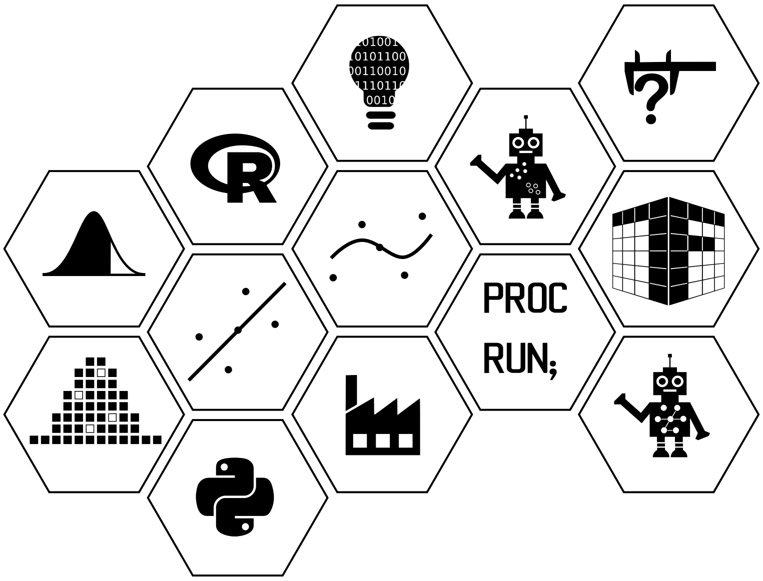
Learning from Data/Data Science Foundations

Week 8: Likelihood III - the normal distribution and combining likelihoods



DATA ANALYTICS GLASGOW

Maximum likelihood estimation: the normal distribution and combining likelihoods

In weeks 5 and 6 we considered maximum likelihood estimation and properties of point estimators for one parameter discrete and continuous distributions. This week we will firstly focus on extending these ideas to the normal distribution (a multiparameter distribution), secondly to combine likelihoods from multiple independent populations, and then we'll consider properties for maximum likelihood estimators in general.

Week 8 learning material aims

The material in week 8 covers:

- maximum likelihood for the normal distribution and multiple independent populations;
- the Hessian matrix and hence sample information for multiparameter models;
- properties of Maximum Likelihood Estimators.

Likelihood for multiparameter distributions

So far only single parameter models have been considered, but the method of likelihood works in a very similar way when a model involves several parameters.

The first video[1] for this week goes through the maximum likelihood steps for example 1 below on the normal distribution:

Video

The normal distribution - likelihood estimators

Duration 11:27

Multiparameter likelihood

Claire Miller, Eilidh Jack and Suzy Whoriskey

Example 1

Normal model

Annual mean temperatures in New Haven

As a simple example consider the 60 year record of mean annual temperature in New Haven, Connecticut.

49.9 52.3 49.4 51.1 49.4 47.9 49.8 50.9 49.3 51.9 50.8 49.6 49.3 50.6 48.4 50.7 50.9 50.6 51.5 52.8 51.8 51.1 49.8 50.2 50.4 51.6 51.8 50.9 48.8 51.7 51.0 50.6 51.7 51.5 52.1 51.3 51.0 54.0 51.4 52.7 53.1 54.6 52.0 52.0 50.9 52.6 50.2 52.6 51.6 51.9 50.5 50.9 51.7 51.4 51.7 50.8 51.9 51.8 51.9 53.0

First of all we can consider a histogram for these data in `R`:

```
airtemp <- c(49.9, 52.3, 49.4, 51.1, 49.4, 47.9, 49.8, 50.9, 49.3, 51.9,
50.8, 49.6, 49.3, 50.6,
48.4, 50.7, 50.9, 50.6, 51.5, 52.8, 51.8, 51.1, 49.8, 50.2, 50.4, 51.6,
51.8, 50.9,
48.8, 51.7, 51.0, 50.6, 51.7, 51.5, 52.1, 51.3, 51.0, 54.0, 51.4, 52.7,
53.1, 54.6,
52.0, 52.0, 50.9, 52.6, 50.2, 52.6, 51.6, 51.9, 50.5, 50.9, 51.7, 51.4,
51.7, 50.8,
51.9, 51.8, 51.9, 53.0)

h <- hist(airtemp, main="", xlab="Temperature, oC")
```

