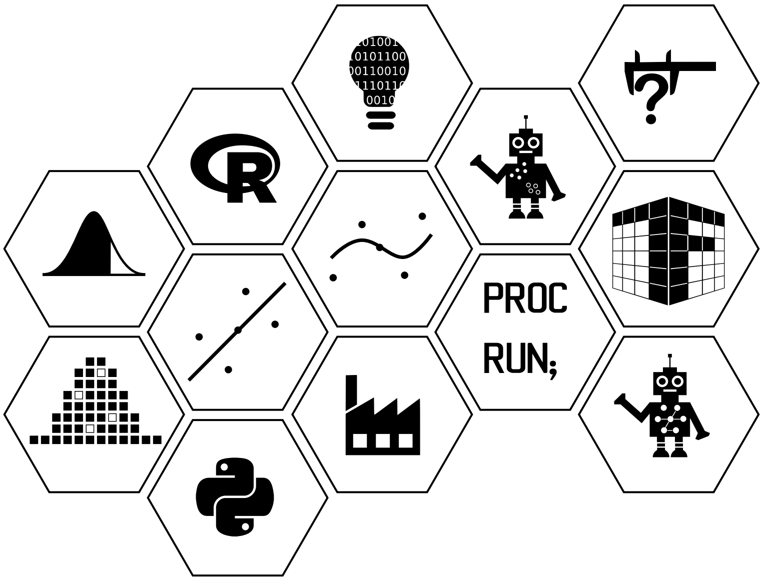
Learning from Data/Data Science Foundations

Week 6: Likelihood II - continuous distributions, optimisation and properties



DATA ANALYTICS GLASGOW

Maximum likelihood estimation, optimisation and properties

In week 5 we motivated and introduced the ideas for `good' point estimators and the approach of maximum likelihood estimation. This week's material will extend these ideas to one parameter continuous distributions, and introduce further properties which will help us to: introduce ideas of interval estimation and testing, and extend likelihood to multiparameter distributions (e.g. normal) in the weeks to follow.

Week 6 learning material aims

The material in week 6 covers:

- maximum likelihood for continuous distributions;
- maximum likelihood estimation on a boundary;
- numerical optimisation;
- further properties of point estimators.

In the video [1] below, an overview is provided of examples for:

- likelihood for a continuous distribution;
- the situation where the maximum occurs on a boundary;
- the situation where no closed form solution for the first derivative of the log-likelihood exists (i.e. we can't find the MLE by hand).

Video

Likelihood topics

Duration 9:56

Likelihood topics

Claire Miller, Eilidh Jack and Suzy Whoriskey

Continuous distributions

Remember that we want to use the method of *Maximum Likelihood* to enable us to find point estimators for parameters of any known statistical distribution.

Continuous case is modelling expected value of a response variable

When X_i is a continuous random variable, then the probability density function of X_i evaluated at x_i does not directly represent the probability of the data (see the **Probability and Stochastic Models** or **Probability and Sampling Fundamentals/Sampling Fundamentals** course for more details). However, $f_i(x_i)$ is (approximately) proportional to the model probability that X_i lies in a small interval around the value x_i. So, it is reasonable to take the likelihood of the unknown parameter, θ, to be

$$f_1(x_1) \times f_2(x_2) \times \cdots \times f_n(x_n)$$

i.e.

$$L(\theta; x_1, \ldots, x_n) \propto \prod_{i=1}^{n} f_i(x_i).$$

As in the discrete case, the maximum of this function can usually be determined by differentiating the log-likelihood. Again, we will shorten $L(\theta; x_1, \ldots, x_n)$ to $L(\theta)$.

General approach:

For data, $x_1, \ldots, x_n$ and general parameter θ:

- Step 1: First, we must evaluate the **likelihood function** $L(\theta, \mathbf{x})$, (you can obtain the pdf for known distributions from the probability distribution sheet).
- Step 2: evaluate the **log-likelihood function** $\ell(\theta)$,
- Step 3: differentiate w.r.t. θ and set equal to 0: $\ell'(\theta) = 0$,
- Step 4: solve for θ,
- Step 5: verify that you have a maximum, $\ell''(\theta) < 0$.

For a continuous distribution, the only step that differs slightly here (from the discrete case) is in step 1. For a continuous distribution the likelihood for the parameter of interest θ is proportional to the product of the probability density functions for n random variables/observations recorded.

Example 1

Exponential model with parameter θ

A business wishes to monitor the useage of their website and so they record the time in days between hits on their website for 6 months:

1, 5, 15, 2, 3, 45, 13, 3, 3, 16, 23, 42, 4, 7, 4

What can we say about the mean time between hits?

The exponential distribution is commonly used to model the time elapsed between events and so let's consider that for example 1.

Model: $X_1, X_2, \ldots, X_{15}$ independent, with each $X_i \sim \text{Expo}(\theta) \quad \theta > 0$

Data: $x_1, x_2, \ldots, x_{15}$

Likelihood:

$$L(\theta) \propto \prod_{i=1}^{15} f_i(x_i),$$

$$L(\theta) \propto \prod_{i=1}^{15} \theta\, e^{-\theta x_i} = \theta^{15} e^{-\theta \sum x_i},$$

Log-likelihood:

$$\ell(\theta) = 15 \log_e \theta - \theta \sum_{i=1}^{15} x_i,$$

$$\ell'(\theta) = \frac{15}{\theta} - \sum_{i=1}^{15} x_i,$$

$$\ell'(\theta) = 0 \text{ when } \theta = \frac{n}{\sum_{i=1}^{15} x_i} = \frac{1}{\bar{x}},$$

$$\ell''(\theta) = -\frac{15}{\theta^2},$$

and this is < 0 for all $\theta > 0$.

Therefore, $\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$.

Using the data for this example, it can be seen that, $\bar{x} = 12.4$ and hence $\hat{\theta}_{MLE} = 1/\bar{x} = 0.081$.

Let's explore the data in `R` to plot the likelihood and log-likelihood function and find the maximum likelihood estimate for $\hat{\theta}$.

