the density function of the Poisson distribution

$$f_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

is used. The log-likelihood method is used, where the log-likelihood is maximised by setting it's derivative to 0 and checking the second derivative. First, the log-likelihood is simplified:

$$\begin{aligned} l_{X_1,\ldots,X_{63}}(\lambda) &= \sum_{i=1}^{63} \ln\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right) \\ &= \sum_{i=1}^{63} \ln\left(\lambda^{X_i}\right) + \sum_{i=1}^{63} \ln\left(e^{-\lambda}\right) + \sum_{i=1}^{63} \ln\left(\frac{1}{X_i!}\right) \\ &= \ln(\lambda) \sum_{i=1}^{63} (X_i) - 63\lambda - \sum_{i=1}^{63} \ln(X_i!)\,. \end{aligned}$$

Now that a form that is easy to derive has been obtained deriving with respect to λ obtains

$$\begin{aligned} \frac{\partial l_{X_1,\ldots,X_{63}}(\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{63}(X_i)}{\lambda} - 63 &\overset{!}{=} 0 \\ \frac{\sum_{i=1}^{63}(X_i)}{\lambda} &\overset{!}{=} 63 \\ \frac{\sum_{i=1}^{63}(X_i)}{63} = \overline{X} &\overset{!}{=} \lambda, \end{aligned}$$

the second derivative is calculated to check if this is indeed the maximum:

$$\begin{aligned} \frac{\partial^2 l_{X_1,\ldots,X_{63}}(\lambda)}{\partial \lambda^2} &= \frac{\partial}{\partial \lambda}\left[\ln(\lambda) \sum_{i=1}^{63} (X_i) - 63\lambda - \sum_{i=1}^{63} \ln(X_i!)\right] \\ &= -\frac{1}{\lambda^2} \sum_{i=1}^{63} (X_i) < 0. \end{aligned}$$

Observe that the second derivative is always negative. Therefore the function is concave, meaning that it has only one local maximum which is also the global maximum. This means that $\lambda = \overline{X}$ is an ML-estimator. The mean is calculated to obtain an ML-estimator for the parameter λ. The mean for both data sets is computed with R:

```
mean_prior <- mean(prior_gas_values)
mean_post <- mean(post_gas_values)
#printing the means to use in our calculations
print(mean_prior)
```

```
## [1] 0.4285714
```

```r
print(mean_post)
```

```
## [1] 1.916667
```

So the ML-estimator λ for the prior gas values data set is equal to 0.4285714, and the ML-estimator λ for the post gas values data set is equal to 1.916667.