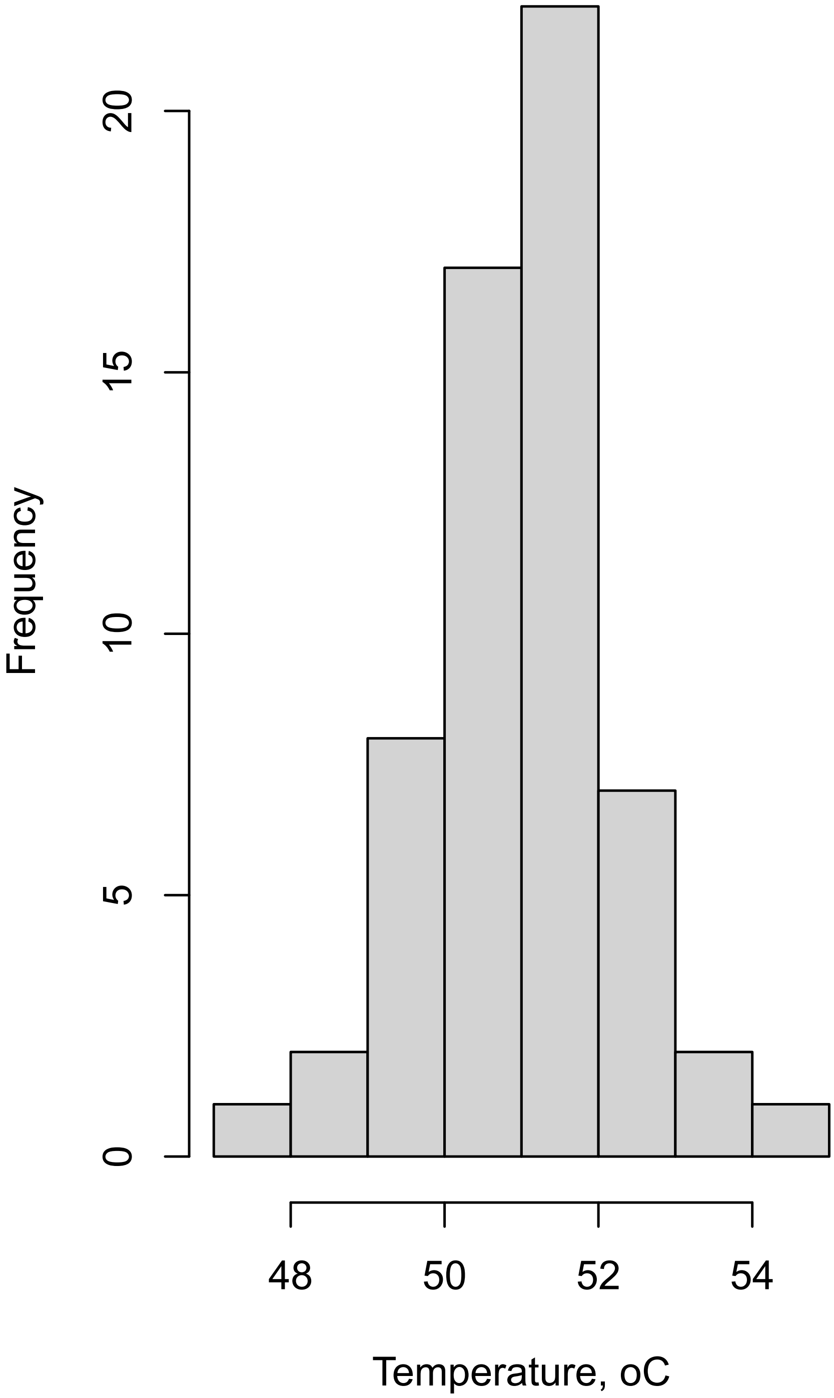


Figure 1

The histogram illustrates that a normal probability model may be reasonable for these data since the histogram appears to follow the shape of a bell-shaped curve. We would then treat the data as observations of independent identically distributed random variables X_i each with p.d.f.

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \quad i = 1, \ldots, n,$$

where μ and σ are unknown parameters, $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

Remember that when we are using likelihood we are aiming to find the best estimators (and hence estimates) for our parameters of our assumed distribution given the data that we have observed. The first video for this week (see page 2) gives an illustration of this for the normal distribution.

So let's find the maximum likelihood estimators (and hence estimates for μ and σ^2) for the normal distribution, and for the New Haven temperature data example.

Likelihood

Since the normal distribution is a continuous distribution our likelihood is proportional to the product of our p.d.fs for each data point, and we have:

$$L(\mu, \sigma) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}.$$

The **log-likelihood** is then

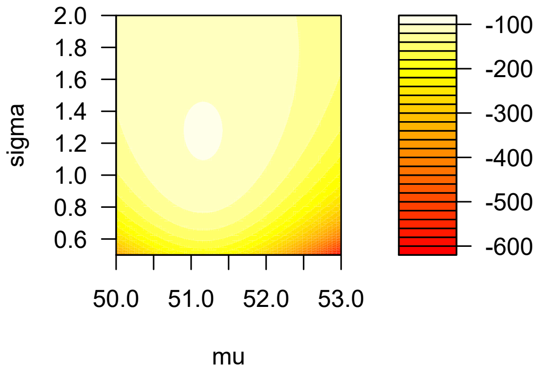
$$\ell(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2).$$

See the first video for this week on page 2 for more details of the log-likelihood derivation.

A contour plot of the log-likelihood function is a useful way of displaying the function.

We can use the following code to do this in `R`

```
sigma <- seq(0.5,2,length=100) ## a sequence of values for sigma
mu <- seq(50,53, length=100) ## a sequence of values for mu
n <- length(airtemp) ## the sample size

## create an empty matrix to store the elements of the log-likelihood

loglik <- matrix(NA, 100, 100)

## compute values of the log-likelihood for the sequence of sigma and mu
values
for (i in 1:100){
  for (j in 1:100){

    loglik[j,i] <- -((n/2)*log(2*pi))-(n*log(sigma[i]))-(sum((airtemp-
mu[j])^2)/(2*((sigma[i])^2)))
  }
}

## plot the log-likelihood
filled.contour(mu, sigma, loglik, color.palette=heat.colors, xlab="mu",
ylab="sigma")
```



We can identify the maximum likelihood estimates simply by examining this function in the figure carefully: :inlineMath[10].

Supplement 1

Visualising likelihood

The following R code can be used to visualise a 3D surface of the log-likelihood for the normal distribution and the temperature in New Haven example over a range of values for both μ and σ.

```
library(rpanel)
rp.likelihood("sum(log(dnorm(data, theta[1], theta[2])))",
airtemp, c(50, 0.5), c(53, 2))
```

The first argument in the function is simply an alternative way to write the log-likelihood function i.e. it can be written as the log of the product of p.d.fs, or (using log rules) the sum of the log of the p.d.fs. the second argument contains our data and the final arguments provide minimum and maximum values (respectively) for the two parameters of interest μ and σ.

Maximisation of $\ell()$ follows the same approach as in the single parameter case.

