1

Observed likelihood function

1.1 The observed likelihood function

Consider a statistical model $f(y; \theta)$ with observed data $y^0$ from some process or system being examined. We are interested in what information is available concerning the true value of $\theta$ in the application, as discussed briefly in Chapter 1. For this we are assuming that the model provides an accurate description of the underlying process, and we are investigating what information follows from that model with data. Of course there may be questions concerning how accurate the model is but that is a separate and important issue not yet to be addressed here.

The model with density function $f(y; \theta)$ has of course the observed value $f^0(\theta) = f(y^0; \theta)$ in the presence of the data. This is just the $y = y^0$ section of the full function $f(y; \theta)$, and is indicated by the mathematics of “plugging in” the value $y = y^0$. We obtain

$$L^0(\theta) = cf(y^0; \theta),$$  \hspace{1cm} (1.1)

which is called the likelihood function, and is a fundamental implication from the model with data $y^0$.

The likelihood function records how the probability associated with the data point $y = y^0$ varies or how it changes as we contemplate various values for the parameter. For such reasons, it is appropriate to leave the function arbitrary to a positive multiplicative constant $c$, thus only relative values of probability, from one $\theta$ to another, are available. Indeed for an absolutely continuous distribution we only have probability in a differential element $dy$ at $y = y^0$ not at a mathematical point so there is then implicitly an arbitrary element. And of course the widely admired sufficiency property only attaches to the likelihood information when this arbitrary constant is included in the definition mak-
Observed likelihood function

Figure 1.1 An observed log-likelihood with location at $\hat{\theta}$ and scaling indicated by the curvature $\kappa(\hat{\theta})$.

For likelihood it is often convenient to work in logarithmic units and use the log-likelihood function

$$\ell^0(\theta) = \log L^0(\theta):$$

$$\ell^0(\theta) = a + \log f(y^0; \theta)$$ (1.2)

For a simple example, see Section 1.3 and the related Figures (1.3) and (1.4).

In the presence of two investigations where the corresponding variables are statistically independent, we can multiply the likelihoods or add the log-likelihoods:

$$L^0(\theta) = L^0_1(\theta)L^0_2(\theta), \quad \ell^0(\theta) = \ell^0_1(\theta) + \ell^0_2(\theta).$$ (1.3)

1.2 Simple things about a likelihood function

We now define key characteristics of an observed likelihood. For a given model with observed data let $\ell(\theta)$ in Figure (1.1) be the

\begin{align*}
\end{align*}
1.2 Simple things about a likelihood function

observed likelihood. The value of $\theta$ that maximizes $\ell(\theta)$ is called the
maximum likelihood value,

$$\hat{\theta} = \text{argsup } L(\theta) = \text{argsup } \ell(\theta)$$

and its observed value is $\hat{\theta}_0$; this indicates where the likelihood is
concentrated.

The curvature of the likelihood function at the maximum is

$$j_{\theta\theta} = -\ell_{\theta\theta}(\theta) = -\frac{\partial^2}{\partial \theta \partial \theta} \ell(\theta)$$

which is the negative second derivative or negative Hessian of $\ell(\theta)$, evaluated at the maximum likelihood value $\hat{\theta}$. This indicates how
tightly or concentrated the likelihood is at its maximum.

The slope of the log-likelihood, at some parameter value $\theta$ of
interest, viewed as a possible true value, is called the score variable

$$s(\theta) = \ell_\theta(\theta) = \frac{\partial}{\partial \theta} \ell(\theta);$$

and we are using a subscript to indicate partial differentiation with
respect to the indicated variable. The score shows how quickly the
log-likelihood is changing at that value of the parameter and is
important for assessing the value $\theta$ as a possible true value for the
parameter.

And now we have one more important characteristic of an ob-
served log-likelihood. It is called the signed likelihood root and
is available for scalar parameters say $\theta$. The rise of log-likelihood
from an interest value $\theta$ to the maximum value is called the log-
likelihood ratio statistic and is given as

$$\hat{\ell} - \ell(\theta) = \ell(\hat{\theta}) - \ell(\theta) = r^2 / 2$$

where $r^2$ is called the deviance. We take the root of $2(\hat{\ell} - \ell) = r^2$
and attach the sign of $\hat{\theta} - \theta$ to obtain the signed-likelihood root

$$r = \text{sign}(\hat{\theta} - \theta)[2(\hat{\ell} - \ell(\theta))]^{1/2}. \quad (1.4)$$

This gives us four widely used characteristics of an observed
likelihood: the maximum likelihood value $\hat{\theta}$, the curvature at the
maximum called the observed information $j_{\theta\theta}$, the score at a value
$\theta$ designated $s(\theta)$, and the signed likelihood root $r$. Each of these
has important roles in likelihood analysis; and each has advantages
as part of providing a visual assessment of an observed likelihood; see Figure (1.1).
1.3 Behaviour of the likelihood function: simple Bartlett properties

How does a likelihood function change in repeated sampling? Does an observed likelihood function shift around, change shape? Some answers come fairly easily. Let $f(y; \theta)$ be the statistical model for the variable $y$ which here could be a scalar variable or even the vector variable for a sample of $n$. And let $\ell(\theta; y)$ be the corresponding log-likelihood function from a general data point $y$. We have of course that

$$\int f(y; \theta) dy = 1$$

for all values of $\theta$. Then from curiosity we might reasonably try differentiating with respect to $\theta$ to see what we might obtain. For this we suppose that the differentiation can be carried through the sample space integral sign and then use

$$\ell_\theta(y; \theta) = \frac{\partial}{\partial \theta} \log f(y; \theta) = \left\{ \frac{1}{f(y; \theta)} \right\} \left\{ \frac{\partial}{\partial \theta} f(y; \theta) \right\}$$

or equivalently

$$\frac{\partial}{\partial \theta} \int f(y; \theta) dy = \int \ell_\theta(y; \theta) f(y; \theta) dy = E\{s(\theta; y); \theta\} = 0. \quad (1.5)$$

This tells us that the slope of the log-likelihood function at the 'true' value of the parameter has mean value zero; and can be referred to as the first Bartlett property of $f(y; \theta)$.

