MATH3091 Statistical Modelling II
Lecture 2: Likelihood Function

Dr. Chao Zheng

4th Feb 2022

2.1 The likelihood function

Setup and notation

Suppose that data consist of n observations $\boldsymbol{y} = (y_1, \dots, y_n)^T$.

The vector $\boldsymbol{y}$ contains observations of random variables

$$\boldsymbol{Y} = (Y_1, \dots, Y_n)^T,$$

which have joint probability density function (p.d.f.) $f_{\boldsymbol{Y}}$ (joint probability function (p.f.) for discrete variables).

We often assume that $Y_1, \dots, Y_n$ are **independent** random variables. Hence

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

Introduction to the likelihood

In parametric statistical inference, we specify a joint distribution $f_{\boldsymbol{Y}}$, for $\boldsymbol{Y}$, which is known except for some parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

Then we use the observed data $\boldsymbol{y}$ to make inferences about $\boldsymbol{\theta}$. In this case, we usually write $f_{\boldsymbol{Y}}$ as $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$, to make explicit the dependence on the unknown $\boldsymbol{\theta}$.

Often we think of the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{y}$ for fixed $\boldsymbol{\theta}$, which describes the relative probabilities of different possible values of $\boldsymbol{y}$, given a particular set of parameters $\boldsymbol{\theta}$.

However, in statistical inference, we have observed $\boldsymbol{y}$, and want to know (infer) $\boldsymbol{\theta}$: e.g, which values of $\boldsymbol{\theta}$ could plausibly have generated the observations $\boldsymbol{y}$?

The likelihood function

In this way, we can think of $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for fixed $\boldsymbol{y}$, which describes the relative *likelihoods* of different possible (sets of) $\boldsymbol{\theta}$, given observed data $y_1, \ldots, y_n$.

For this likelihood, we write it as

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}),$$

which is a function of the unknown parameter $\boldsymbol{\theta}$. For convenience, we often drop $\boldsymbol{y}$ from the notation, and write $L(\boldsymbol{\theta})$.

Notes on the likelihood

1. Frequently it is more convenient to consider the *log-likelihood* function $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.
2. Nothing in the definition of the likelihood does not requires $y_1, \ldots, y_n$ to be observations of independent random variables, although we shall frequently make this assumption.
3. Any factors which depend on $y_1, \ldots, y_n$ alone (and not on $\boldsymbol{\theta}$) can be ignored when writing down the likelihood. Such factors give no information about the relative likelihoods of different possible values of $\boldsymbol{\theta}$.

Modelling births in the UK

Between 2011 and 2015 there were $n = 3827170$ live births recorded in the UK. Let

$$Y_i = \begin{cases} 1 & \text{if child } i \text{ is female} \\ 0 & \text{if child } i \text{ is male} \end{cases},$$

$i = 1, \ldots, n$.

How could we model Y_i?

- $Y_i \sim \text{Bernoulli}(p)$, where p is unknown
- $Y_i \sim \text{Bernoulli}(0.5)$
- $Y_i \sim \text{geometric}(p)$, where p is unknown
- $Y_i \sim \text{geometric}(0.5)$
- $Y_i \sim \text{Poisson}(\lambda)$, where λ is unknown
- None of the above

Modelling births in the UK

Between 2011 and 2015 there were $n = 3827170$ live births recorded in the UK. 1863820 of these children were recorded as female and 1963350 as male.

Suppose we model $Y_i \sim \text{Bernoulli}(p)$. How would you estimate p?