In `R` :

```r
hittime <- c(1, 5, 15, 2, 3, 45, 13, 3, 3, 16, 23, 42, 4, 7, 4)

## The number of observations is given by:
n <- length(hittime)

## The maximum likelihood estimate (MLE) was found to be:
thetahat <- 1/mean(hittime)

## To plot the log-likelihood set up a sequence of values for theta around the MLE
theta <- seq(0.01, 0.2, length = 50)

## The log-likelihood can be found using:
loglik <- n * log(theta) - theta * sum(hittime)

## This can then be plotted against theta:
plot(theta, loglik, type = "l", lwd=3, ylab="log-likelihood",
xlab=expression(theta))
abline(v=thetahat, lwd=3, col="yellow")
```

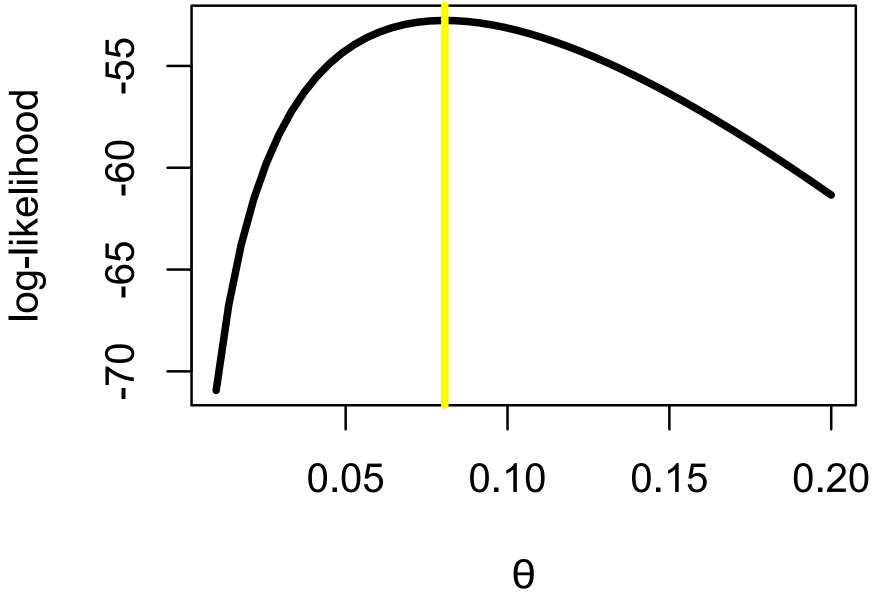


Figure 1

Task 1

Data were recorded on the time (intervals in service-hours) between the failures of the air-conditioning equipment in a Boeing 720 aircraft. The observations were:

`50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88, 46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14.`

We are interested in the mean time between failures.

Assuming these data follow an Expo(θ) distribution, find the MLE for θ by hand and plot the likelihood and log-likelihood functions over a range of values for θ in \texttt{R}.

It is always important to bear in mind, though, that the MLE of θ might be found on the boundary of the range of θ. The following example demonstrates this, in a very important special case.

Example 2

A uniform model

An example for a uniform distribution could be people arriving at a bus stop to wait for a bus that comes by regularly. They do not know what time the bus came by last. The arrival time of the next bus is a continuous uniform distribution e.g. $[0, \theta]$ measured in minutes.

Model: $X_1, X_2, \ldots, X_n$ independent, with each $X_i \sim \mathrm{U}(0, \theta) \quad 0 \leq x_i \leq \theta$

Data: $x_1, x_2, \ldots, x_n$

Likelihood:

$$L(\theta) \propto \prod_{i=1}^{n} f_i(x_i),$$

$$L(\theta) \propto \prod_{i=1}^{n} \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n \quad 0 \leq x_i \leq \theta.$$

Strangely, this doesn't depend on the x_i's!

Actually, it does, because we require $\theta \geq x_i$ for each x_i, otherwise the contribution to the likelihood function is 0. So we must have $\theta \geq \max x_i$.

Log-likelihood:

$$\ell(\theta) = -n \log_e(\theta), \text{ for } \theta \geq \max x_i,$$

$$\ell'(\theta) = -\frac{n}{\theta},$$

and this is < 0 for all $\theta > 0$.

We cannot solve this equation for θ, so this is a case where the maximum lies at an end-point. The MLE must lie at the left-hand end-point of the valid range. Here, $\hat{\theta}_{MLE} = \max(x_i)$.

If we plot the log-likelihood function we can see that the maximum occurs on the left hand boundary. An illustration of this is provided in the first video for this week (see the link on page 2).

Task 2

Uniform distribution cont.....

Suppose that $x_1, x_2 \ldots, x_n$ are observations of the independent random variables $X_1, X_2, \ldots, X_n$ respectively. Obtain the maximum likelihood estimator of the unknown parameter θ in the model for $X_1, X_2, \ldots, X_n$, $X_i \sim \mathrm{U}(-\theta, \theta)$ where $0 < \theta$, and $-\theta \leq x_i \leq \theta$.

We will consider other examples for continuous distributions (e.g. normal) when we consider likelihood for multiparameter distributions in week 8.

Computation of MLEs

We have found closed-form expressions for $\hat{\theta}$ for simple cases. However, to find the maximum of $L(\theta; \mathbf{x})$ can be an optimisation problem [2]. Optimisation supplementary materials: https://moodle.gla.ac.uk/pluginfile.php/5664316/mod_resource/content/3/Numerical%20methods%20in%20R.pdf

Complications arise here when it is not possible to find the root of the equation $\ell'(\theta) = 0$ in closed form. In such cases, it is often possible to find the root numerically, for example, by an iterative procedure known as the Newton-Raphson method (also known as Newton's method).

This is based on the idea that, if $\theta^{(0)}$ is a first approximation to a root of the general equation

$$\ell'(\theta) = 0,$$

then a better approximation to the root is given by

$$\theta^{(1)} = \theta^{(0)} - \frac{\ell'(\theta^{(0)})}{\ell''(\theta^{(0)})}.$$

A series of increasingly better approximations may then be obtained, using the iterative formula, that for iteration j:

$$\theta^{(j+1)} = \theta^{(j)} - \frac{\ell'(\theta^{(j)})}{\ell''(\theta^{(j)})}.$$

This iterative procedure is stopped when the numerical approximation has converged to the correct value of the root of the equation (to a pre-determined level of accuracy), e.g. when

$$-0.001 \leq \frac{\ell'(\theta^{(j)})}{\ell''(\theta^{(j)})} \leq 0.001.$$

In order to start the Newton-Raphson method for finding the root of the equation, and hence the MLE of θ, we need a first approximation.