First differentiate ℓ w.r.t. each of the parameters:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),\\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

Here we are using partial differentiation\footnote{see the Preliminary Maths course for more details here}. There is a full explanation of the derivation in the first video for this week (see page 2).

Now we set both these derivatives to zero and **solve** the resulting pair of simultaneous equations for μ and σ:

$$\begin{aligned}\mu &= \bar{x},\\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

Likelihood estimates

$$\begin{aligned}\hat{\mu} &= \bar{x} = 51.16\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = 1.575^{**}\end{aligned}$$

** Note: that when we perform maximum likelihood estimation we get a slightly different estimator for $\hat{\sigma}^2$ than the one we introduced for s^2 in the learning material for week 2. Our original definition for s^2 is the unbiased estimator and the expression for $\hat{\sigma}^2$ from maximum likelihood is a biased estimator. This is not necessarily problematic. For now, it's just something to be aware of and you will learn more about this in later courses. Our original definition of $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is actually referred to as the REstricted Maximum Likelihood (REML) estimator.

The first video for this week (see page 2) shows an example of the fitted normal distribution curve compared to the data.

Checking that we have found MLEs

We must also check that the estimates are **maximum** likelihood estimates. This involves evaluating the second derivative matrix, or **Hessian (H)**, of $\ell()$ w.r.t. each of the parameters at $\hat{\mu}, \hat{\sigma}$.

The second video for this week goes through this process in detail for this example:

Video

The Hessian matrix

Duration 14:47

The Hessian matrix

Claire Miller, Eilidh Jack and Suzy Whoriskey

The **Hessian** matrix is:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{pmatrix}_{\hat{\mu}, \hat{\sigma}}$$

For the turning point at $\hat{\mu}, \hat{\sigma}$ to be a maximum, $\mathbf{H}$ must be *negative definite*. This means that the quantity $\mathbf{a}^{\mathbf{T}}\mathbf{H}\mathbf{a}$ is negative for every (non-zero) vector $\mathbf{a}$. (In other words, the second derivative in every direction is negative.)

Therefore,

$$\begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ell}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{2}{\sigma^3}\sum(x_i-\mu) \\ -\frac{2}{\sigma^3}\sum(x_i-\mu) & \frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum(x_i-\mu)^2 \end{pmatrix}$$

$$\begin{aligned} \mathbf{H} &= \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{2}{\sigma^3}\sum(x_i-\mu) \\ -\frac{2}{\sigma^3}\sum(x_i-\mu) & \frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum(x_i-\mu)^2 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{\hat\sigma^2} & -\frac{2}{\hat\sigma^3}\sum(x_i-\bar x) \\ -\frac{2}{\hat\sigma^3}\sum(x_i-\bar x) & \frac{n}{\hat\sigma^2}-\frac{3}{\hat\sigma^4}\sum(x_i-\bar x)^2 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{\hat\sigma^2} & 0 \\ 0 & \frac{n}{\hat\sigma^2}-\frac{3}{\hat\sigma^4}n\hat\sigma^2 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{\hat\sigma^2} & 0 \\ 0 & \frac{n}{\hat\sigma^2}-\frac{3n}{\hat\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{\hat\sigma^2} & 0 \\ 0 & \frac{-2n}{\hat\sigma^2} \end{pmatrix} \end{aligned}$$

For any (non-zero) vector $\mathbf{a} = (a_1, a_2)^T$, the quantity $\mathbf{a^T H a} = -(a_1^2\frac{n}{\hat\sigma^2} + a_2^2\frac{2n}{\hat\sigma^2})$. The matrix $\mathbf{H}$ is therefore negative definite and so in this case we do have *maximum* likelihood estimates.

Supplement 2

In the course **predictive modelling** you have seen how to use least squares to estimate the parameters in a regression model. For example, suppose for a response Y and explanatory variable X, we have data $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ and we are interested to fit the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (1)$$

where $\epsilon_i \sim N(0, \sigma^2)$. An alternative approach to estimating β_0 and β_1 is to use maximum likelihood and this can be done here by modifying the above expressions slightly. Another way to express model (1) above is in terms of $\mathbb{E}(Y|X) = \beta_0 + \beta_1 x_i = \mu$, i.e. $Y|X \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Therefore, to estimate the parameters here using likelihood we would have:

$$L(\beta_0, \beta_1, \sigma) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right\}.$$

These ideas will be extended in the course **Advanced Predictive Models**.

Supplement 3

In week 7 we introduced results to provide us with approximate confidence intervals for parameters in one parameter distributions. The first of these was the Wilks interval using the idea of a likelihood interval. We saw that a 14.65% likelihood interval provides us with an approximate 95% confidence interval for the parameter of any such distribution. Now that we have derived the likelihood for the normal distribution, we can show that in the case of the normal distribution these intervals are exact for μ when σ is known. This derivation also helps us to make the connection to the distributional result that we utilised in week 7 based on the χ^2 distribution.

See the material at Likelihood intervals supplement for details here.

Combining Likelihoods

We can also be in the situation that we have more than one parameter to estimate when we are interested in comparing a question of interest for more than one population. Let's consider the following example.

Example 2

In an experiment to investigate the bacteria present in the home, 15 households had the number of bacteria (in 100,000's) recorded from a sample taken from the kitchen sink (control surface) and 15 different housesholds had the number of bacteria recorded from a sample taken from the kitchen sink immediately after cleaning (treated surface). The data are shown below.

```
Control surface: 37 35 31 36 25 43 41 33 25 37 27 30 32 35 38

Treated surface: 4 4 3 3 3 5 3 4 5 4 4 3 4 4 3
```

We will assume that these data follow independent Poisson probability models.