- 0.5
- $1863820/3827170 = 0.486$
- $1963350/3827170 = 0.513$
- $1863820/1963350 = 0.949$
- $1963350/1863820 = 1.053$
- None of the above

Modelling births in the UK

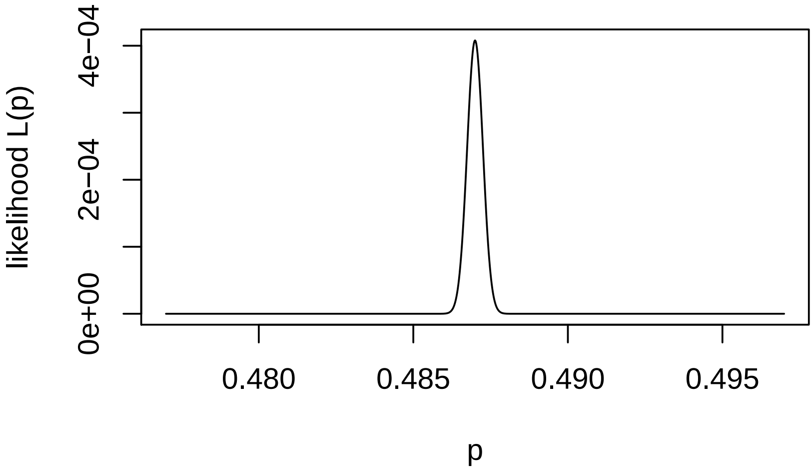
Between 2011 and 2015 there were $n = 3827170$ live births recorded in the UK. 1863820 of these children were recorded as female and 1963350 as male.

Suppose we model $Y_i \sim \text{Bernoulli}(p)$. How would you estimate p?

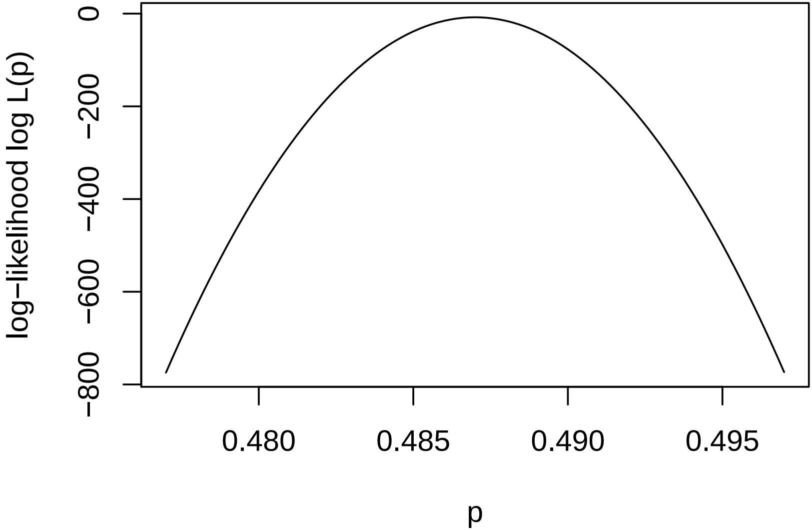
- 0.5
- $1863820/3827170 = 0.486$
- $1963350/3827170 = 0.513$
- $1863820/1963350 = 0.949$
- $1963350/1863820 = 1.053$
- None of the above

We estimate p as $\bar{y} = \frac{1863820}{3827170} = 0.486$.

Modelling births in the UK: likelihood function



Modelling births in the UK: log-likelihood function



Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, independent identically distributed (i.i.d.) Bernoulli(p) random variables. Here $\theta = (p)$ and the likelihood is

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

Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, independent identically distributed (i.i.d.) Bernoulli(p) random variables. Here $\theta = (p)$ and the likelihood is

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

The log-likelihood is

$$\ell(p) = \log L(p) = n\bar{y}\log p + n(1-\bar{y})\log(1-p).$$

Plotting the log-likelihood with R

```
lfun <- function(p, y) {
  ybar <- mean(y)
  n <- length(y)
  n * ybar * log(p) + n * (1 - ybar) * log(1 - p)
}
```

