Maximum likelihood (ML) estimation

Maria Cobo and Maria Roy

21th of October 2023

1 Exercise

ML estimation of a one-parameter distribution. Let X be a random variable with probability density:

$$f(x|\beta) = \beta e^{(-\beta x+1)} \; with \; x \; > \; 1/\beta$$

We consider a random sample of n observations of this distribution.

a) Write down the likelihood function for a sample of n observations of this distribution

The likelihood function for a sample of n observations is the product of the individual probability density functions for each observation. Therefore, we have to multiply individual PDFs for each observation together.

We assume that the observations are independent and identically distributed. Therefore, we can write the likelihood function as:

$$L(\beta \mid x_1, \; x_2 \; ..., \; x_n) \; = \; \prod_{i=1}^{n} \; f(x_i \mid \beta)$$

Now, we substitute this with our probability density function:

$$L(\beta \mid x_1, \; x_2 \; ..., \; x_n) \; = \; \prod_{i=1}^{n} \; \beta e^{-\beta x_i+1} \; with \; x \; > \; \frac{1}{\beta}$$

b) Obtain the log-likelihood function

In order to obtain the log-likelihood function, we must apply the log operator. This transformation will simplify calculations and it won't change the parameter value that maximizes the likelihood function. Therefore, the log-likelihood function is:

$$l(\beta \mid x) \; = \; lnL(\beta \mid x_1, \; x_2 \; ..., \; x_n) \; = \; ln[\sum_{i=1}^{n} \; \beta e^{-\beta x_i+1}] \; = \; nln(\beta) - \beta \sum x_i + 1$$

for i = 1 to n

c) Find the stationary point(s) of the log-likelihood function analytically

To find the stationary point(s) of the log-likelihood function we need to do the derivative and equal it to zero. The results we get will let us know which are the potential maximum likelihood estimates for β.

$$\frac{\partial}{\partial\beta}\ ln\ L(\beta \mid x_1,\ x_2\ ...,\ x_n)\ =\ \frac{\partial}{\partial\beta}(n\ ln(\beta)\ -\ \beta\sum_{i=1}^{n}x_i\ +\ 1)$$

$$\frac{\partial}{\partial\beta}\ ln\ L(\beta \mid x_1,\ x_2\ ...,\ x_n)\ =\ \frac{n}{\beta}\ -\ \sum_{i=1}^{n}x_i$$

$$\frac{n}{\beta}\ -\ \sum_{i=1}^{n}x_i\ =\ 0 \rightarrow\ \frac{n}{\beta}\ =\ \sum_{i=1}^{n}x_i\ \rightarrow\ \beta\ =\ \frac{n}{\sum_{i=1}^{n}x_i}$$

d) Determine whether the stationary point(s) are maxima or minima

To determine whether the stationary point(s) are maxima or minima we need to perform the second derivative on the log-likelihood function.

$$\frac{\partial^2}{\partial\beta^2}\ lnL(\beta \mid x_1,\ x_2\ ...,\ x_n)\ =\ -\frac{n}{\beta^2}$$

The second derivative is negative therefore, we can assume that n $>$ 0 and $\beta >$ 0. The stationary point that we found is a maximum. Therefore, this stationary point is the maximum likelihood estimate for the parameter β.

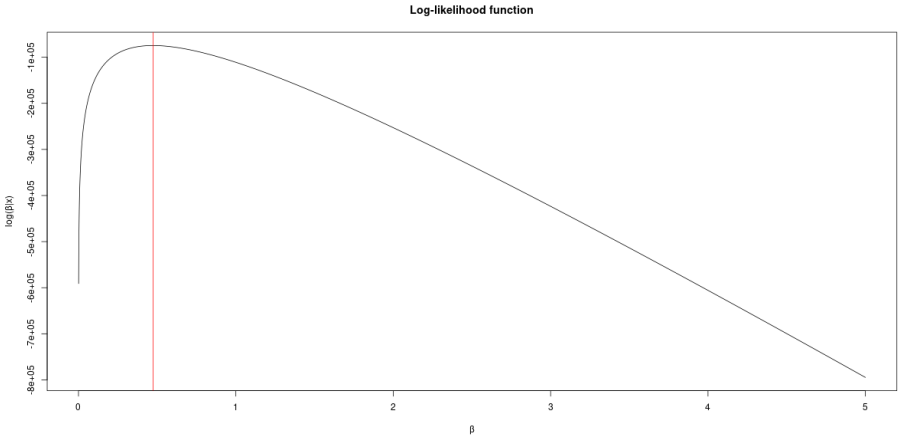
e) Download the file Sample.dat, which contains sample of observations from this probability distribution. Determine the sample size and calculate the value of the ML estimator for this sample