Updating the point graphs

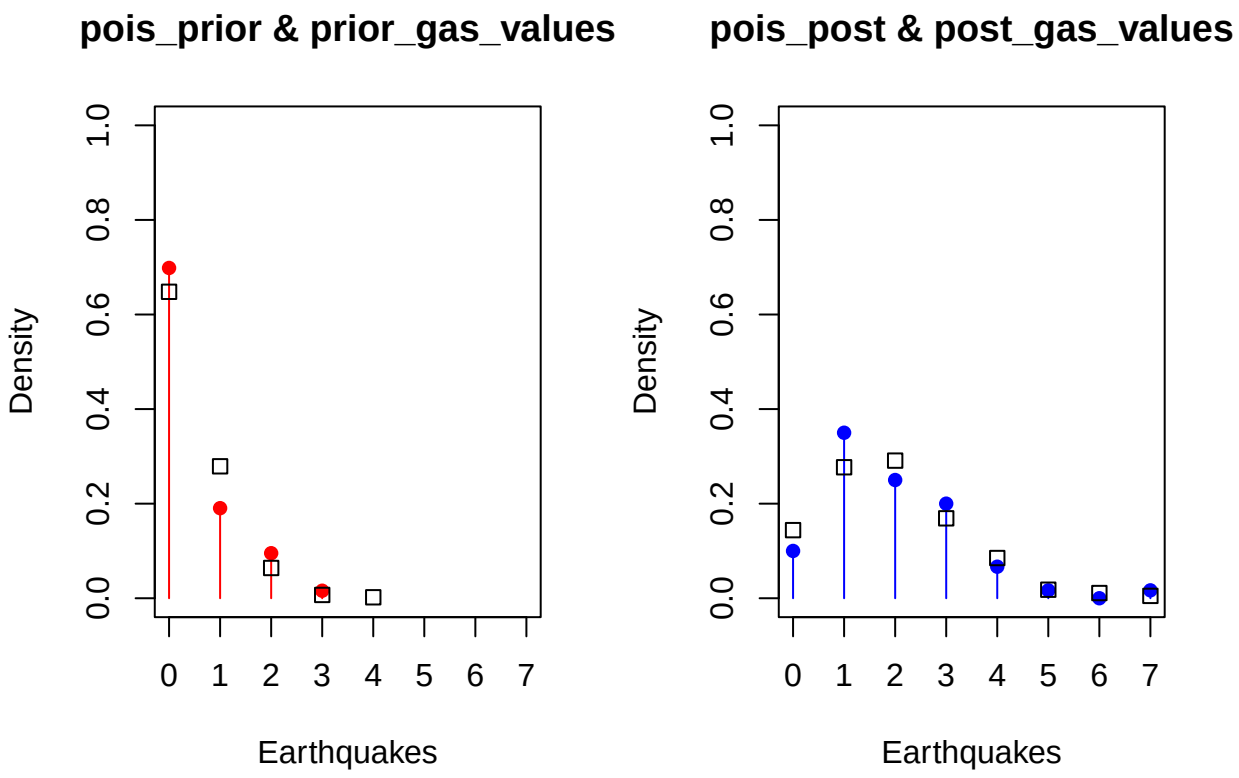
The point graphs can be updated to include our poisson distributions with the estimated parameters.

```r
par(mfrow=c(1, 2))
#we plot the Poisson distribution over the previous plots.
#First we establish the points for both Poisson distributions
pois_prior_values <- rpois(1000, 0.4285714)
pois_prior_freqs_table <-
  table(factor(pois_prior_values,
               levels = 0:max(pois_prior_values)))/length(pois_prior_values)
pois_prior_freqs <- stack(pois_prior_freqs_table)$values
pois_prior_x <- c(0:(length(pois_prior_freqs)-1))

pois_post_values <- rpois(1000, 1.916667)
pois_post_freqs_table <-
  table(factor(pois_post_values,
               levels = 0:max(pois_post_values)))/length(pois_post_values)
pois_post_freqs <- stack(pois_post_freqs_table)$values
pois_post_x <- c(0:(length(pois_post_freqs)-1))

#Create another blank plot for our prior values and the Poisson values
plot(x = 1,
     type = "n",
     xlim = c(0, max_amount),
     ylim = c(0, 1),
     pch = 16,
     xlab = "Earthquakes",
     ylab = "Density",
     main = "pois_prior & prior_gas_values")
points(prior_x, prior_freqs, pch = 16, lwd = 1, col = "red")
segments(x0 = prior_x, y0 = 0, x1 = prior_x, y1 = prior_freqs, col = "red")
points(pois_prior_x, pois_prior_freqs, pch = 0, lwd = 1, col = "black")

#Create another blank plot for our post values and the Poisson values
plot(x = 1,
     type = "n",
     xlim = c(0, max_amount),
     ylim = c(0, 1),
     pch = 16,
     xlab = "Earthquakes",
     ylab = "Density",
     main = "pois_post & post_gas_values")
points(post_x, post_freqs, pch = 16, lwd = 1, col = "blue")
segments(x0 = post_x, y0 = 0, x1 = post_x, y1 = post_freqs, col = "blue")
points(pois_post_x, pois_post_freqs, pch = 0, lwd = 1, col = "black")
```



Computing the mean squared errors

The difference between the prior frequencies and the poisson prior frequencies is squared for all available points. Then the mean is taken which results in the mean squared error for the model. The same is done for the post gas extraction data.

```
difference_prior <- ( prior_freqs - pois_prior_freqs[1:length(prior_freqs)] )^2
mse_prior <- mean(difference_prior)
print(mse_prior)
```

```
## [1] 0.002858114
```

```
difference_post <- ( post_freqs - pois_post_freqs[1:length(post_freqs)] )^2
mse_post <- mean(difference_post)
print(mse_post)
```

```
## [1] 0.00131275
```

Calculating the confidence intervals

In this section the two-sided asymptotic 0.95 confidence interval for both of our parameters is calculated.

To derive confidence intervals for our calculated λ's, the central limit theorem is used. The following holds

$$\sqrt{n}\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{\mathcal{D}} N(0,1).$$

This leads to the following inequality

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\lambda}} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

This inequality cannot be easily solved for λ. To make this process easier, an estimator for $\sqrt{\lambda}$ is calculated. This can be done by calculating the log-likelihood estimator for $\sqrt{\lambda}$. Define $\mu = \sqrt{\lambda}$, then it follows to find the log-likelihood estimator for μ. The corresponding probability density function is given by

$$f_\lambda(x) = \frac{\mu^{2n}}{x!} e^{-\mu^2}.$$

The log-likelihood is then given by

$$\begin{aligned} l_{X_1,\ldots,X_{63}}(\mu) &= \sum_{i=1}^{63} \ln\left(\frac{\mu^{2X_i} e^{-\mu^2}}{X_i!}\right) \\ &= \sum_{i=1}^{63} \left(2X_i - \mu^2 - \ln(X_i!)\right) \\ &= 2\ln(\mu) \sum_{i=1}^{63} X_i - 63\mu^2 - \sum_{i=1}^{63} \ln(X_i!). \end{aligned}$$