In this example, we have data for two independent samples X, Y, e.g. ($X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$) and hence the joint probability distribution, using the property of independent probabilities from the course [*Probability and Stochastic Models* and *Probability and Sampling Fundamentals/Sampling Fundamentals*], is:

$$P(X,Y;\boldsymbol{\lambda}) = P(X;\lambda_1)P(Y;\lambda_2).$$

The likelihoods therefore combine in a very simple way:

$$L(\boldsymbol{\lambda}) = L_1(\lambda_1)L_2(\lambda_2),$$

where L_1 and L_2 denote the likelihood functions of the two separate samples.

Similarly, the log-likelihood functions combine as

$$\ell(\boldsymbol{\lambda}) = \ell_1(\lambda_1) + \ell_2(\lambda_2).$$

For example 2 above, we can define the random variables $X_1, \ldots, X_{15}$ and $Y_1, \ldots, Y_{15}$ to describe the observed values for the control and treated surfaces respectively. The model is that the X_i's and Y_i's are all independent and that $X_i \sim \mathrm{Poi}(\lambda_1)$ while $Y_i \sim \mathrm{Poi}(\lambda_2)$, where λ_1 and λ_2 are the mean numbers of bacteria for control and treated surfaces respectively.

Likelihood

$$\begin{aligned} L(\boldsymbol{\lambda};\mathbf{x},\mathbf{y}) &= K_1\ K_2 \prod_{i=1}^{15} e^{-\lambda_1}\ \lambda_1^{x_i} \prod_{i=1}^{15} e^{-\lambda_2}\ \lambda_2^{y_i} \\ &= K_1\ K_2 e^{-15\lambda_1}\ \lambda_1^{\sum_{i=1}^{15} x_i} e^{-15\lambda_2}\ \lambda_2^{\sum_{i=1}^{15} y_i} \end{aligned}$$

where K_1 and K_2 are constants that don't depend on λ_1 or λ_2.

Log-likelihood

$$\begin{aligned} \ell(\boldsymbol{\lambda};\mathbf{x},\mathbf{y}) &= -15\lambda_1 + \sum_{i=1}^{15} x_i \log(\lambda_1) - 15\lambda_2 + \sum_{i=1}^{15} y_i \log(\lambda_2) \\ &= -15\lambda_1 + 505\log(\lambda_1) - 15\lambda_2 + 56\log(\lambda_2) \end{aligned}$$

ignoring the constants, which will disappear when we differentiate. The sums of the observations in the control and treated groups are 505 and 56 respectively.

Partial derivatives

The first derivatives are

$$\frac{\partial \ell}{\partial \lambda_1} = -15 + \frac{505}{\lambda_1} = 0 \text{ when } \lambda_1 = 505/15 = 33.67$$

and

$$\frac{\partial \ell}{\partial \lambda_2} = -15 + \frac{56}{\lambda_2} = 0 \text{ when } \lambda_2 = 56/15 = 3.73.$$

Hessian matrix

The partial derivates are

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial {\lambda_1}^2} &= -\frac{505}{\lambda_1^2} \\
\frac{\partial^2 \ell}{\partial {\lambda_2}^2} &= -\frac{56}{\lambda_2^2} \\
\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} &= 0.
\end{aligned}$$

Therefore,

$$\mathbf{H} = \begin{pmatrix} -\frac{505}{\lambda_1^2} & 0 \\ 0 & -\frac{56}{\lambda_2^2} \end{pmatrix}.$$

It follows that the Hessian matrix is negative definite, as it is diagonal with only negative entries. We have therefore established that the solutions to the likelihood equations are indeed the MLEs.

Sample information

As we saw in weeks 6 & 7, the second derivative information can be used to provide information on the uncertainty in our parameter estimates, and we'll return to this idea at the end of this week's material. Here our sample information is a matrix $\mathbf{K}(\mathbf{x})$ defined by $-\mathbf{H}$, and evaluated at our MLEs.

Task 1

Evaluate the sample information matrix $\mathbf{K}(\mathbf{x})$ for example 2 above at the MLEs.

Informally, it therefore appears as though there is a large difference between the bacteria present on the two surfaces, since there is a large difference between λ_1 and λ_2. However, we will re-visit this in weeks 9 & 10 where we'll re-visit interval estimation, and introduce, hypothesis testing for likelihood to formally assess this.

Task 2

Patients with high blood pressure were randomly allocated to receive one of two treatments. The patients had been treated with cognitive behaviour therapy (CBT) and were then given no further treatment (NFT) or CBT and beta-blockers (BB). Six weeks later it was noted whether or not each of the patients showed a decrease in their blood pressure. For the NFT group 15 out of 50 patients showed a decrease after 6 weeks, whereas this happened for 26 out of 50 of the BB patients.

1. Formulate the likelihood and log-likelihood functions;
2. Find the likelihood estimates;
3. Check that you have found maximum likelihood estimates;
4. What can we say about the probabilities of blood pressure decrease in the two patient groups?

(Hint: assume a Binomial distribution for each group, with individual sample sizes n_i and individuals parameters θ_i)

Maximising likelihood functions numerically

It was shown earlier that sometimes the likelihood functions cannot be solved algebraically. We introduced the Newton-Raphson algorithm to solve the likelihood functions numerically.

These ideas can be extended to vector parameters. For **vector parameters** in the Newton-Raphson method, good initial estimates are required, as in the one parameter case, for each of the parameters in the model. The following formula is then used to find successively better estimates; the iterations proceed until convergence.