In many examples the Newton-Raphson method converges to a solution in very few steps as long as a sensible first approximation is available.

In practice, application of the Newton-Raphson procedure is cumbersome to do by hand. It is easier to use the `optimize` or `optim` function in `R`. We'll show a few examples throughout the course using `optim` as an illustration since this more general function can be used for estimating more than one parameter.

Example 3

As part of a project on the occurence and distribution of a species of hoverfly in Trinidad, data were collected on the numbers of larvae found on 20 randomly sampled flowering bracts of Heliconia plants. The observed numbers were `9, 14, 3, 3, 8, 7, 7, 6, 7, 0, 6, 0, 5, 1, 3, 12, 2, 4, 0, 11` One possible model for the number of larvae (y_i) is that they are observations of independent random variables with probability function

$$p_i(y_i) = \theta^{y_i}(1-\theta) \qquad 0 < \theta < 1$$

where the y_i are non-negative integers.

The likelihood and log-likelihood can be written as (assuming a discrete probability function):

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta) = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^n,$$

$$\ell(\theta) = \sum_{i=1}^{n} y_i \log(\theta) + n \log(1-\theta).$$

We could evaluate this by hand to find that, $\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n+\sum_{i=1}^{n} y_i} = 0.84375$. However, let's evaluate this in `R` using the `optim` function to see if we obtain the same result. In `R` :

```
## The data
Heliconia <- c(9, 14, 3, 3, 8, 7, 7, 6, 7, 0, 6, 0, 5, 1, 3, 12, 2, 4,
0, 11)

## Find the sample size:
n <- length(Heliconia)

## Set up a range of values for
theta <- seq(0.4, 1, length=100)

## Compute the log-likelihood:
LogLik <- (sum(Heliconia))*log(theta)+n*log(1-theta)

## Plot the log-likelihood:
plot(theta, LogLik, type="l", lwd=3, ylab="log-likelihood",
xlab=expression(theta))
```

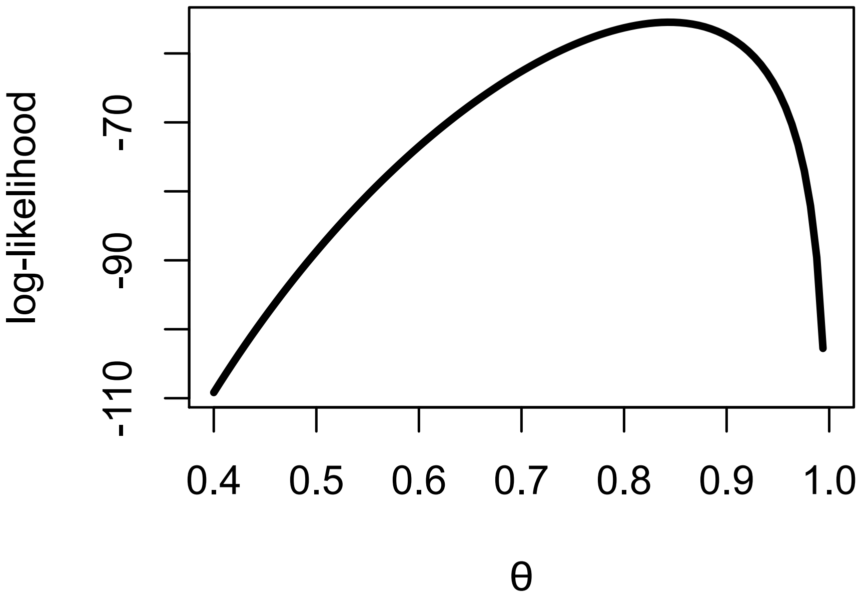


Figure 3

```
## The plot suggest that a plausible initial estimate of theta is around
0.84 or 0.85.

## Construct a function for the log-likelihood:
Helc <- function(theta,y,n){
  sum(y)*log(theta)+n*log(1-theta)
}

## Optimise the function Helc to estimate the parameter:
optim(par=0.844, fn=Helc, method="BFGS",control=list(fnscale= -1),
y=Heliconia, n=20)
```

R Console

```
$par
[1] 0.8437497

$value
[1] -55.47506

$counts
```

```
function gradient
12        3

$convergence
[1] 0

$message
NULL
```

Here are a few details on the arguments in this function:

- {`par`}: is the array of initial parameter values, corresponding to the parameters of your function (here we only have one initial estimate for theta of 0.844, and this can be estimated from the plot of the log-likelihood);

- {`fn`}: is the name of the function that you want to maximize;

- {`method`}: selects one of several maximization methods:'BFGS' is a variant on the Newton method known as a 'quasi-Newton' method;

- {`control`}: is a list containing a number of optional control parameters: `fnscale` should be set to -1 to make the function perform maximisation rather than minimization.

After all the {`optim`} arguments, you then supply the (named) arguments for your function, just `y` and `n` in this case, the data and the number of data points respectively. For more information see `?optim`.

Looking at the `R` output here, it tells us that the function has converged to a parameter value for θ of 0.8437, `par` in the output. This is the most important part of the output for us, the rest of the output gives: the value of the log-likelihood at convergence (`value`), information on the number of iterations (`counts`), and 0 for `convergence` indicates successful completion.

In practice of course you can use this function in the situation where we can solve for the MLE by hand but also in the situation where there is no closed form solution.

Task 3

For the air conditioning example in task 1, the log-likelihood function in `R` is:

`loglikfn <- function(theta,x,n){n*log(theta)-theta*sum(x)}` This function is simply

$$\ell(\theta) = n\log_e(\theta) - \theta\sum_{i=1}^{n} x_i$$

which we've shown in task 1 to be maximized when $\theta = 0.016$.

1. Use the `optim` function in `R` to also show this. Initially, take a starting value of close to 0.016.

2. Here we've started with an estimate for the parameter θ which is quite close to the value that we already know should be the MLE. Try different values for the initial parameter here to see the sensitivity of the `optim` function to the starting values.

