IMS Homework assignment 1

Gijs Turkenburg (s3573532) and Sam Noordam (s3680657)

2023-09-28

Introduction

In this document we will establish a suitable distribution to estimate the amount of rainfall in the netherlands. We will use the provided datasat "maxrain.Rdata", which contains the maximal hourly precipitation amount (in mm) measured within one year in De Bilt between 1906 and 2022. We will plot this data and try to find a distribution, where the parameters are estimated with the moment estimator. We will then plot our found distribution over our data to see if it matches, and based on the chosen distribution we will calculate the probability that a sewage system has to cope with more than 300 rain in one hour. We will finally compare this value with the empirical probability of the data set.

Ordering the data set

Because our dataset has a structure we can't work with, we extract just the values of rainfall from every year:

```
#load the data into the program
load("maxrain.Rdata")
#seperate the rain into a parameter years and values, where we will use values
data_splitser <- stack(maxrain) #contains seperated data
values <- data_splitser$values #extracts the values
years <- data_splitser$ind #extracts the years
values <- na.omit(values) #finally we omit any incomplete cases, just in case
typeof(values) #values has a "double" type
```

```
## [1] "double"
```

```
is.vector(values) #values is indeed a vector, like we want it to be.
```

```
## [1] TRUE
```

```
length(values) #we later want to know how much data points we have
```

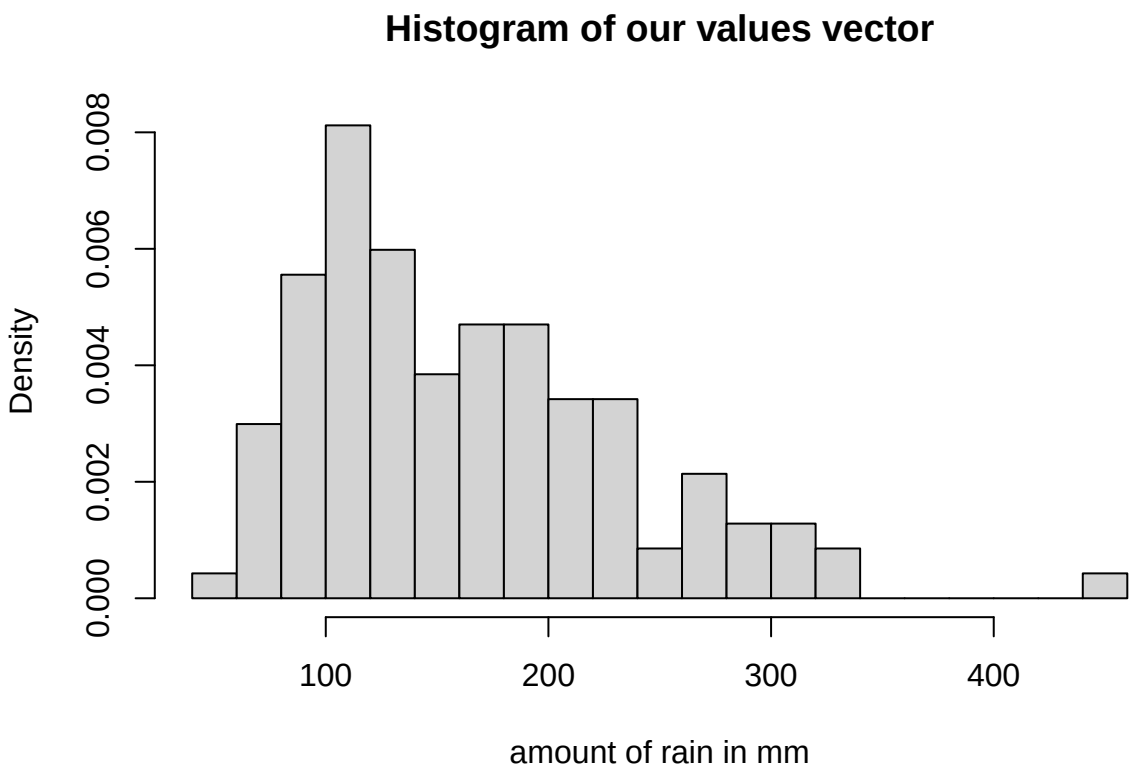
```
## [1] 117
```

We now have a vector "values" containing all the values we will use.

Designing our histogram

We will now compute a histogram with this values vector.

```
#we compute our first histogram to see what the data looks like
hist(values, breaks = 20, freq = FALSE, main = "Histogram of our values vector"
     ,xlab = "amount of rain in mm") #we choose 20 breaks arbitrarily
```



We can see a couple of things:

1. We have a (couple of) datapoint(s) that is really far from the rest, we would like to remove this outlier(s) to get a more accurate histogram,
2. The shape is not symmetrical, it seems to resemble a sort of wave: going up fast, peaking at around 120, and subsiding slowly,
3. The amount of bars we choose seems about right.