The iterative formula is:

$$\hat{\boldsymbol{\theta}}^{(j+1)} = \hat{\boldsymbol{\theta}}^{(j)} - \mathbf{H}^{-1}\mathbf{g}.$$

where,

$$\mathbf{g} = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1}\ \hat{\boldsymbol{\theta}}^j \\ \frac{\partial \ell}{\partial \theta_2}\ \hat{\boldsymbol{\theta}}^j \\ . \\ . \end{pmatrix} \text{ and } \mathbf{H} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2}\ \hat{\boldsymbol{\theta}}^j & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\ \hat{\boldsymbol{\theta}}^j & . & . \\ \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\ \hat{\boldsymbol{\theta}}^j & \frac{\partial^2 \ell}{\partial \theta_2^2}\ \hat{\boldsymbol{\theta}}^j & . & . \\ . & . & . & . \\ . & . & . & . \end{pmatrix}.$$

As we have seen in week 6, in practice the `optim()` function in `R` is a very powerful command that can be used for numerical optimisation.

Task 3

For the air temperature data in example 1, check that you can also find the parameter estimates for μ and σ using the `optim` function in `R`.

Properties of MLEs

The preceding sections have shown how very general and powerful the method of maximum likelihood estimation is. In this section we will consider some of the theoretical properties of maximum likelihood estimates, which strengthen the case for using this method and the associated techniques that we have introduced in previous weeks. Many of these properties were introduced earlier for point estimators in general.

The main properties of Maximum Likelihood Estimators are:

- Invariance
- Consistency
- Asymptotic[2] normality
- Asymptotic Efficiency - roughly this means that among all well-behaved estimators, the MLE has the smallest variance (at least for large samples).

Several of these will be discussed in more detail below. However, only the main results are provided. Outline proofs of the results are provided in supplementary material (see page 12 for a link to this material). All that is required here is that you are aware of the ideas of these results. **You will not be asked to state the details of the results or to prove the results in the class test. Example 4 is an important example though applying these results, which we'll use the ideas of again in revision tasks and later weeks.**

Invariance

Consider an observation $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top$ of a vector of random variables with joint p.m.f. or p.d.f. $f(\mathbf{x}, \theta)$, where θ is a parameter with MLE $\hat{\theta}$. If β is a parameter such that $\beta = g(\theta)$ where g is any function, then the maximum likelihood estimate of β is $\hat{\beta} = g(\hat{\theta})$, and this property is known as *invariance*. So, when working with maximum likelihood estimation, we can adopt whatever parameterization is most convenient for performing calculations, and simply transform back to the most interpretable parameterization at the end.

Example 3

Invariance

For example, suppose we are interested in finding an estimator and hence estimate for the parameter β in a probability distribution, and we know that there is a relationship between the parameter β and another parameter (say) θ e.g. that $\beta = \frac{1}{\theta}$. If it is easier for us to find the estimate of θ, then we can do this first using maximum likelihood estimation to obtain $\hat{\theta}_{MLE}$ and then compute $\hat{\beta}_{MLE} = \frac{1}{\hat{\theta}_{MLE}}$ in order to estimate β.

Consistency

Maximum likelihood estimators are often not unbiased (and we saw an example of this for $\hat{\sigma}^2$ for the normal distribution in example 1), but under quite mild regularity conditions they are *consistent*. This means that as the sample size on which the estimate is based tends to infinity the maximum likelihood estimator tends in probability to the true parameter value. Consistency therefore implies asymptotic unbiasedness, but it actually implies slightly more than this, for example that the variance of the estimator is decreasing with sample size.

Properties of the expected log-likelihood

The following results explain the approach of maximum likelihood estimation, that we've been using to find estimates for our parameters, through illustrating that the **expected log-likelihood has a maximum at the true parameter value.**

Let $x_1, x_2, \ldots, x_n$ be independent observations from a p.d.f. $f(\mathbf{x}, \theta)$ where θ is an unknown parameter with true value θ_T. Treating θ as unknown, the likelihood and log-likelihood for θ are:

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

$$\ell(\theta) = \log_e\left(\prod_{i=1}^{n} f(x_i, \theta)\right) = \sum_{i=1}^{n} \log_e[f(x_i, \theta)] = \sum_{i=1}^{n} \ell_i(\theta),$$

where ℓ_i is the log-likelihood given only the single observation x_i. Treating ℓ as a function of random variables $X_1, X_2, \ldots, X_n$ means that ℓ is itself a random variable (and the ℓ_i are independent random variables). Hence we can consider expectations of ℓ and its derivatives.

Result 1:

$$\mathbb{E}_T\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right) = 0,$$

where the subscript (T) on the expectation is to emphasize that the expectation is w.r.t. $f(x, \theta_T)$. Result 1 has the following obvious consequence in Result 2, since $\left(\mathbb{E}_T\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right)\right)^2 = 0:$

Result 2:

$$\text{Var}\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right) = \mathbb{E}_T\left[\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right)^2\right] = I_\theta.$$

(Remember, Var(X) = $\mathbb{E}(X^2) - (\mathbb{E}(X))^2$)

In week 6 it was stated that I_θ is Fisher's information and that:

Result 3:

$$I_\theta \equiv \mathbb{E}_T\left[\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right)^2\right] = -\mathbb{E}_T\left[\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta_T}\right].$$

If the data tie down θ very closely (and accurately) then the log-likelihood will be sharply peaked in the vicinity θ_T (i.e. high I_θ), whereas data containing little information about θ will lead to an almost flat likelihood and low I_θ.

Now notice that result 1 says that the expected log-likelihood has a turning point at θ_T, while since I_θ is non-negative, result 3 indicates that this turning point is a maximum.

