Week 12: Maximum Likelihood Estimation

MATH697

Sahir Bhatnagar

November 21, 2017

McGill University

Introduction

Statistical Analyses

- Statistical analysis involves the informal/formal comparison of **hypothetical or predicted behaviour** with **experimental results**

Statistical Analyses

- Statistical analysis involves the informal/formal comparison of **hypothetical or predicted behaviour** with **experimental results**
- For example, we wish to be able to compare the predicted outcomes of an experiment, and the corresponding probability model, with a data histogram. We will use both *qualitative* and *quantitative* approaches.

- Suppose that an experiment or **trial** is to be repeated n times under **identical conditions**. Let X_i be the random variable corresponding to the outcome of the ith trial, and suppose that each of the n random variables $X_1, \dots, X_n$ takes values in sample space Ω.

- Suppose that an experiment or **trial** is to be repeated n times under **identical conditions**. Let X_i be the random variable corresponding to the outcome of the ith trial, and suppose that each of the n random variables $X_1, \dots, X_n$ takes values in sample space Ω.
- Often, assumptions can reasonably be made about the experimental conditions that lead to simplifications of the joint probability model for the random variables. Essentially, the assumption of **identical experimental conditions** for each of the n trials implies that the random variables corresponding to the trial outcomes are **identically distributed**, that is, in the usual notation, the PMF/PDF of X_i is denoted $f(x)$ dropping the subscript on the function f.

- Another common assumption is that the random variables $X_1, \dots, X_n$ are **independent**. Thus $X_1, \dots, X_n$ are usually treated as **i.i.d.** random variables.

- Another common assumption is that the random variables $X_1, \ldots, X_n$ are **independent**. Thus $X_1, \ldots, X_n$ are usually treated as **i.i.d.** random variables.
- In practice, it is commonly assumed that f takes one of the familiar forms (*Binomial*, *Poisson*, *Exponential*, *Normal* etc.).

- Another common assumption is that the random variables $X_1, \ldots, X_n$ are **independent**. Thus $X_1, \ldots, X_n$ are usually treated as **i.i.d.** random variables.
- In practice, it is commonly assumed that f takes one of the familiar forms (*Binomial*, *Poisson*, *Exponential*, *Normal* etc.).
- Thus f depends on one or more parameters ($\theta, \lambda, (\mu, \sigma)$ etc.). The role of these parameters could be indicated by re-writing the function $f(x)$ as

$$f(x) \equiv f(x; \theta) \qquad x \in \Omega \tag{1}$$

where θ here is a **parameter**, which may possibly be vector-valued.

- It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$.

- It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$.
- In the general case represented by (1) above, we have parameter $\theta \in \Theta$ where Θ is some subset of $\mathbb{R}^d$ and $d = 1, 2$, say, is the number of parameters. We refer to Θ as the **parameter space**. In practice, of course, parameter θ is **unknown** during the experiment

- It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$.
- In the general case represented by (1) above, we have parameter $\theta \in \Theta$ where Θ is some subset of $\mathbb{R}^d$ and $d = 1, 2$, say, is the number of parameters. We refer to Θ as the **parameter space**. In practice, of course, parameter θ is **unknown** during the experiment
- For a Normal model, we have $\Theta \equiv (\mu, \sigma^2)$

Objectives of a Statistical Analysis

- After the experiment has been carried out, a sample of **observed data** will have been obtained. Suppose that we have observed outcomes $x_1, \ldots, x_n$ on the n trials (that is, we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$), termed a **random sample**.

Objectives of a Statistical Analysis

- After the experiment has been carried out, a sample of **observed data** will have been obtained. Suppose that we have observed outcomes $x_1, \ldots, x_n$ on the n trials (that is, we have observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$), termed a **random sample**.
- This sample can be used to answer qualitative and quantitative questions about the nature of the experiment being carried out.

Objectives of a Statistical Analysis

The objectives of a statistical analysis can be summarized as follows:

- SUMMARY : **Describe** and **summarize** the sample $\{x_1, ..., x_n\}$ in such a way that allows a specific probability model to be proposed.
- INFERENCE : **Deduce** and **make inference about** the parameter(s) of the probability model θ.
- TESTING : **Test** whether θ is " **significantly**" larger/smaller/different from some specified value.
- GOODNESS OF FIT : **Test** whether the probability model encapsulated in the mass/density function f, and the other model assumptions are **adequate** to explain the experimental results.

Parameter Estimation

- It is often of interest to draw **inference from data** regarding the parameters of the **proposed probability distribution**

- It is often of interest to draw **inference from data** regarding the parameters of the **proposed probability distribution**
- Recall that many aspects of the standard distributions studied are **controlled by the distribution parameters**. It is therefore important to find a simple and yet general technique for parameter estimation

Parameter Estimation

- It is often of interest to draw **inference from data** regarding the parameters of the **proposed probability distribution**
- Recall that many aspects of the standard distributions studied are **controlled by the distribution parameters**. It is therefore important to find a simple and yet general technique for parameter estimation
- The technique we focus on is the method of **maximum likelihood**, first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.