We will first remove the outlier(s) by sorting our data and removing the highest value(s)

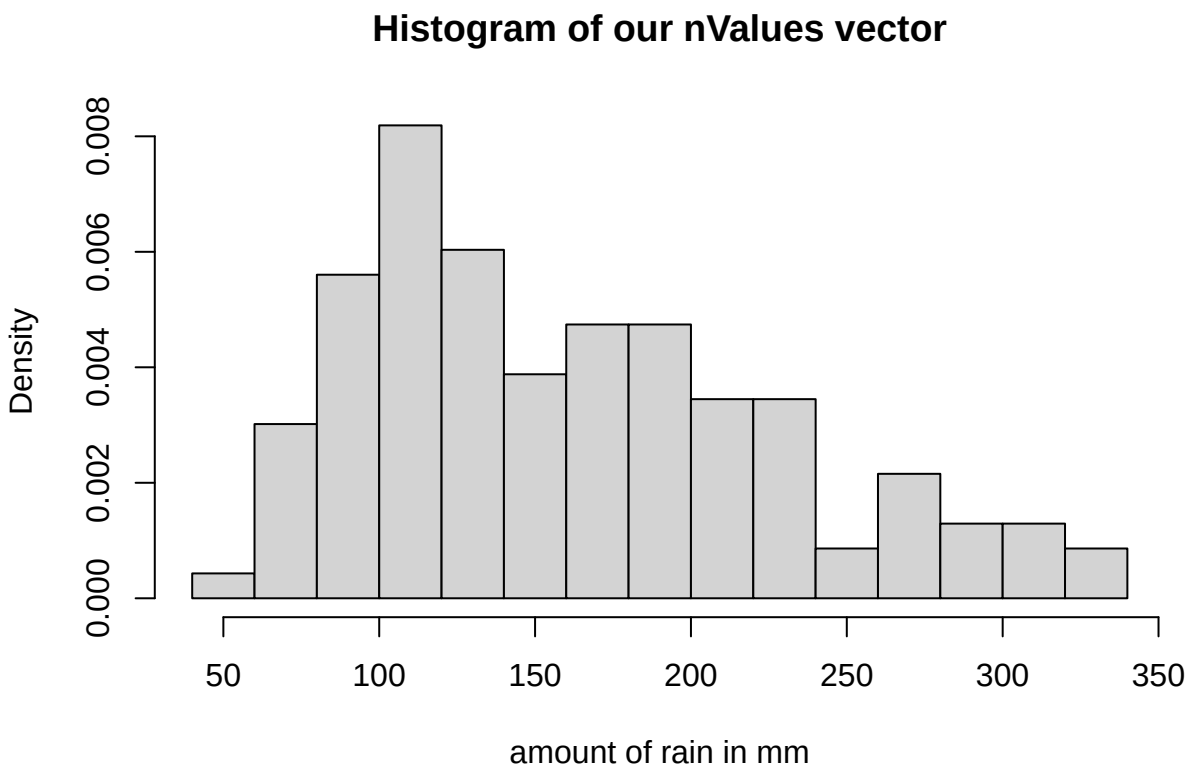
```
#sort the values to see how many we need to remove
sortedValues <- sort(values)
print(sortedValues)
```

```
##   [1]  59  62  71  76  76  77  78  80  81  81  82  84  84  85  88  92  92  92
##  [19]  92  95  98 101 102 103 105 105 105 106 106 107 107 107 108 108 110 110
##  [37] 111 111 114 114 123 123 125 126 127 127 131 132 132 135 136 137 138 140
##  [55] 145 146 146 147 148 153 157 157 160 161 162 164 164 164 168 173 176 178
##  [73] 179 180 186 186 187 189 190 193 195 196 198 199 199 202 204 207 209 210
##  [91] 215 217 219 223 224 229 230 232 234 234 236 244 247 265 269 270 275 280
## [109] 282 282 296 310 310 317 326 333 441
```

```
#We see one outlier, 441, so we remove it from the data to create a vector new values:
nValues <- values[values < max(values)]
```

Now we create a new histogram with our new vector, and we can see that we can work with this histogram.

```
#we choose 10 breaks because it makes a nice shape we can work with
hist(nValues, breaks = 10, freq = FALSE, main = "Histogram of our nValues vector"
     ,xlab = "amount of rain in mm")
```



Finding a distribution and calculating the moment estimator

As previously stated, our distribution doesn't seem symmetrical, and peaks early and then slowly subsides. We have a couple of options to choose from:

- Frechet distribution: This distribution seems to match, however trying to find the moment esimator is very difficult so we opt for a different distribution,
- Gumbel distribution: This distribution also seems to match, however calculating with it seems too difficult,
- Negative-binomial: We opt for this distribution. It matches the shape and we can calculate with it