First, with 'read.table()' we read the table that contains the samples of observations. To determine the sample size, we look at the number of rows that our table has and it is 100000. Then, with this information, we can calculate the maximum likelihood estimation for a parameter β. We divide the number of rows by the sum of the data observed and we get the value of β that maximizes the likelihood function based on our data. This value is 0.4736756. The **sample size** for this data set is **100000** and the value of the **maximum likelihood estimator** is **0.4736756**.

f) Plot the log-likelihood function, and assess graphically if your ML estimate coincides with the maximum of this function

To plot the log-likelihood first, we need to determine the β values to see how the log-likelihood function changes across these. We set a range that goes from 0.001 to 5 with an increment of 0.001. Then, we define the log likelihood function where 'x' are the observed points and 'beta' are the beta values. Moreover, we return the maximum log likelihood function that we calculated before: $nln(\beta) - \beta\ \sum_{i=1}^{n}\ x_i + n$ and plot our log likelihood function with our specified β values.

Finally, we add a vertical line to the x-coordinate of the maximum likelihood estimation, this represents the value of the parameter that maximizes the likelihood function. And, we can see that our ML estimate coincides with the maximum of this function.



g) Determine an expression for the Fisher information by calculating $-E = \frac{\partial^2 l}{\partial \beta^2}$

Fisher information corresponds to the negative second derivative of the log-likelihood function with respect to β.
From question d) we know that: $\partial^2 l/\partial \beta^2 \; ln \; L(\beta) = \frac{-n}{\beta^2}$.
Fisher information I(β) is the expected value of this second derivative which is denoted as $-E(\partial^2 l/\partial \beta^2 \; ln \; L(\beta)$. In this case, E is the expectation of $-n/\beta^2$:

$$I(\beta) = -E(\frac{n}{\beta^2})$$

Since β is a constant and not a random variable, the expectation operator does not apply to it.

$$I(\beta) = -\frac{n}{\beta^2}$$

If we compute this function with the beta we get from our sample size, we get that: I(β) = 4.483.489'38. This value represents how much information the data contains about parameter β. A larger Fisher information indicated that the data is more informative about the parameter.

h) Use the Fisher information for obtaining an expression for the variance of the maximum likelihood estimator β^{ML}

$$I(\beta) = E\frac{n}{\beta^2} \rightarrow Var(\beta) \geq \frac{1}{I_x(\beta)} \rightarrow Var(\beta) \geq \frac{\beta^2}{n} \rightarrow Var(\beta) \geq 2.24e^{-06}$$

i) Using the asymptotic normality of the ML estimator, give an expression of a confidence interval for β

$$\hat{\beta_n} \rightarrow N(\beta, \frac{1}{I_x(\beta)} = N(\beta, \frac{\beta^2}{n}) \rightarrow CI = \hat{x} \pm z\frac{s}{\sqrt{n}} \rightarrow CI = \hat{\beta_{ML}} \pm z\sqrt{V(\hat{\beta})}$$

where $\hat{x} = \hat{\beta_{ML}}$, and $\frac{s}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{V(\hat{\beta}}$

j) Calculate a 95 confidence interval for parameter β**, using the data set that you have downloaded**

$$CI(\beta)_z = \hat{\beta_{ML}} \pm z\sqrt{V(\hat{\beta})} \rightarrow CI(\beta)_{0.95} = z \pm 1.96 \frac{\beta}{\sqrt{n}}$$