The log-likelihood estimator for μ can be found by setting the derivative of this expression to zero as follows:

$$\begin{aligned} \frac{\partial l_{x_1,\ldots,X_{63}}(\mu)}{\partial \mu} = \frac{2}{\mu} \sum_{i=1}^{63} X_i - 126\mu &\stackrel{!}{=} 0 \\ \frac{2}{126} \sum_{i=1}^{63} X_i &\stackrel{!}{=} \mu^2 \\ \mu &= \sqrt{\overline{X}}. \end{aligned}$$

To determine if this gives the maximum, the second derivative is calculated.

$$\frac{\partial^2 l_{X_1,\ldots,X_{63}}(\mu)}{\partial \mu^2} = -\frac{2}{\mu^2} \sum_{i=1}^{63} X_i - 126 < 0.$$

This is always negative. Therefore, $l_{X_1,\ldots,X_{63}}(\mu)$ is concave, meaning it has only one local maximum which is also the global maximum.

Therefore, $\mu = \sqrt{\overline{X}}$ is the corresponding log-likelihood estimator for μ, and thus $\sqrt{\lambda} = \sqrt{\overline{X}}$ holds.

With this estimator, the inequality can be solved for λ. To do this, the interval is split. Each inequality is solved individually for λ. Using $n = 63$ as the size and replacing $\sqrt{\lambda}$ with $\sqrt{\overline{X}}$, it follows that

$$\begin{aligned} \Phi^{-1}\left(\frac{\alpha}{2}\right) &\leq \sqrt{63} \frac{\overline{X} - \lambda}{\sqrt{\overline{X}}} \\ \frac{\sqrt{\overline{X}}}{\sqrt{63}} \Phi^{-1}\left(\frac{\alpha}{2}\right) &\leq \overline{X} - \lambda \\ \lambda &\leq \overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}} \Phi^{-1}\left(\frac{\alpha}{2}\right) \end{aligned}$$

and

$$\begin{aligned} \sqrt{63} \frac{\overline{X} - \lambda}{\sqrt{\overline{X}}} &\leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ \overline{X} - \lambda &\leq \frac{\sqrt{\overline{X}}}{\sqrt{63}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ \overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) &\leq \lambda. \end{aligned}$$

The inequality then becomes

$$\overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \leq \lambda \leq \overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}}\Phi^{-1}\left(\frac{\alpha}{2}\right),$$

which corresponds to the following interval:

$$\left(\overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \overline{X} - \frac{\sqrt{\overline{X}}}{\sqrt{63}}\Phi^{-1}\left(\frac{\alpha}{2}\right)\right).$$

This interval is now applied to the two data sets.

```
#We compute the confidence intervals for both data sets.
#prior:
prior_left <- (mean_prior - (sqrt(mean_prior)/sqrt(63))*qnorm((1-(0.05/2)),0,1))
prior_right <- (mean_prior - (sqrt(mean_prior)/sqrt(63))*qnorm((0.05/2),0,1))
#post
post_left <- (mean_post - (sqrt(mean_post)/sqrt(63))*qnorm((1-(0.05/2)),0,1))
post_right <- (mean_post - (sqrt(mean_post)/sqrt(63))*qnorm((0.05/2),0,1))
#output
print(paste("CI for prior gas values: (", prior_left, ",", prior_right, ")"))
```

```
## [1] "CI for prior gas values: ( 0.266916323703404 , 0.590226533439453 )"
```

```
print(paste("CI for post gas values: (", post_left, ",", post_right, ")"))
```

```
## [1] "CI for post gas values: ( 1.5748044445121 , 2.25852888882124 )"
```

Final conclusion

With the creation of these models, it can be concluded that the amount of earthquakes after 1962, when gas extraction began in Groningen, has increased. This is based on multiple observations, starting with the fact that the calculated confidence intervals don't overlap. The fact that these intervals are disjoint means that the two data sets can be seen as independent, meaning that there should be a reason for the increase (which is likely because of the gas extraction that began in 1963). This increase can also be observed by looking at the means. There is an increase of more than 400% from the data prior to 1963 and after. Also, the comparison to the Poisson distributions (which has low mean square errors, so they are quite accurate) yielded in a higher parameter for the data set from after 1962. This parameter is the mean of the Poisson distribution, which means that an increase can also be seen with this method.