For the calculations for the moment estimator we want to know the mean of our samples:

```
#we calculate the mean by summing the values and dividing by the length of our set
sumVals <- sum(nValues)
dataMean <- 1/length(nValues) * sumVals
print(dataMean)
```

```
## [1] 161.8448
```

```
dataVar <- var(nValues)
print(dataVar)
```

```
## [1] 4535.106
```

We can calculate the first and second moments of the negative binomial distribution using the booklet of distributions. We derive the formula for the second moment by using the standard formula $\text{var}(X) = \mathbb{E}(X)^2 - \mathbb{E}(X^2)$. We get the following formula:

$$\mathbb{E}\left(X^2\right) = \text{var}\left(X\right) + \mathbb{E}\left(X\right)^2.$$

We find that the first and second moments are given by the following expressions

$$\mathbb{E}(X) = \frac{r(1-p)}{p}, \tag{1}$$

$$\begin{aligned}\mathbb{E}\left(X^2\right) &= \operatorname{var}(X) + \mathbb{E}(X)^2 \\ &= \frac{r(1-p)}{p^2} + \left(\frac{r(1-p)}{p}\right)^2 \\ &= \frac{(p-1)\,r\,((p-1)\,r-1)}{p^2}\end{aligned} \tag{2}$$

We solve equation (1) for r in terms of p and $\mathbb{E}(X)$:

$$\begin{aligned}\mathbb{E}(X) &= \frac{r(1-p)}{p} \\ \iff p \cdot \mathbb{E}(X) &= r(1-p) \\ \iff \frac{p}{1-p} \cdot \mathbb{E}(X) &= r \\ \iff r &= \frac{p}{1-p} \cdot \mathbb{E}(X).\end{aligned}$$

By plugging in this expression for r into equation (2) we can derive a formula for $\mathbb{E}\left(X^2\right)$ solely in terms of p and $\mathbb{E}(X)$:

$$\begin{aligned}\mathbb{E}\left(X^2\right) &= \frac{(p-1)\,r\,((p-1)\,r-1)}{p^2} \\ &= \frac{(p-1)\left(\frac{p}{1-p}\mathbb{E}(X)\right)\left((p-1)\left(\frac{p}{1-p}E(X)\right)-1\right)}{p^2} \\ &= \frac{-p\mathbb{E}(X)\,(-p\mathbb{E}(X)-1)}{p^2} \\ &= \frac{(p\mathbb{E}(X)+1)\,\mathbb{E}(X)}{p}.\end{aligned}$$

We can rearrange this formula as follows:

$$\begin{aligned}\mathbb{E}\left(X^2\right) &= \frac{(p\mathbb{E}(X)+1)\,\mathbb{E}(X)}{p} \\ p\mathbb{E}\left(X^2\right) &= (p\mathbb{E}(X)+1)\,\mathbb{E}(X) \\ p\mathbb{E}\left(X^2\right) &= p\mathbb{E}(X)^2 + \mathbb{E}(X) \\ p\mathbb{E}\left(X^2\right) - p\mathbb{E}(X)^2 &= \mathbb{E}(X) \\ p\left(\mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2\right) &= \mathbb{E}(X) \\ p &= \frac{\mathbb{E}(X)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2}.\end{aligned}$$

Which in turn gives us the following expression for r in terms of $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$:

$$\begin{aligned} r &= \frac{p}{1-p} \cdot \mathbb{E}(X) \\ &= \frac{\frac{\mathbb{E}(X)}{\mathbb{E}(X^2)-\mathbb{E}(X)^2}}{1-\frac{\mathbb{E}(X)}{\mathbb{E}(X^2)-\mathbb{E}(X)^2}} \cdot \mathbb{E}(X) \\ &= \frac{\mathbb{E}(X)}{\mathbb{E}(X^2)-\mathbb{E}(X)^2-\mathbb{E}(X)} \cdot \mathbb{E}(X) \\ &= \frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)-\mathbb{E}(X)^2-\mathbb{E}(X)}. \end{aligned}$$

Therefore we have derived the following expressions for p and r solely in terms of the first and second moments:

$$\begin{aligned} r &= \frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)-\mathbb{E}(X)^2-\mathbb{E}(X)}, \\ p &= \frac{\mathbb{E}(X)}{\mathbb{E}(X^2)-\mathbb{E}(X)^2}. \end{aligned}$$

With the values calculated from the given data, we have

$$\begin{aligned} \mathbb{E}(X) &\approx 5162.05835544 \\ \mathbb{E}(X^2) &\approx 161.8448. \end{aligned}$$