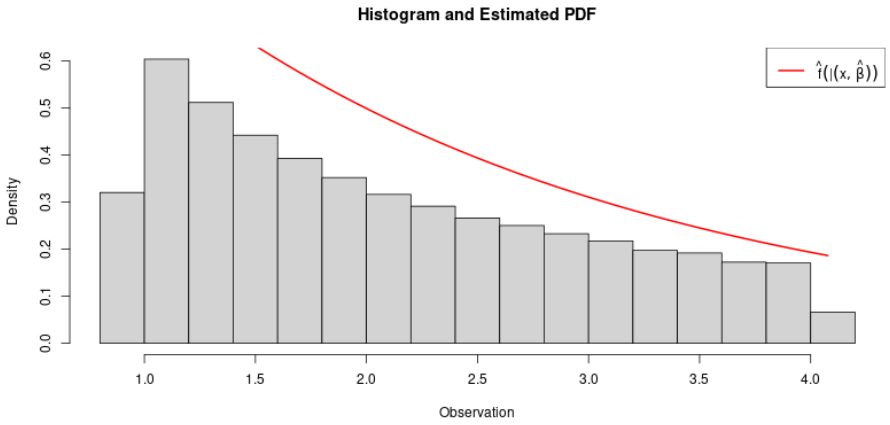
From the data set we have downloaded, the confidence interval is defined by:

$$CI(\beta)_{0.95} = 0.4737 \pm 1.96 \frac{0.4737}{\sqrt{100000}} \rightarrow CI(\beta)_{0.95} = (0.46929, 0.47811)$$

k) Do you think it is tenable that $\beta = 1$?

Based on the ML estimator, the calculated CI and how well the distribution fits our data, $\beta = 1$ is not tenable. The most plausible value for β is around 0.4737, which is supported by the CI, that ranges from (0.46929, 0.47811) and it does not include 1. When β is within this interval, it indicates a plausible range of values based on our data.

l) Make a histogram of the data, using function hist, using the argument freq=FALSE. Over plot the histogram with the estimated probability density f(x|β), using the maximum likelihood estimate. What do you observe?



We can observe that the estimated probability density f(x|β) it does not fit the density of the different values that are observed in the sample we downloaded, since any values fall under the curve.

In simpler terms, the estimated probability density function is not a good fit for the actual data. It may suggest that the model used to estimate the distribution does not accurately represent the real-world data, and there may be some other factors that need to be considered to better match the observed values in the sample.

2 Exercise. Hardy-Weinberg law

The Hardy-Weinberg law in genetics says that the proportions of genotypes AA, Aa and aa are θ^2, $2\theta\ (1-\theta)$, and $(1-\theta)^2$, respectively, where $\theta \in [0, 1]$. Suppose that in a sample of n individuals from the population (small relative to the size of the population), we observe

x_1 individuals of type AA, x_2 individuals of type Aa and x_3 of type aa.

(a) What distribution do the counts (x_1, x_2, x_3) follow?

The counts (x_1, x_2, x_3) for the genotype in a sample from the population follow a multinomial distribution. The multinomial distribution is a generalization of the binomial distribution to multiple categories. In this case, there are three categories corresponding to the three genotypes (AA, Aa, aa) and each individual in the sample can fall into one of these categories. The multinomial distribution describes the probabilities of observing different counts of individuals in each category.

(b) Record the likelihood function, the log-likelihood function and the score function for θ.

Likelihood Function. L(θ) is the probability of observing the specific counts (x_1, x_2, x_3) given the parameter θ. The likelihood function is given by:

$$L(\theta) \ = \ [\frac{n!}{x_1!\ x_2!\ x_3!}]\ (\theta^2)^{x_1}\ (2\theta(1\ -\ \theta))^{x^2}\ [(1\ -\ \theta)^2]^{x_3}$$

Log-Likelihood Function. $ln(L(\theta))$ is obtained by taking the natural logarithm of the likelihood function.

$$\begin{aligned} ln(L(\theta)) \ &= \ ln(n!)\ -\ ln(x_1!)\ -\ ln(x_2!)\ -\ ln(x_3!)\ + \\ &x_1\ ln(\theta^2)\ +\ x_2\ ln(2\theta\ (1\ -\ \theta))\ +\ x_3\ ln[(1\ -\ \theta)^2] \end{aligned}$$