Therefore, these results show that the **expected log-likelihood has a maximum at the true parameter value.**

The results generalize immediately to vector parameters. In this case result 3 is:

Result 3 (vector parameter)

$$\mathbf{I}_\theta \equiv \mathbb{E}_T\begin{pmatrix} \left(\frac{\partial \ell}{\partial \theta_1}\right)^2 & \frac{\partial \ell}{\partial \theta_1}\frac{\partial \ell}{\partial \theta_2} & . \\ \frac{\partial \ell}{\partial \theta_2}\frac{\partial \ell}{\partial \theta_1} & \left(\frac{\partial \ell}{\partial \theta_2}\right)^2 & . \\ . & . & . \end{pmatrix} = -\mathbb{E}_T\begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & . \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_1^2} & . \\ . & . & . \end{pmatrix}.$$

Large sample distribution of $\hat{\theta}$

In the large sample limit (as $n \to \infty$) we can establish the following pivotal function for θ, which will enable us to compute approximate confidence intervals in later weeks:

$$\left(\hat{\theta} - \theta_T\right) \sim N(0, I_\theta^{-1}).$$

The result generalizes to vector parameters:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \mathbf{I}_\theta^{-1}),$$

in the large sample limit.

What to look for in a good estimator:

We introduced the **Cramer-Rao lower bound** on the variance of an unbiased estimator for properties of point estimators in week 6.

Let θ be a parameter and $\hat{\theta}$ an **unbiased estimator** of θ, meaning that $\mathbb{E}(\hat{\theta}) = \theta$.

Then, in summary, the Cramer-Rao result says that the variance of $\hat{\theta}$ can not be smaller than $I_\theta{}^{-1}$

- the inverse of the information about θ ($\boldsymbol{I_\theta}{}^{-1}$ in the vector parameter case).

This result offers rather strong support for the method of maximum likelihood estimation, for as we have seen above, in the large sample limit MLEs are unbiased and have exactly $I_\theta{}^{-1}$ variance.

Usually, of course, $\boldsymbol{I_\theta}$ will not be known any more than $\boldsymbol{\theta}$ is and will have to be estimated by plugging $\hat{\boldsymbol{\theta}}$ into the expression for $\boldsymbol{I_\theta}$. In fact, often the **sample information matrix** $\mathbf{K}(\mathbf{x})$, which is just the negative of the Hessian ($-\mathbf{H}$) of the log-likelihood evaluated at the MLE, is an adequate approximation to the information matrix $\boldsymbol{I_\theta}$ itself.

This provides:

$$\hat{\boldsymbol{\theta}}_{MLE} \sim N(\boldsymbol{\theta_T}, \mathbf{K}(\mathbf{x})^{-1}),$$

i.e. $\mathbf{K}(\mathbf{x})^{-1}$ provides the variance/covariance matrix for $\hat{\boldsymbol{\theta}}_{MLE}$ or in the one parameter case:

$$\hat{\theta}_{MLE} \sim N\left(\theta_T, \frac{1}{k(\mathbf{x})}\right),$$

and $1/k(\mathbf{x})$ provides the variance for $\hat{\theta}$, as we have seen in weeks 6 & 7.

Example 4

Return to example 2, on household bacteria. We can now estimate the variance for the parameters λ_1 and λ_2.

We found that, the Hessian matrix was:

$$\mathbf{H} = \begin{pmatrix} -\frac{505}{\lambda_1^2} & 0 \\ 0 & -\frac{56}{\lambda_2^2} \end{pmatrix},$$

and the sample information matrix was:

$$\mathbf{K(x)} = \begin{pmatrix} 0.4455 & 0 \\ 0 & 4.0250 \end{pmatrix}.$$

Since we have a diagonal matrix we can simply take the reciprocal of the diagonal elements\footnote{see the Preliminary Maths course for a reminder on inverting matrices}.

Therefore,

$$\mathbf{K(x)^{-1}} = \begin{pmatrix} 2.245 & 0 \\ 0 & 0.248 \end{pmatrix},$$

This gives:

Var($\hat{\lambda}_1$) = 2.245

Var($\hat{\lambda}_2$) = 0.248,

and hence we now have information on the uncertainty in our parameter estimates, which were $\hat{\lambda}_1 = 33.67$ and $\hat{\lambda}_2 = 3.73$, which we will be able to use for inference in weeks 9 & 10. (Note the off-diagonal terms here are 0. This is a result of the independence assumption between our two populations X and Y.)

Supplement 4

For more details on all of the above results and outline proofs see the material at MLE properties

Learning outcomes for week 8

By the end of week 8 you should be able to:

- apply the principle of maximum likelihood to obtain point estimates of parameters in multiparameter statistical models;
- define (and derive) the Hessian matrix and sample information for multiparameter models;
- state the properties of Maximum Likelihood Estimators.

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

Task 4

Random variables X and Y have joint p.d.f.

$$f(x,y) = (\alpha+1)(\beta+1)x^{\alpha}y^{\beta} \quad 0 \leq x \leq 1,\ 0 \leq y \leq 1$$

Assume that you have n independent pairs of observatons (x_i, y_i), $i = 1, \ldots, n$.

- Find the maximum likelihood estimators of α and β.
- Find expressions for the approximate variances of your estimators, in terms of $\hat{\alpha}$ and $\hat{\beta}$.

Task 5

Suppose that observations (x_{1i}, x_{2i}) are available for $i = 1, \ldots, n$ and that an appropriate model for these is that each pair is an independent observation of the random variables with joint p.d.f.

$$f(x_1, x_2) = \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2} \text{ where } x_1 > 0,\ x_2 > 0,\ \lambda_1 > 0,\ \lambda_2 > 0.$$

- Write down the likelihood and corresponding log-likeihood of $\lambda_1,\ \lambda_2$.
- Find the maximum likelihood estimates for λ_1 and λ_2.
- Obtain the Hessian matrix.

Task 6

A clinical trial was conducted to compare four drugs used to treat hypertension. Each patient was randomly allocated to receive one of the four drugs during the course of the trial. The numbers of side effects which were recorded for the patients during the course of the trial are given below.