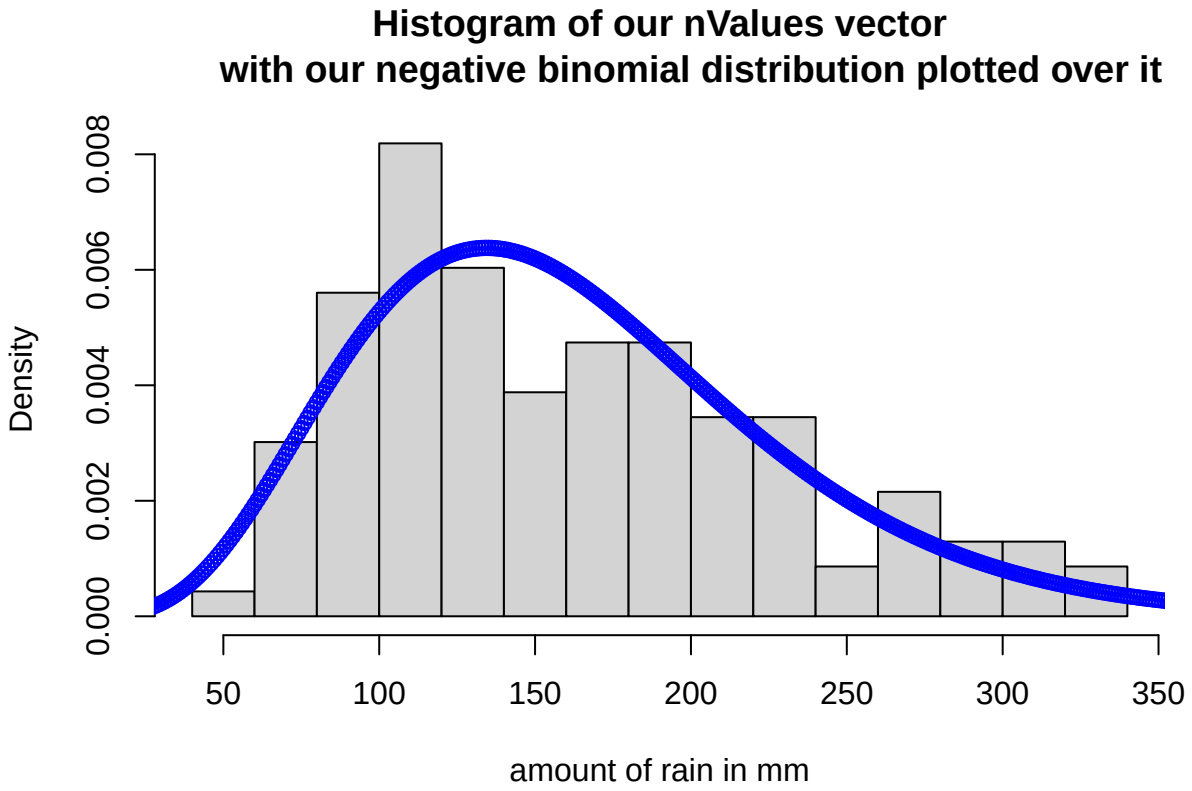
This gives us the following values for p and r:

$$\begin{aligned} p &\approx 0.0356871090428 \\ r &\approx 5.98952276201. \end{aligned}$$

Plotting and matching the distribution

Now that we have estimated our parameters, we will compare the distribution to our data set

```
hist(nValues, breaks = 10, freq = FALSE, main = "Histogram of our nValues vector
      with our negative binomial distribution plotted over it"
     ,xlab = "amount of rain in mm") #our histogram again
points(1:500, dnbinom(1:500, 6, 0.0357), col="blue", pch=10)
```



We can see that the shape matches the shape of our data pretty well. Some things to note:

- The peak of the negative binomial distribution with our estimator parameters is slightly off compared to the peak in the data
- If you compare the length of the middle of each bar to the blue point plot, you will get an answer that is near 0.
- We should expect a lower probability when we do the calculation for +300mm rainfall, the points graph is under the bars of the histogram.

Probability for 300mm of rainfall and comparing with the empirical value

We will now compute the probability that the sewage system has to cope with more than 300mm of rain. We compute:

```
print(1-pnbinom(300, 6, 0.0357))
```

```
## [1] 0.0367948
```

So our probability is 0.0368 or about 3,68% according to our found negative binomial distribution. Now we will compare this value with the empirical probability, which we will compute:

```
#we calculate the empirical probability for rainfall over 300mm
vals_over_300_no_outlier <- sum(nValues > 300)
emp_value_no_outlier <- vals_over_300_no_outlier/length(nValues)
print(emp_value_no_outlier)
```

```
## [1] 0.04310345
```

We get 0,0431 which is 4,31% and we see that this percentage is a bit higher, which checks out with the analysis of our chosen distribution and estimators. Also keep in mind that we chose to erase the outlier to get a cleaner histogram.