Additional Properties of Point Estimators

In week 5, the properties of unbiased and consistent were introduced for point estimators. Using these two properties may still leave a number of candidate estimators. A further property is to look for minimum variance unbiased estimators (MVUEs). The words *efficient* and *efficiency* when applied to estimators refer to the variances of the estimators. The lower the variance of an unbiased estimator, the more efficient it is.

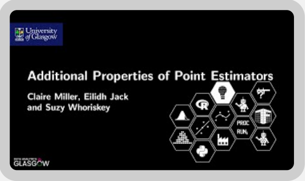
Please note: you will not be expected to carry out examples of (or state full definitions for) the properties in this section of the notes in the class test. However, it's important that you are aware of them at least intuitively for other courses that follow. The only property that we will use reasonably extensively throughout the rest of the course is the definition for **sample information (Definition 3)**, which is important for interval estimation, and so you should be able to define this.

The video below provides an illustration of the properties described below in the definitions and theorems, which combine to enable us to identify minimum variance unbiased estimators:

Video

Additional properties of point estimators

Duration 4:51



Additional Properties of Point Estimators

Claire Miller, Eilidh Jack
and Suzy Whoriskey

Definition 1

Mean Squared Error (MSE)

One way of considering bias and variance together is via a combined measure called the mean squared error.

The quality of a point estimator is sometimes assessed by the mean squared error (MSE). The MSE of an estimator $\hat{\theta}$ of a parameter θ is defined to be:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right].$$

This can be written as:

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left[\mathrm{bias}(\hat{\theta})\right]^2.$$

Supplement 1

Here, as supplementary material, is an outline derivation of the statements for Mean Squared Error above.

$$\begin{aligned}
\mathrm{Var}\left[(\hat{\theta}-\theta)\right] &= \mathrm{E}\left[(\hat{\theta}-\theta)^2\right] - \left[\mathrm{E}(\hat{\theta}-\theta)\right]^2 \\
\mathrm{E}\left[(\hat{\theta}-\theta)^2\right] &= \mathrm{Var}\left[(\hat{\theta}-\theta)\right] + \left[\mathrm{E}(\hat{\theta}-\theta)\right]^2 \\
&= \mathrm{Var}(\hat{\theta}) + \left[\mathrm{E}(\hat{\theta})-\theta\right]^2 \\
&= \mathrm{Var}(\hat{\theta}) + \left[\mathrm{bias}(\hat{\theta})\right]^2
\end{aligned}$$

Note:

- in line 1, here we are using an expression for the variance;
- in line 3, we are using the fact that the variance of a random variable plus a constant is just the variance of the random variable;
- in line 4, we are using the fact that the expectation of a constant is a constant, since θ is the true value of the parameter and hence is a constant.

Definition 2

Efficiency

An unbiased estimator is said to be efficient if it has the minimum possible variance; the efficiency of an unbiased estimator is the ratio of the minimum possible variance to the variance of the estimator.

Minimum possible variance

It is usual to take the following well-known lower bound to the variance of unbiased estimators as the *minimum variance*.

Theorem 1

The Cramer-Rao inequality (and lower bound)

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from the distribution with p.d.f $f(x;\theta)$. Subject to certain regularity conditions on $f(x;\theta)$, we have that for any unbiased estimator $\hat{\theta}$ for θ,

$$\mathrm{Var}\left[\hat{\theta}\right] \geq I_\theta^{-1}$$

where

$$I_\theta = E\left[\left(\frac{d\ell(\theta)}{d\theta}\right)^2\right] = -E\left[\left(\frac{d^2\ell(\theta)}{d\theta^2}\right)\right]^{**}$$

**This will be proved later in the course.

$L(\theta;\mathbf{x})$ is the likelihood function defined earlier and $\ell = \log_e(L(\theta))$. I_θ^{-1} is known as the Cramer-Rao lower bound, and the corresponding inequality is the Cramer-Rao inequality. I_θ is sometimes known as the Fisher information about θ in the observations and $\frac{d\ell(\theta)}{d\theta}(\mathrm{i.e.}\, \ell'(\theta))$ is sometimes known as the score function $s(\mathrm{x};\theta)$.

All the work that we have done so far for the likelihood has only enabled us to come up with point estimators (or point estimates) for our paramaters from a known probability distribution. In addition to obtaining information on a point estimate for a parameter, we also want to capture the variability/uncertainty in that estimate, since we know that different samples of data will provide us with slightly different point estimates. The theorem above tells us that we can get an indication of the information provided by the data, i.e. how certain are we about our parameter estimate?, by investigating the likelihood, and hence, log-likelihood function in more detail. Therefore, we can define the *sample information* using the theorem above.