Parameter Estimation

- It is often of interest to draw **inference from data** regarding the parameters of the **proposed probability distribution**
- Recall that many aspects of the standard distributions studied are **controlled by the distribution parameters**. It is therefore important to find a simple and yet general technique for parameter estimation
- The technique we focus on is the method of **maximum likelihood**, first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.
- Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable properties

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data
- Suppose a sample $x_1, \ldots, x_n$ has been obtained from a probability model specified by mass or density function $f(x; \theta)$ depending on parameter(s) θ lying in parameter space Θ

Maximum Likelihood Estimation (MLE)

- Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data
- Suppose a sample $x_1, \dots, x_n$ has been obtained from a probability model specified by mass or density function $f(x; \theta)$ depending on parameter(s) θ lying in parameter space Θ
- The **maximum likelihood estimate** or **m.l.e.** is produced as follows:

Maximum Likelihood Estimation (MLE)

STEP 1 Write down the **likelihood function**, $L(\theta)$, where

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

that is, the product of the n mass/density function terms (where the ith term is the mass/density function evaluated at x_i) viewed as a function of θ

STEP 2 Take the natural log of the likelihood, and collect terms involving θ

STEP 3 Find the value of $\theta \in \Theta$, $\hat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. If θ is a single parameter, find $\hat{\theta}$ by solving

$$\frac{d}{d\theta}\ \{\log L(\theta)\} = 0$$

in the parameter space Θ. If θ is vector-valued, say $\theta = (\theta_1, ..., \theta_d)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ by simultaneously solving the d equations given by

$$\frac{\partial}{\partial\theta_j}\ \{\log L(\theta)\} = 0 \qquad j = 1, \ldots, d$$

in parameter space Θ.

STEP 4 Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. If

$$\frac{d^2}{d\theta^2}\ \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of θ

Maximum Likelihood Estimation (MLE)

STEP 4 Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. If

$$\frac{d^2}{d\theta^2}\ \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of θ

This procedure is a systematic way of producing parameter estimates from sample data and a probability model; it can be shown that such an approach produces estimates that have good properties. After they have been obtained, the estimates can be used to carry out *prediction* of behaviour for future samples.

Likelihood and Maximum Likelihood Estimator

Definition 1 (Likelihood Function)

Let $X_1, \ldots, X_n$ have joint PMF or PDF

$$f(x_1, \ldots, x_n; \Theta)$$

where the parameters $\Theta \equiv (\theta_1, \ldots, \theta_m)$ have unknown values and $\mathbf{x} = x_1, \ldots, x_n$ are the observed sample values. The **likelihood function** is regarded as a function of Θ

$$L(\Theta; \mathbf{x}) = f(x_1, \ldots, x_n; \Theta) \tag{2}$$

Definition 2 (Maximum Likelihood Estimator)

$$\widehat{\Theta}(\mathbf{x}) = \arg\max_{\Theta} L(\Theta; \mathbf{x}) \tag{3}$$

Invariance Principle: if $\widehat{\Theta}(\mathbf{x})$ is a MLE for Θ, then $g(\widehat{\Theta}(\mathbf{x}))$ is a MLE for $g(\theta)$

Example 3 (Poisson MLE)

A sample $x_1, \ldots, x_n$ is modelled by a Poisson distribution with parameter denoted λ

$$f(x;\theta) \equiv f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, \ldots$$

for some $\lambda > 0$. Find the MLE of λ **analytically**.

Poisson MLE using `optim`

We can use the `stats::optim` function in R to find the MLE, provided we have a likelihood function. The `optim` can maximize (or minimize) an objective function using many different algorithms. This is referred to as **solving the objective function numerically**. Simulate some sample data generated from a Poisson distribution and solve for the MLE:

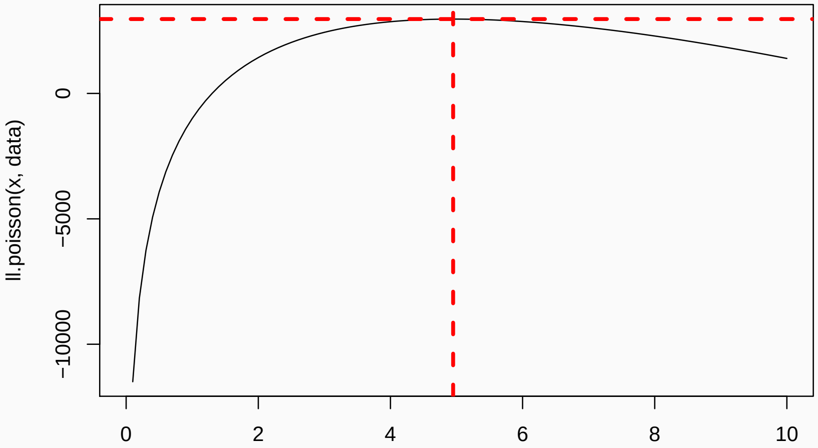
```
# define the objective function
ll.poisson <- function(lambda, x) {
 sum(x) * log(lambda) - length(x) * lambda
}
data <- rpois(1000, 5) # generate some data
# by default optim finds the min, but the negative min is the max
# therefore we need to use list(fnscale = -1)
opt <- optim(par = 2, fn = ll.poisson, method = "BFGS",
             control = list(fnscale = -1), x = data)

c(opt$par, mean(data))
```

```
## [1] 5.051999 5.052000
```

Poisson MLE using `optim`

We plot the objective function (in this case, it's the log-likelihood) and dotted red lines representing the value of the objective function at the value of λ that maximizes the log-likelihood

```
curve(ll.poisson(x, data), 0,10, xlab = "lambda")
abline(h = opt$value, v = opt$par, lty = 2, lwd = 3, col = "red")
```



Example 4 (Bernoulli MLE)

A sample $x_1, \ldots, x_n$ is modelled by a Bernoulli distribution with unknown parameter denoted p

$$f(x;\theta) \equiv f(x;p) = p^x(1-p)^{1-x} \qquad x = 0, 1 \ldots$$

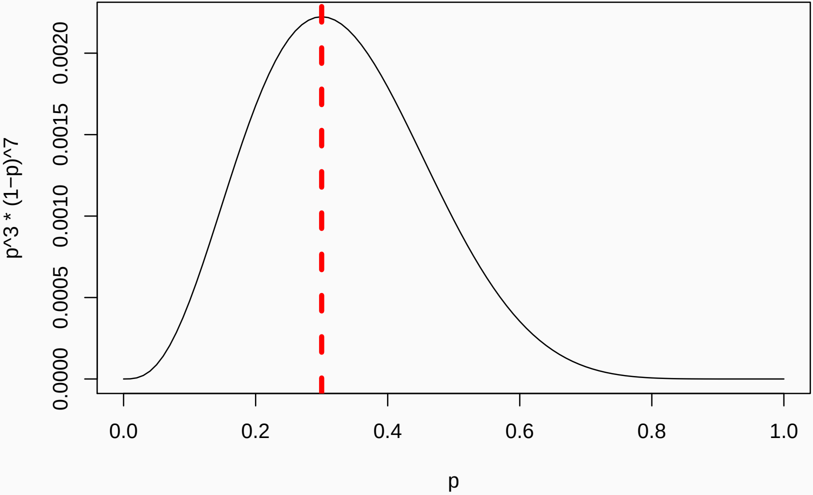
for some $p > 0$. Find the MLE of p.

Example 5 (Bike Helmets)

A sample of ten new bike helmets manufactured by a company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not. Let $p = P(\text{flawed helmet})$ and define $X_1, \ldots, X_{10}$ by $X_i = 1$ if the ith helmet is flawed and zero otherwise. Then the observed x_i's are $1, 0, 1, 0, 0, 0, 0, 0, 0, 1$. For what value of p is the observed sample **most likely to have occurred**? Would anything change if we had been told only that among the ten helmets there were three that were flawed?