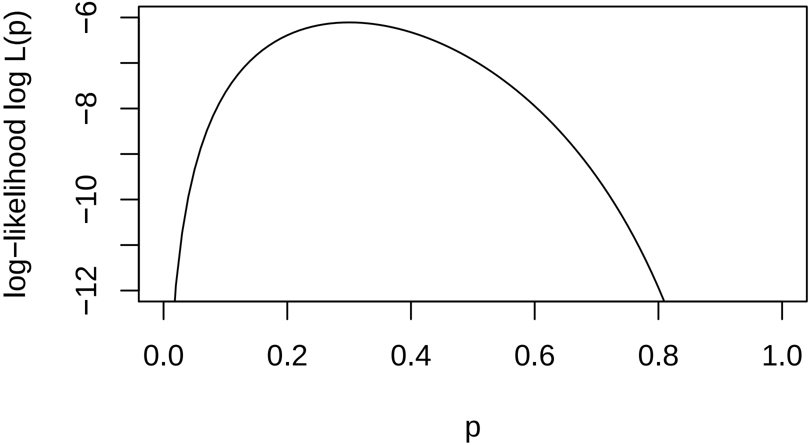
e.g. suppose ***y*** is

```
y <- c(1, 0, 0, 0, 1, 0, 0, 0, 1, 0)
```

We can plot the log-likelihood with

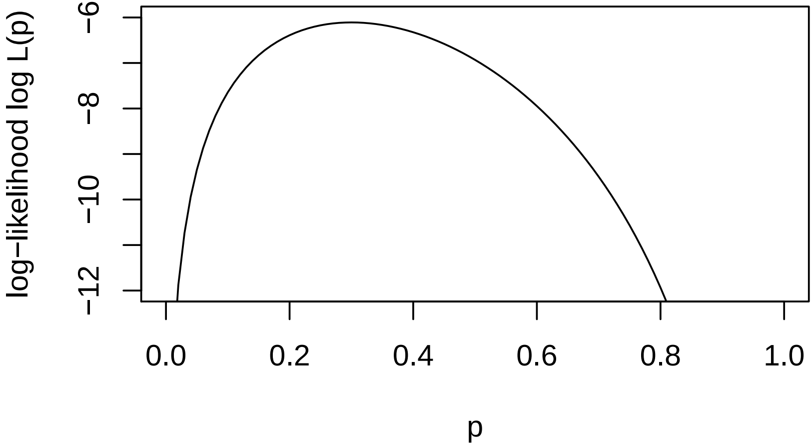
```
curve(lfun(x, y), from = 0, to = 1, ylim = c(-12, -6),
      xlab = "p", ylab = "log-likelihood log L(p)")
```

Plotting the log-likelihood



How would you estimate p in this example?

Plotting the log-likelihood



How would you estimate p in this example? How confident are you of your estimate? (e.g. could $p = 0.5$ plausibly have generated the data?)

2.1.1 Maximum likelihood estimation

Maximum likelihood estimation

We call the value of $\boldsymbol{\theta}$ which maximises the likelihood $L(\theta)$ the *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

$\hat{\boldsymbol{\theta}}$ depends on $\boldsymbol{y}$, as different observed data samples lead to different likelihood functions.

The corresponding function of $\boldsymbol{Y}$ is called the *maximum likelihood estimator* and is also denoted by $\hat{\boldsymbol{\theta}}$.

Some properties of the MLE

As $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the MLE for any component of $\boldsymbol{\theta}$ is given by the corresponding component of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$.

The MLE for any function of parameters $g(\boldsymbol{\theta})$ is given by $g(\hat{\boldsymbol{\theta}})$.

As log is a strictly increasing function, the value of $\boldsymbol{\theta}$ which maximises $L(\boldsymbol{\theta})$ also maximises $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. It is almost always easier to maximise $\ell(\boldsymbol{\theta})$.

"Usual" recipe for finding the MLE

1. Write down the likelihood $L(\boldsymbol{\theta})$.
2. Take logs to find the log-likelihood $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.
3. Find a stationary point $\hat{\boldsymbol{\theta}}$ by differentiating $\ell(\boldsymbol{\theta})$ with respect to $\theta_1, \ldots, \theta_p$, and solving the resulting p simultaneous equations.
4. Check that the stationary point $\hat{\boldsymbol{\theta}}$ is a maximum (rather than a minimum or point of inflection) of the log-likelihood.