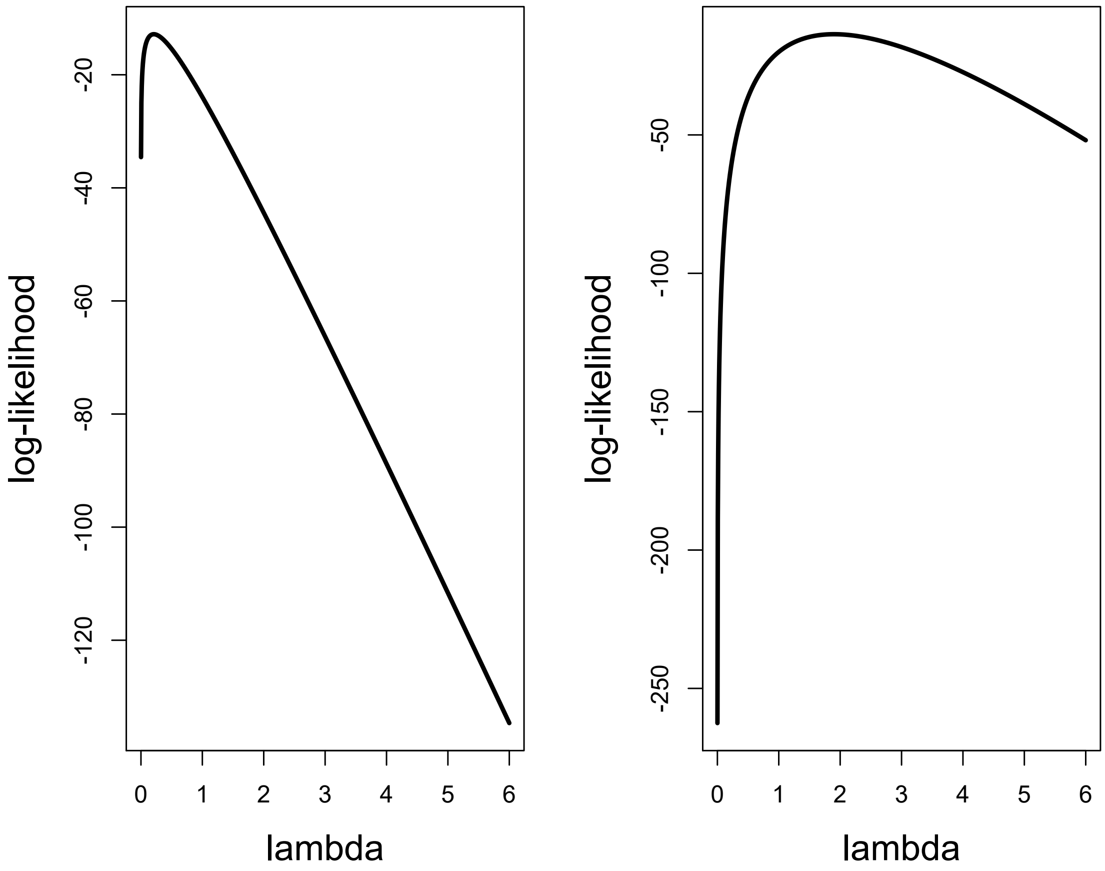
Definition 3

Sample Information (Important)

Since we know that the second derivative of the log-likelihood function will be negative at the maximum, we define the **sample information**, denoted by $k(\mathbf{x})$, as

$$k(\mathbf{x}) = -\left[\left(\frac{d^2\ell(\hat{\theta}_{MLE})}{d\hat{\theta}^2_{MLE}}\right)\right] = -\ell''(\hat{\theta}_{MLE}).$$

We can illustrate the usefulness of this using the two example plots below:



LOW VARIANCE ^^ (curve is tight around MLE)

The sample information by definition provides a value for the curvature at the MLE. In the plot on the left the curve is quite tight around the MLE i.e. there is high curvature. In the plot on the right the curve is more spread out around the MLE i.e. low curvature. The strength of the curvature gives us an indication as to how certain we are around our parameter estimate, and hence we will be able to use this information (along with the Cramer-Rao lower bound) to give us an estimate of the variability in our estimator. A log-likelihood function with high curvature suggests we can estimate quite precisely i.e. with low variance. A log-likelihood function with low curvature suggests we are less accurate in our estimation and hence we will get an estimate with high variance.

We'll return to this idea at several other points throughout the course.

Definition 4

Sufficiency

Sufficiency is important for obtaining minimum variance unbiased estimators.

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from $f(x; \theta)$. Suppose further that $t(x_1, x_2, \ldots, x_n)$ is a function of the observations $x_1, x_2, \ldots, x_n$ and not of θ and that $T(X_1, X_2, \ldots, X_n)$ is the corresponding random variable.

T is then a statistic, and T is sufficient for θ (a sufficient statistic for θ) if the conditional distribution of $X_1, X_2, \ldots, X_n$ given the value T, does not depend on θ.

What this means is that if T is sufficient for θ, then it contains all the information about θ which is contained in the sample data (e.g. X_i); once the value of T is known we can squeeze no more information out of the sample data (e.g. X_i) regarding θ.

This definition of sufficiency does not indicate how to go about finding a sufficient statistic. The following *factorisation theorem* can help with this.

Theorem 2

Factorisation theorem

For $T(X_1, X_2 \ldots X_n)$ to be sufficient for a parameter θ, the joint probability function factors in the form:

$$f(\mathbf{x} \mid \theta) = g[T(\mathbf{x}), \theta] h(\mathbf{x})$$

i.e. the density f can be factored into a product such that one factor, h, does not depend on θ and the other factor, which does depend on θ, depends on x only through $T(\mathbf{x})$.

Example 4

For example for the Poisson distribution (from week 5):

$X_1, \ldots, X_n, X_i \sim \text{Poi}(\lambda)$

Likelihood:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \quad \lambda \geq 0 \end{aligned}$$

and hence $\sum x_i$ is a sufficient statistic.

Theorem 3

The Rao-Blackwell Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample of observations from a distribution with pdf $f(x;\theta)$. Suppose that T is a sufficient statistic for θ and that $\hat{\theta}$ is any unbiased estimator for θ. Define $\hat{\theta}_T = E[\hat{\theta} \mid T]$. Then,

- $\hat{\theta}_T$ is a function of T alone.
- $E(\hat{\theta}_T) = \theta$
- $\text{Var}(\hat{\theta}_T) \leq \text{Var}(\hat{\theta}).$

The Rao-Blackwell theorem gives a quantitative rationale for basing an estimator of a parameter θ on a sufficient statistic if one exists.

Example 5

For example for the Poisson distribution:

$X_1, \ldots, X_n, X_i \sim \text{Poi}(\lambda)$

We've seen in the previous example that $\sum x_i$ is a sufficient statistic.

We've seen (in week 5) that $\hat{\lambda}_{MLE} = \bar{x}$ and is an unbiased estimator, and hence $\hat{\lambda}$ is a function of a sufficient statistic and is a minimum variance unbiased estimator for λ.

We will revisit many of these properties later when we consider properties of Maximum Likelihood Estimators specifically to enable us to develop theory for interval estimation and hypothesis testing for likelihood.

Learning outcomes for week 6

By the end of week 6 you should be able to:

- apply the principle of maximum likelihood to obtain point estimates of parameters in one-parameter continuous statistical models;
- state approaches to use for numerical optimisation in the cases where a closed-form solution is not possible, and make appropriate use of computational approaches here;
- state further properties of point estimators and define the term *sample information*.