Score Function. $\frac{\partial\ ln(L)}{\partial\ \theta}$ represents the derivative of the log-likelihood function with respect to θ. It is used to find the ML estimator. The score function is given by:

$$\frac{\partial\ ln(L)}{\partial\ \theta} \ = \ x_1\ \frac{2\theta}{\theta^2} + x_2\ (2(1\ -\ \theta)\ -\ 2\theta)\ +\ x_3\ (\frac{-2\ (1\ -\ \theta)}{(1\ -\ \theta)^2})$$

(c) Record the form of the ML estimator for θ

To find the ML estimator for θ, set the score function equal to zero: $\frac{\partial ln(L)}{\partial\ \theta} \ = \ 0$ Then, solve for θ. The form of the ML estimator for θ depends on the specific values of x_1, x_2, x_3. The MLE represents the value of θ that maximizes the likelihood of observing the given genotype counts in the sample.

3 Exercise

Find an example of a published research where they study/apply/use MLE, include the link, with the retrieved data and do a summary of 5-6 text lines.

This article talks about a software package called STEM and how it can estimate the maximum likelihood (ML) species tree from a sample of gene trees. It explains how, when the ML estimate of the species tree is requested, STEM returns the MT for the particular user-specified values of and the gene-specific

rates. Finally, the study emphasizes how STEM can compute ML branch lengths on any given species tree, which reduces the search for high-likelihood trees to a discrete (topology only) space, as well as allows evaluation of any species tree of interest.

This other article discusses the significance of microarray technology in gene expression analysis and cancer diagnosis. It emphasizes the importance of integrating clinical data with microarray data to enhance patient care and treatment. The study explores various statistical methods to identify differentially expressed genes in diseases. Also, the paper demonstrates the equivalence between Maximum Likelihood Generalized Eigenvalue Decomposition (MLGEVD) and generalized ridge regression. Finally, the aim of the study is to improve the accuracy of clinical decision-making processes.

4 R code

Here we display the R code used to compute the values for some previous exercises and to do the plot:

```
# Load data
sample <- read.table("ML/sample.dat", quote="\"",
    comment.char="")
n = length(sample$V1) # get the sample size = 100000
# Calculate beta = 0.4736756
beta = 100000/sum(sample$V1)

# Define the log-likelihood function
log_likelihood <- function(beta, sample) {
  return(n * log(beta) - beta * sum(sample) + n)
}

# f) Estimation of beta using the log-likelihood
    function
# create sequence of beta values
beta_values <- seq(0.001, 5, by = 0.001)
# calculate log-likelihood for each beta value
log_values <- sapply(beta_values, function(beta) log_
    likelihood(beta, sample))
# creates a plot
plot(beta_values, log_values, type = "l", main = "Log-
    likelihood function", xlab = "$\beta$", ylab = "log
    ($\beta$|x)")
# add red point to highlight specific beta value and
    its associate log-likelihood
points(beta, log_likelihood(beta, sample), col = "red"
    )
# add a vertical red line to the specific beta value
abline(v=beta, col = "red")

# defining observed variance of the data
```

```
variance <- 2.2439169e-06
conf_level <- 0.95 # set confidence level
# calculate the Z-score for a given confidence level
z <- qnorm(1 - (1 - conf_level) / 2)

# calculate the margin of error using the Z-score and
    the square root of the variance
margin_of_error <- z * sqrt(variance)

# calculate the upper and lower bounds of the CI for
    beta
lower_bound <- beta - margin_of_error
upper_bound <- beta + margin_of_error

# defining function that estimates the density for
    given data x and parameter beta
estimated_density <- function(x, beta) {
  return(beta * exp(-beta * x + 1))
}

# create a histogram of the data
hist(sample$V1, freq = FALSE, xlab = "Observation",
    main = "Histogram and Estimated PDF")

# create a sequence of x values from 0 to the maximum
    value with 1000 evenly spaced points
# length.out is set to allow a smoother and more
    detailed representation of the estimated PDF
x_val <- seq(0, max(sample$V1), length.out = 1000)

# calculate the estimated probability density function
    (PDF)
pdf_val <- estimated_density(x_val, beta)

# add a red line to the plot that represents the
    estimated PDF
lines(x_val, pdf_val, col = "red", lwd = 2)

# add a legend to the plot
legend("topright", legend = expression(hat(f)(x|hat(
    beta))), col = "red", lwd = 2)
```