Example (Bernoulli)

$y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and the log-likelihood is

$$\ell(p) = n\bar{y}\log p + n(1-\bar{y})\log(1-p).$$

Differentiating with respect to p,

$$\frac{\partial}{\partial p}\ell(p) = \frac{n\bar{y}}{p} - \frac{n(1-\bar{y})}{1-p}$$

so the MLE $\hat{p}$ solves

$$\frac{n\bar{y}}{\hat{p}} - \frac{n(1-\bar{y})}{1-\hat{p}} = 0.$$

Solving this for $\hat{p}$ gives $\hat{p} = \bar{y}$. Note that

$$\frac{\partial^2}{\partial p^2}\ell(p) = -n\bar{y}/p^2 - n(1-\bar{y})/(1-p)^2 < 0$$

everywhere, so the stationary point is clearly a maximum.

Checking with R

In our test case, where $\boldsymbol{y}$ is

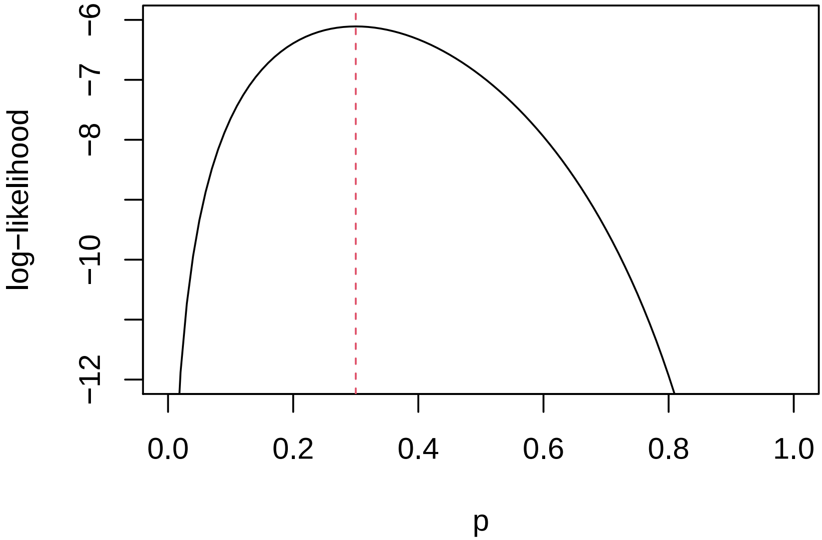
```
y
```

```
##  [1] 1 0 0 0 1 0 0 0 1 0
```

let's plot out the likelihood, and add on the MLE:

```
curve(lfun(x, y), from = 0, to = 1, ylim = c(-12, -6),
      xlab = "p", ylab = "log-likelihood")
ybar <- mean(y)
points(ybar, lfun(ybar, y))
```

Checking with R



Is this a sensible estimate?

For one particular dataset, we have just found $\hat{p}$ as 0.3. This is the maximum likelihood **estimate** for this particular dataset.

Is this a sensible estimate?

For one particular dataset, we have just found $\hat{p}$ as 0.3. This is the maximum likelihood **estimate** for this particular dataset.

To check that this is a sensible way to estimate p, we could generate data $y_1, \ldots, y_n$ from the model for some known value of p, and estimate p with $\hat{p} = \bar{y}$.