Review exercises, selected video solutions and written answers to all tasks/review exercises are provided overleaf.

Review exercises

Task 4

In a study to investigate the short term price movements in FTSE 100 shares, the price change (in pounds) was recorded for 20 of the companies represented on the index. The following data are obtained:

```
-0.70 -0.56 -0.47 -0.34 -0.33 -0.32 -0.27 -0.26 -0.12 0.98
  -0.01  0.02  0.05  0.08  0.14  0.15  0.20  0.23  0.27 0.45
```

Assume that these data are from an independent random sample, and follow a $\mathsf{N}(0, \phi)$ distribution.

Find the maximum likelihood estimate of ϕ.

Note: it might be useful to plot the log-likelihood against ϕ to check your result. The dataset is available at: http://www.stats.gla.ac.uk/~claire/RData_2021.html in Week 6 - review exercises (`R` Data file) in the object `FTSE` .

Task 5

In a study looking at the effect of rising average temperatures on insect pest populations, the time from egg hatch to adulthood was recorded for 20 fruitflies of the species *Drosophila Simulans*. The recorded times (T_i) in days are:

```
11.7 15.0 17.5 14.1 12.6 13.2 11.4 14.5 11.4 13.8
       13.0 14.9 12.0 13.5 13.4 13.0 12.6 10.3 11.6 11.4
```

The minimum development time implied by the physiological maximum development rate is about 10 days. A reasonable model for these data is that they are observations of independent r.v.s. $T_i = 10 + X_i$ where X_i has a distribution with p.d.f.

$$f(x_i) = \frac{1}{2\beta^3} x_i^2 e^{-x_i/\beta} \qquad x_i > 0, \;\; \beta > 0$$

Find the maximum likelihood estimate of β.

Task 6

Given data $x_1, \ldots, x_n$ obtained as a random sample from the $\mathrm{Ga}(2, \theta)$ distribution, find the maximum likelihood estimate of θ.

Task 7

The leaves of a certain species of plant are examined for insect infestation. When at least one insect is present on a leaf, the total number of insects on the leaf is recorded. Otherwise, no record is kept of that particular leaf. The table below shows the results from examining 100 leaves.

no. of insects, k	**1**	**2**	**3**	**4**	**5**	**6**	**7**
frequency, n_k	26	27	18	18	8	2	1

It is believed that the number of insects infesting a leaf follows a Poisson distribution. Since the zero category is missing (no recorded data), the recorded data follow what is known as a Truncated Poisson distribution.

Model: $X_1, X_2, \ldots, X_n$ independent, with each X_i following a Truncated Poisson distribution

Data: $x_1, x_2, \ldots, x_n$, where n_k observed values are equal to $k\,(k = 1, 2, \ldots)$

The probability mass function for each random variable is:

$$\frac{e^{-\lambda}\lambda^x}{(1 - e^{-\lambda})x!}$$

If the data were not truncated, then the MLE would occur at the sample mean, $\bar{x}$. This is a sensible first approximation to the MLE for the truncated case.

Use the `optim` function in `R` to find $\hat{\lambda}_{MLE}$. The data can be found at: http://www.stats.gla.ac.uk/~claire/RData_2021.html in Week 6 - review exercises (`R` Data file) in the object `Insects` . (Hint: you can follow Example 3 in the Week 6 Learning Material to set up your function, using the range of around 2.2 to 2.8 for your values of λ.)

Answer 1

Air conditioning failures:

Model: $X_1, X_2, \ldots, X_{24}$ independent, with each $X_i \sim \mathrm{Expo}(\theta) \quad \theta > 0$

Data: $x_1, x_2, \ldots, x_{24}$

Likelihood:

$$L(\theta) \propto \prod_{i=1}^{24} f_i(x_i),$$

$$L(\theta) \propto \prod_{i=1}^{24} \theta\, e^{-\theta x_i} = \theta^{24} e^{-\theta \sum x_i},$$

Log-likelihood:

$$\ell(\theta) = 24 \log_e \theta - \theta \sum_{i=1}^{24} x_i,$$

$$\ell'(\theta) = \frac{24}{\theta} - \sum_{i=1}^{24} x_i,$$

$$\ell'(\theta) = 0 \text{ when } \theta = \frac{n}{\sum_{i=1}^{24} x_i} = \frac{1}{\bar{x}},$$

$$\ell''(\theta) = -\frac{24}{\theta^2},$$

and this is < 0 for all $\theta > 0$.

Therefore, $\hat{\theta}_{MLE} = \frac{1}{\bar{x}} = 0.016.$

```
aircond <-
c(50,44,102,72,22,39,3,15,197,188,79,88,46,5,5,36,22,139,210,97,30


## The number of observations is given by:
n <- length(aircond)

##  The maximum likelihood estimate (MLE) was found to be:
thetahat <- 1/mean(aircond)

##  To plot the log-likelihood set up a sequence of values for
theta around the MLE
theta  <- seq(0.005, 0.025, length = 50)

## The log-likelihood, and associated plot, can be found using:
loglik <- n * log(theta) - theta * sum(aircond)
plot(theta, loglik, type="l", lwd=3, ylab="log-likelihood",
xlab=expression(theta))
```