Drug	**1**	**2**	**3**	**4**
Number of patients	29	27	29	26
Number of side effects	10	85	55	39

It may be assumed that the data are described well by independent Poisson probability models $X_i \sim \text{Po}(n_i \theta_i)$, where,

X_i denotes the number of side effects for the ith drug,

n_i denotes the number of patients on the ith drug,

θ_i denotes the mean number of side effects per patient on the ith drug.

The general model for these data is:

$$X_1 \sim \mathrm{Po}(29\theta_1),\ X_2 \sim \mathrm{Po}(27\theta_2),\ X_3 \sim \mathrm{Po}(29\theta_3),\ X_4 \sim \mathrm{Po}(26\theta_4)$$

and hence, since the models are independent, the likelihood is:

$$L(\theta, x) = K\theta_1^{10}e^{-29\theta_1}\theta_2^{85}e^{-27\theta_2}\theta_3^{55}e^{-29\theta_3}\theta_4^{39}e^{-26\theta_4}$$

Compute the Hessian matrix.

Answer 1

When the parameters in the negative Hessian are replaced by the MLEs, we obtain the sample information matrix as

$$\mathbf{K(x)} = \begin{pmatrix} \frac{505}{(33.67)^2} & 0 \\ 0 & \frac{56}{(3.73)^2} \end{pmatrix} = \begin{pmatrix} 0.4455 & 0 \\ 0 & 4.0250 \end{pmatrix}$$

Answer 2

An appropriate model for the data is

$$X_1 \sim \mathrm{Bi}(50, \theta_1)$$

$$X_2 \sim \mathrm{Bi}(50, \theta_2)$$

where X_1 and X_2 denote the numbers of patient experiencing a decrease in the NFT and BB groups respectively.

The likelihood function is

$$L(\theta_1, \theta_2; \mathbf{x}) = K_1\theta_1^{15}(1-\theta_1)^{35}K_2\theta_2^{26}(1-\theta_2)^{24}$$

where K_1 and K_2 are constants.

The log-likelihood function is

$$\ell(\theta_1, \theta_2; \mathbf{x}) = 15\log_e(\theta_1) + 35\log_e(1-\theta_1) + 26\log_e(\theta_2) + 24\log_e(1-\theta_2) + K$$

The likelihood equations are

$$\frac{15}{\theta_1} - \frac{35}{1-\theta_1} = 0$$

$$\frac{26}{\theta_2} - \frac{24}{1-\theta_2} = 0$$

The maximum likelihood esimates are

$$\hat{\theta}_1 = 15/50 = 0.30$$

$$\hat{\theta}_2 = 26/50 = 0.52$$

This is confirmed by the Hessian matrix, **H**

$$\begin{pmatrix} \frac{\partial^2\ell}{\partial\theta_1^2} & \frac{\partial^2\ell}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2\ell}{\partial\theta_1\partial\theta_2} & \frac{\partial^2\ell}{\partial\theta_2^2} \end{pmatrix} = \begin{pmatrix} -\frac{15}{\theta_1^2} - \frac{35}{(1-\theta_1)^2} & 0 \\ 0 & -\frac{26}{\theta_2^2} - \frac{24}{(1-\theta_2)^2} \end{pmatrix}$$

which is clearly negative-definite at all values of (θ_1, θ_2) in $0 < \theta_i < 1, i = 1, 2$.

The sample information matrix, $\mathbf{K(x)} = -\mathbf{H}$, evaluates this at the MLE's.

Notice that since $15 = 50\hat{\theta}_1$, the top left diagonal entry of $\mathbf{K(x)}$ is:

$$\begin{aligned} \frac{15}{\hat{\theta}_1^2} + \frac{35}{(1-\hat{\theta}_1)^2} &= \frac{50\hat{\theta}_1}{\hat{\theta}_1^2} + \frac{50(1-\hat{\theta}_1)}{(1-\hat{\theta}_1)^2} \\ &= \frac{50}{\hat{\theta}_1} + \frac{50}{(1-\hat{\theta}_1)} \\ &= \frac{50}{\hat{\theta}_1(1-\hat{\theta}_1)} \end{aligned}$$

Informally, it therefore appears as though there is a difference between the two treatment groups. However, we will need the methods that we will introduce in weeks 9 & 10 for further interval estimation and hypothesis testing for likelihood to formally assess this.

Answer 3

Air temperature example, using `optim`

```
## The data
airtemp <- c(49.9, 52.3, 49.4, 51.1, 49.4, 47.9, 49.8, 50.9,
49.3, 51.9, 50.8, 49.6, 49.3, 50.6,
48.4, 50.7, 50.9, 50.6, 51.5, 52.8, 51.8, 51.1, 49.8, 50.2,
50.4, 51.6, 51.8, 50.9,
48.8, 51.7, 51.0, 50.6, 51.7, 51.5, 52.1, 51.3, 51.0, 54.0,
51.4, 52.7, 53.1, 54.6,
52.0, 52.0, 50.9, 52.6, 50.2, 52.6, 51.6, 51.9, 50.5, 50.9,
51.7, 51.4, 51.7, 50.8,
51.9, 51.8, 51.9, 53.0)

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

## Construct a function for the log-likelihood:
ltemp <- function(x,y, n){
  -((n/2)*log(2*pi))-(n*log(x[2]))-(sum((y-x[1])^2)/(2*
((x[2])^2)))
}

## Optimise the function ltemp to estimate the parameters:
optim(par=c(51.1, 1.27), fn=ltemp,
method="BFGS",control=list(fnscale= -1), y=airtemp, n=60)
```

R Console

```
$par
[1] 51.160005  1.255017

$value
[1] -98.76524

$counts
function gradient
17         5

$convergence
```

```
[1] 0


$message
NULL
```