Is $\hat{p}$ close to p?

```
n <- 100
theta <- 0.6
y <- rbinom(n, 1, theta)
p_hat <- mean(y)
p_hat
```

```
## [1] 0.62
```

Repeating the simulation process

When we generate a new dataset from the model, and compute $\hat{p}$ again, we get a different answer:

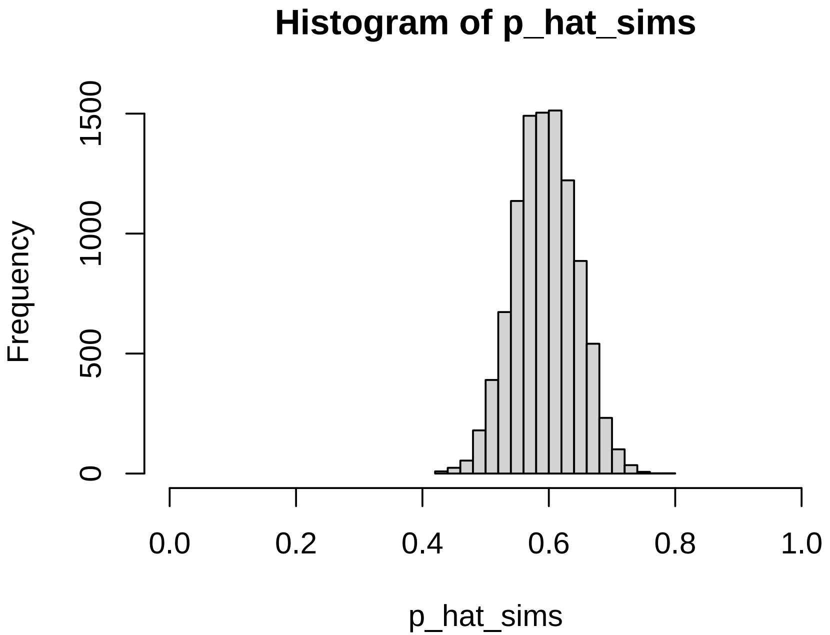
```
y <- rbinom(n, 1, theta)
p_hat <- mean(y)
p_hat
```

```
## [1] 0.63
```

The distribution of estimates from repeated simulations

We could do this 10000 times, and look at the range of estimates which we get:

```
p_hat_sims <- replicate(10000, mean(rbinom(n, 1, theta)))
hist(p_hat_sims, xlim = c(0, 1))
```


Histogram of p_hat_sims
Frequency
0
500
1000
1500
0.0
0.2
0.4
0.6
0.8
1.0
p_hat_sims

The mean of estimates from repeated simulations

```r
mean(p_hat_sims)
```

```
## [1] 0.599103
```

This is very close to the true value of p, 0.6.

It seems that $\hat{p}$ does not systematically underestimate or overestimate p.

Estimates to estimators

In this case, $\hat{p} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the maximum likelihood **estimator** of θ. It is a random variable, and we can look at its distribution.

This is a short-cut to the process of repeated simulation from the model described above.

For instance, we can check that $E(\hat{p}) = E(\bar{Y}) = p$ (for any p), so the estimate is **unbiased**, and we can find $\mathrm{Var}(\hat{p}) = \mathrm{Var}(\bar{Y})$ to see how spread out the esimates will be.

Estimates to estimators

In this case, $\hat{p} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the maximum likelihood **estimator** of θ. It is a random variable, and we can look at its distribution.

This is a short-cut to the process of repeated simulation from the model described above.

For instance, we can check that $E(\hat{p}) = E(\bar{Y}) = p$ (for any p), so the estimate is **unbiased**, and we can find $\text{Var}(\hat{p}) = \text{Var}(\bar{Y})$ to see how spread out the esimates will be.

We can also check what happens to the distribution of $\hat{p}$ as the number of samples n grows large.

There are **general results** about the distribution of the maximum likelihood estimator as n grows large, which we will study soon.

Conclusion

- We have reintroduced the likelihood: "the probability that we would have seen the data we actually did, for each value of the parameter".
- We have reviewed the "usual" recipe for finding maximum likelihood estimates: find a stationary point of the log-likelihood (and check it is a maximum).
- Likelihood is very general: we can find a likelihood function for any probability model. The difficult part is often to choose an appropriate model.