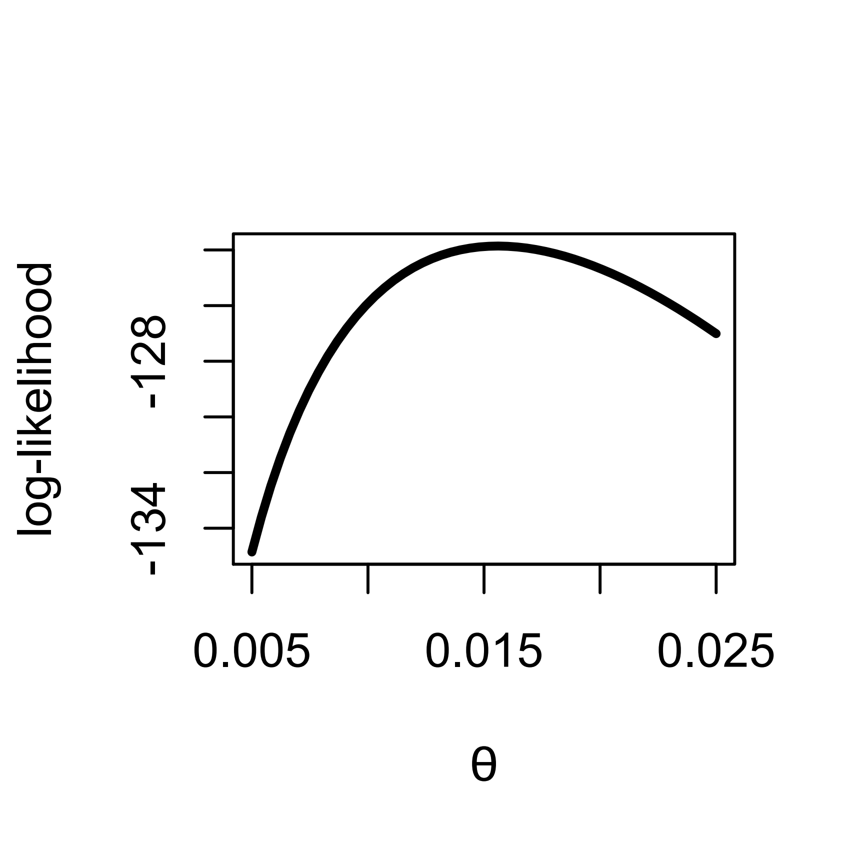


Figure 2

Answer 2

$$f(x_i) = \frac{1}{2\theta}, \qquad -\theta \leq x_i \leq \theta$$

The restriction on the range of the x_i values means that $\theta \geq \mid x_i \mid$ for $i = 1, 2, \ldots, n$. Hence, $\theta \geq \max \mid x_i \mid$. Now,

$$L(\theta; x_1, \ldots, x_n) \propto \prod_{i=1}^{n} \left(\frac{1}{2\theta} \right) = \left(\frac{1}{2\theta} \right)^n, \quad \theta \geq \max \mid x_i \mid$$

$L(\theta)$ does not have a turning point in this interval. The maximum value occurs on the boundary of its range, i.e. at $\max \mid x_i \mid$. Hence,

$$\hat{\theta}_{MLE} = \max \mid x_i \mid$$

Answer 3

Air conditioning example:

```
aircond <-
c(50,44,102,72,22,39,3,15,197,188,79,88,46,5,5,36,22,139,210,97,30

n <- length(aircond)
loglikfn <- function(theta,x,n){n*log(theta)-theta*sum(x)}
optim(par=0.01, fn=loglikfn,
method="BFGS",control=list(fnscale= -1), x=aircond, n=n)
```

R Console

```
$par
[1] 0.01561576
```

```
$value
[1] -123.86

$counts
function gradient
40          7

$convergence
[1] 0

$message
NULL
```

In the output here the main thing to focus on is the `par`, which gives us our estimate of our maximum for θ. This is reported at 0.0156, which we can see is very close to the MLE which you found by hand in task 1 of 0.016.

Here we've started with an estimate for the parameter θ which is quite close to the value that we already know should be the MLE. Try different values for the initial parameter here to see the sensitivity of the `optim` function to the starting values e.g. `par=0.05`.

Answer 4

FTSE example

Model: $X_1, X_2, \dots, X_n$ independent, with each $X_i \sim \mathrm{N}(0, \phi)$

Data: $x_1, x_2, \dots, x_n, \quad n = 20$

$L(\phi) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{\phi}} \frac{1}{\sqrt{2\pi}}\, e^{-x_i^2/(2\phi)} = K\phi^{-n/2}\, e^{-\sum_{i=1}^{n} x_i^2/(2\phi)}$

$\ell(\phi) = -\frac{1}{2} n \log\phi - \sum_{i=1}^{n} x_i^2/(2\phi) + \log(K)$

$\ell'(\phi) = -\frac{1}{2} n/\phi + \sum_{i=1}^{n} x_i^2/(2\phi^2)$

POTENTIAL HELP:
https://medium.com/@lorenzojcducv/maximum-likelihood-for-the-normal-distribution-966df16fd031

$\ell'(\phi) = 0 \text{ when } \phi = \sum_{i=1}^{n} x_i^2/n$

$\ell''(\phi) = \frac{1}{2} n/\phi^2 - \sum_{i=1}^{n} x_i^2/\phi^3$ and this is $-\frac{1}{2}\frac{n^3}{(\sum_{i=1}^{n} x_i^2)^2} = -\frac{n}{2\hat{\phi}^2}$, when $\phi = \sum_{i=1}^{n} x_i^2/n$.

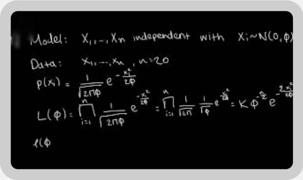
Therefore, $\hat{\phi} = \sum_{i=1}^{n} x_i^2/n$. Evaluating this gives, $\hat{\phi}_{MLE} = 0.144$.

STANDARD DEVIATION^^

Video

Video model answers

Duration 7:29


Model: $X_1, \ldots, X_n$ independent with $X_i \sim N(0, \phi)$
Data: $x_1, \ldots, x_n$, $n=20$
$p(x_i) = \frac{1}{\sqrt{2\pi\phi}} e^{-\frac{x_i^2}{2\phi}}$
$L(\phi) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\phi}} e^{-\frac{x_i^2}{2\phi}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\phi}} e^{-\frac{x_i^2}{2\phi}} = K\phi^{-\frac{n}{2}} e^{-\frac{\sum x_i^2}{2\phi}}$
$l(\phi$

Answer 5

Finding the MLE:

Dropping the constants which don't depend on β, the log-likelihood is

$$l(\beta) \propto -3n\log(\beta) - \sum_i x_i/\beta.$$

Hence

$$l'(\beta) = -\frac{3n}{\beta} + \frac{\sum_i x_i}{\beta^2} \text{ and } l''(\beta^2) = \frac{3n}{\beta^2} - \frac{2\sum_i x_i}{\beta^3}.$$

Setting $l'(\beta) = 0$

$\Rightarrow \hat{\beta} = \frac{\sum_i x_i}{3n}$,

so $\sum_i x_i = 60.9 \Rightarrow \hat{\beta} = 1.015$. Since $X_i = T_i - 10$.