Answer 4

$$L(\alpha,\beta) \propto \prod_{i=1}^{n}(\alpha+1)(\beta+1)x_i^{\alpha}y_i^{\beta}$$

$$L(\alpha,\beta) \propto (\alpha+1)^n(\beta+1)^n\prod_{i=1}^{n}x_i^{\alpha}\prod_{i=1}^{n}y_i^{\beta}$$

$$\ell(\alpha,\beta) = n\log_e(\alpha+1) + n\log_e(\beta+1) + \sum\alpha\log_e(x_i) + \sum\beta\log_e(y_i)$$

$$\frac{\partial\ell}{\partial\alpha} = \frac{n}{\alpha+1} + \sum\log_e(x_i) \quad \frac{\partial\ell}{\partial\beta} = \frac{n}{\beta+1} + \sum\log_e(y_i)$$

Setting both these to zero and solving yields:

$$\hat{\alpha} = \frac{-n}{\sum\log_e(x_i)} - 1 \quad \hat{\beta} = \frac{-n}{\sum\log_e(y_i)} - 1$$

$$\frac{\partial^2\ell}{\partial\alpha\partial\beta} = 0 \; \frac{\partial^2\ell}{\partial\beta^2} = \frac{-n}{(\beta+1)^2} \; \frac{\partial^2\ell}{\partial\alpha^2} = \frac{-n}{(\alpha+1)^2}$$

$$\Rightarrow \operatorname{var}(\hat{\alpha}) \simeq (\hat{\alpha}+1)^2/n \quad \operatorname{var}(\hat{\beta}) \simeq (\hat{\beta}+1)^2/n.$$

Answer 5

$$L(\boldsymbol{\lambda}) \propto \prod_{i=1}^{n} \lambda_1 \lambda_2 e^{-\lambda_1 x_{1i} - \lambda_2 x_{2i}} = \lambda_1^n \lambda_2^n e^{-\lambda_1 \sum x_{1i} - \lambda_2 \sum x_{2i}}$$

$$\ell(\boldsymbol{\lambda}) = n \log_e \lambda_1 + n \log_e \lambda_2 - \lambda_1 \sum x_{1i} - \lambda_2 \sum x_{2i}$$

$$\frac{\partial \ell}{\partial \lambda_j} = \frac{n}{\lambda_j} - \sum_i x_{ji} \quad j = 1, 2$$

Setting, $\frac{\partial \ell}{\partial \lambda_j} = 0$

$$\Rightarrow \hat{\lambda}_j = \frac{n}{\sum_i x_{ji}} \quad j = 1, 2.$$

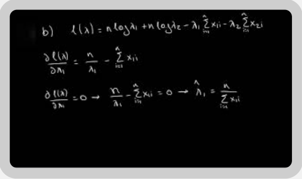
$$\frac{\partial^2 \ell}{\partial \lambda_j^2} = \frac{-n}{\lambda_j^2} \quad j = 1, 2$$

$$\boldsymbol{H} = \begin{pmatrix} -n/\lambda_1^2 & 0 \\ 0 & -n/\lambda_2^2 \end{pmatrix}$$

Video

Video model answers

Duration 7:36



b) $\ell(\lambda) = n \log \lambda_1 + n \log \lambda_2 - \lambda_1 \sum_{i=1}^{n} x_{1i} - \lambda_2 \sum_{i=1}^{n} x_{2i}$

$\frac{\partial \ell(\lambda)}{\partial \lambda_1} = \frac{n}{\lambda_1} - \sum_{i=1}^{n} x_{1i}$

$\frac{\partial \ell(\lambda)}{\partial \lambda_1} = 0 \rightarrow \frac{n}{\lambda_1} - \sum_{i=1}^{n} x_{1i} = 0 \rightarrow \hat{\lambda}_1 = \frac{n}{\sum_{i=1} x_{1i}}$

Answer 6

The derivatives are

$$\ell'(\theta_1) = 10/\theta_1 - 29 \quad \text{etc.}\ldots.$$

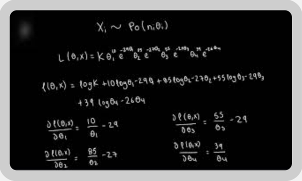
$\boldsymbol{H}$ provides the matrix of second derivatives....

$$\boldsymbol{H} = \begin{pmatrix} -10/\theta_1^2 & 0 & 0 & 0 \\ 0 & -85/\theta_2^2 & 0 & 0 \\ 0 & 0 & -55/\theta_3^2 & 0 \\ 0 & 0 & 0 & -39/\theta_4^2 \end{pmatrix}$$

Video

Video model answers

Duration 6:21



$X_i \sim Po(n_i\theta_i)$

$L(\theta,x) = K\theta_1^{10}e^{-29\theta_1}\theta_2^{85}e^{-27\theta_2}\theta_3^{55}e^{-29\theta_3}\theta_4^{39}e^{-26\theta_4}$

$\ell(\theta,x) = \log K + 10\log\theta_1 - 29\theta_1 + 85\log\theta_2 - 27\theta_2 + 55\log\theta_3 - 29\theta_3$
$+ 39\log\theta_4 - 26\theta_4$

$\frac{\partial \ell(\theta,x)}{\partial\theta_1} = \frac{10}{\theta_1} - 29$

$\frac{\partial \ell(\theta,x)}{\partial\theta_2} = \frac{85}{\theta_2} - 27$

$\frac{\partial \ell(\theta,x)}{\partial\theta_3} = \frac{55}{\theta_3} - 29$

$\frac{\partial \ell(\theta,x)}{\partial\theta_4} = \frac{39}{\theta_4}$

Footnotes

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

[2] 'asymptotic' here meaning 'as sample size tends to infinity'. ↩