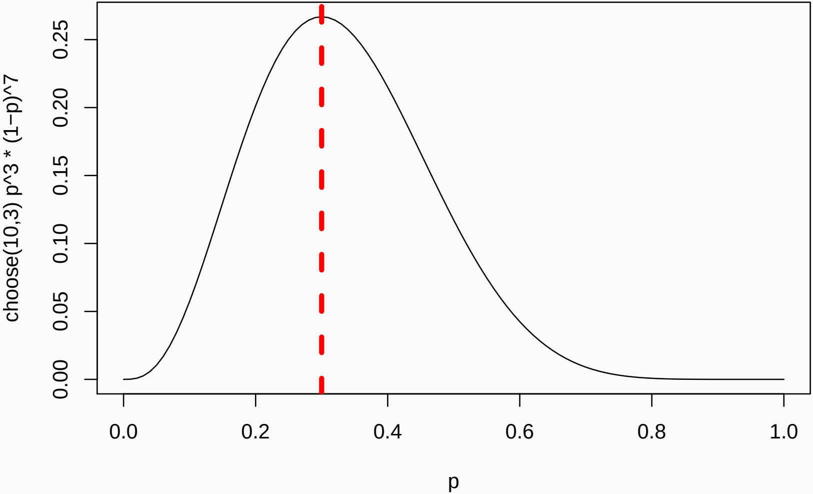
Bernoulli MLE Example

```
bern <- function(x) x^3 * (1-x)^7
curve(bern(x), 0,1, ylab = "p^3 * (1-p)^7", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



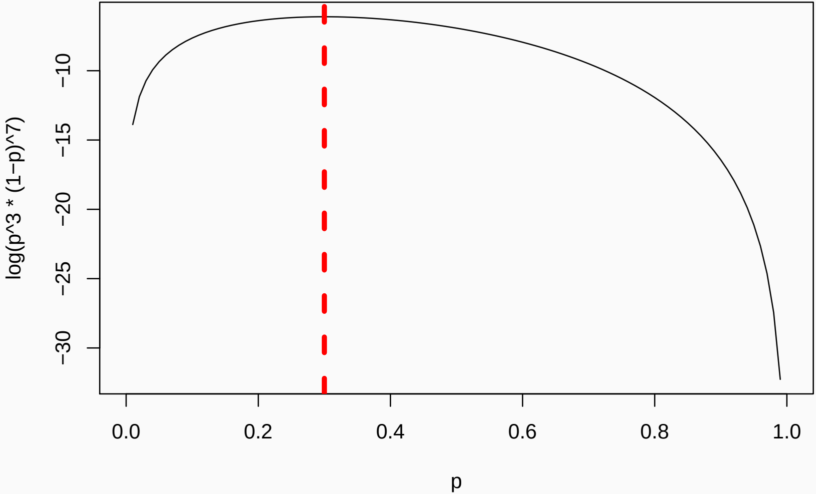
Bernoulli MLE Example

```
binom <- function(x) choose(10,3) * x^3 * (1-x)^7
curve(binom(x), 0,1, ylab = "choose(10,3) p^3 * (1-p)^7", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



Bernoulli MLE Example - Log Likelihood

```r
bern <- function(x) x^3 * (1-x)^7
curve(log(bern(x)), 0,1, ylab = "log(p^3 * (1-p)^7)", xlab = "p")
abline(v = 0.3, lty = 2, col = "red", lwd = 4)
```



Example 6 (Normal MLE)

A sample $x_1, \ldots, x_n$ is modelled by a Normal distribution with unknown parameters $\Theta \equiv (\mu, \sigma^2)$

$$f(x; \Theta) \equiv f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad x \in \mathbb{R}$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$. Find the MLE of μ and σ^2.

Normal MLE using optim

```
ll.normal = function(theta, x) {
  # theta = (mu, sig2)
  # needs to be defined this way for optim to work
  mu = theta[1]
  sig2 = theta[2]
  n = length(x)
  (n / 2) * log(1 / sig2) - sum((x - mu) ^ 2) / (2 * sig2)
}

# generate N(0,1) data
data = rnorm(1000, mean=2, sd=4)
opt <- optim(par = c(0,1), fn = ll.normal, method = "BFGS",
             control = list(fnscale = -1), x = data)
opt$par
```

```
## [1]  2.113731 16.038960
```

```
c(mean(data), sum((data - mean(data))^2)  / length(data) )
```

```
## [1]  2.113731 16.038961
```

Example 7 (Simple Linear Regression)

Suppose that Y_i, $i = 1, \ldots, n$ are n independent normal random variables, each corresponds to a known explanatory variable x_i and has the form

$$Y_i = \beta x_i + \epsilon_i$$

The ϵ_i are independent and from a normal distribution with mean 0 and variance σ^2. Find the MLEs for the parameters $\Theta = (\beta, \gamma)$ where $\gamma = \sigma^2$. How is the maximum likelihood estimation for β related to the least squares estimation in linear regression? Use the `data(women)` to verify your answer

```r
data(women) # from the datasets package

# -1 because we dont want an intercept
summary(lm(height ~ -1 + weight, data = women))
```

Session Info

```r
devtools::session_info()
```

```
## setting  value
## version  R version 3.4.1 (2017-06-30)
## system   x86_64, linux-gnu
## ui       X11
## language en_US
## collate  en_US.UTF-8
## tz       Canada/Eastern
## date     2017-11-21
##
## package     * version    date       source
## abind         1.4-5      2016-07-21 cran (@1.4-5)
## arm           1.9-3      2016-11-27 cran (@1.9-3)
## assertthat    0.2.0      2017-04-11 CRAN (R 3.4.1)
## backports     1.1.0      2017-05-22 cran (@1.1.0)
## base        * 3.4.1      2017-07-08 local
## bindr         0.1        2016-11-13 CRAN (R 3.4.1)
## bindrcpp      0.2        2017-06-17 CRAN (R 3.4.1)
## blme          1.0-4      2015-06-14 cran (@1.0-4)
## broom         0.4.2      2017-02-13 CRAN (R 3.4.1)
```