The second derivative of the log likelihood at $\hat{\beta}$ is -58.2 which confirms that a maximum has been located.

Answer 6

Finding the MLE:

Model: $X_1, X_2, \ldots, X_n$ independent, with each $X_i \sim \text{Ga}(2, \theta)$

Data: $x_1, x_2, \ldots, x_n$

$L(\theta) \propto \prod_{i=1}^{n} \theta^2 x_i \, e^{-\theta x_i} = \theta^{2n} \, e^{-\theta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i$

$\ell(\theta) = 2n \log_e \theta - \theta \sum_{i=1}^{n} x_i + K$

$\ell'(\theta) = \frac{2n}{\theta} - \sum_{i=1}^{n} x_i$

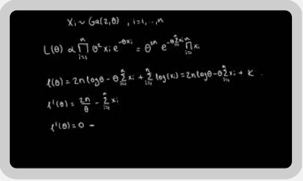
$\ell'(\theta) = 0 \text{ when } \theta = \frac{2n}{\sum_{i=1}^{n} x_i} = \frac{2}{\bar{x}}$

$\ell''(\theta) = -\frac{2n}{\theta^2}$

and this is < 0 for all $\theta > 0$.

Therefore, $\hat{\theta}_{MLE} = \frac{2}{\bar{x}}$

Video



$X_i \sim Ga(2,\theta)$, $i=1,\dots,n$

$L(\theta) \propto \prod_{i=1}^{n} \theta^2 x_i e^{-\theta x_i} = \theta^{2n} e^{-\theta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i$

$\ell(\theta) = 2n \log\theta - \theta \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log(x_i) = 2n\log\theta - \theta \sum_{i=1}^{n} x_i + K$

$\ell'(\theta) = \frac{2n}{\theta} - \sum_{i=1}^{n} x_i$

$\ell'(\theta) = 0 -$

Video model answers for task 6

Duration 3:38

Answer 7

Insects:

```
## The data
Insects <- c(rep(1,26), rep(2, 27), rep(3, 18), rep(4, 18),
rep(5, 8), 6, 6, 7)

## Find the sample size:
n <- length(Insects)

## Set up a range of values for
lambda <- seq(2.2, 2.8, length=100)

## Compute the log-likelihood (ignoring constants, i.e.
anything that
## doesn't depend on lambda):
LogLik <- (-n*lambda) + (sum(Insects)*log(lambda)) - (n*log(1-
exp(-lambda)))

## Plot the log-likelihood:
```

```
plot(lambda, LogLik, type="l", lwd=3, ylab="log-likelihood",
xlab=expression(lambda))
```

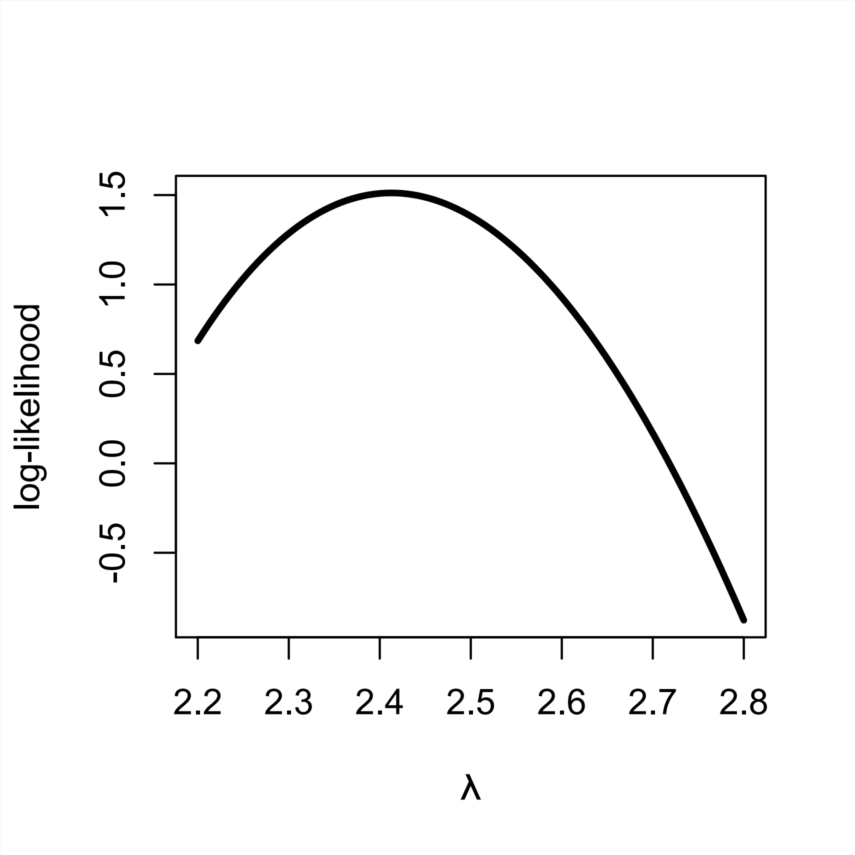


Figure 4

The plot suggest that a plausible initial estimate of lambda is around 2.41 or 2.42.

```
## Construct a function for the log-likelihood:
myfunction <- function(lambda, y, n){
  (-n*lambda) + (sum(y)*log(lambda)) - (n*log(1-exp(-lambda)))
}

## Optimise the function to estimate the parameter:
optim(par=2.4, fn=myfunction,
method="BFGS",control=list(fnscale= -1),
           y=Insects, n=100)
```

R Console

```
$par
[1] 2.41261
```

```
$value
[1] 1.511979

$counts
function gradient
13         3

$convergence
[1] 0

$message
NULL
```

From the `R` output, the function has converged to a parameter value for λ of 2.41.

(Note: a constant, K, has not been included in the log-likelihood for simplicity since this doesn't change the position (λ) of the MLE).

Footnotes

1. Many thanks to Suzy Whoriskey for all her contributions to the development of the course material ↩

2. See supplementary material from the R programming/Statistical computing course provided on the Learning from Data/Data Science Foundations Moodle pages for more details here. ↩