To get a better picture of this, see Figure (1.2). This exhibits a possible observed log-likelihood function and the horizontal axis uses $\theta$ as a free variable standing for possible values of the parameter. If however we want to think of repetition from some true distribution we will take $\theta_*$ as the label for that true value of the parameter. And then to exhibit log-likelihoods we will take advantage of the arbitrary additive constant with an observed log-likelihood and use a representative log-likelihood that has value 0 at the true value $\theta_*$. In repetitions then the observed log-likelihood will always go through the $\theta$-axis at the point $\theta_*$. Then (3.5) implies that that the slope of the log-likelihoods at that crossing point will some times be positive and sometimes be negative but will have
1.3 Behaviour of the likelihood function: simple Bartlett properties

Figure 1.2 A log-likelihood function for consideration under the true value $\theta^*$; at that true the arbitrary additive constant has been chosen so the log-likelihood has value zero. The slope can be presented as $s(\theta^*) = i^{1/2}(\theta^*)z$; and the negative second derivative is $i(\theta^*)$

a mean or average value exactly equal to zero, $E(l_\theta(\theta^*; y); \theta^*) = 0$

or just (3.5) in the more flexible notation

Now suppose we try partial differentiation again but now applied to the new formula formula (3.5):

$$\frac{\partial}{\partial \theta} \int l_\theta(\theta; y)f(y; \theta)dy = \int \{l_{\theta\theta} + l_{\theta}^2\} f(y; \theta)dy = 0 \quad (1.6)$$

Then subject to some integrability conditions on $l_{\theta\theta}(\theta; y)$ and $l_\theta^2(\theta; y)$ we can split the integral and rewrite the result as

$$i_{\theta\theta}(\theta) = E\{-l_{\theta\theta}; \theta\} = \text{var}\{l_\theta; \theta\} \quad (1.7)$$

where the differentiation of the product inside the integral produces two terms, one with a second derivative and the other with an extra score function. This $i_{\theta\theta}(\theta)$ is called the expected infor-
mation concerning \( \theta \) and (3.7) shows that it can be calculated as
the mean negative curvature at the true \( \theta \) or as the variance of
the score at that true value. This has many uses, in particular
mean values are often easier to calculate than variances. And it
does link curvature of log-likelihood to variability in the slope of
log-likelihood; this is referred to as the second Bartlett property.

Formulas (3.6) and (3.8) obtained by differentiating the norming
equation \( \int f(y; \theta)dy = 1 \) are Bartlett identities and their usefulness indicates the power of the norming built into a statistical
density model.

1.4 First-order distribution of the log-likelihood
function

Certainly knowing how the slope of a log-likelihood behaves in
repetitions under the true parameter value of the parameter is
important information: the slope has mean value zero and the variance is given by the negative mean curvature of the log-likelihood.
But knowing more would be helpful, such as what log-likelihoods
to expect and with what probabilities, that is the actual distribution
describing the potential log-likelihoods. For this, consider
a statistical model \( f(y; \theta) \) for a variable \( y \), and then the model
for \( (y_1, \ldots, y_n) \) where the variables \( y_i \) are independent, each with
that initial model. A full observed log-likelihood can be Taylor
expanded about the true value say \( \theta_\ast \), which for notational conve-
nience we take here to be zero, \( \theta_\ast=0 \). The expansion to first order
gives

\[
\ell(\theta) = 0 + s(0)\theta - j(0)\theta^2/2 + O(n^{1/2}) \tag{1.8}
\]

where \( s, j \) depend, although not indicated, on the vector variable
\( (y_1, \ldots, y_n) \); and where the individual log-likelihoods each have
the value 0 at the true \( \theta_\ast = 0 \); see Figure 1.2.

The score \( s(0) \) at the true \( \theta_\ast = 0 \) is a sum of independent
identically distributed components, each with mean 0 and variance
\( i(\theta_\ast) \); thus the score has a limiting Normal distribution with mean
0 and variance \( V = V(\theta_\ast) = ni(\theta_\ast) \); we can thus write \( s(0) = V^{1/2}z \) where \( z \) denotes a standard normal variable. Meanwhile the observed information \( j(0) \) is a sum of \( n \) independent components
each with mean value \( i(\theta_\ast) \) and thus to first order can be written
as \( H = H(\theta^*) = \text{ni}(\theta^*) \). It follows then that (3.8) can be rewritten to first order as

\[
\ell(\theta) = V^{1/2}z\theta - H\theta^2/2 + \ldots \quad (1.9)
\]

\[
= i^{1/2}(\theta^*)z\delta - i(\theta^*)\delta^2/2 + O(n^{-1/2}) \quad (1.10)
\]

where the parameter departure \( \theta = \delta V^{-1/2} \) is expressed in terms of \( \delta = n^{1/2} \) so as to scale appropriately in the \( O(n^{-1/2}) \) neighbourhood of the true value. Thus to first order the log-likelihood is an ordinary parabola through the origin, and when scaled in terms of \( \delta \) in root information \( i^{1/2}(\theta^*) \) units about the true has slope \( z \) and negative curvature equal unity.

### 1.5 Distribution under model misspecification

When examining a statistical model \( f(y; \theta) \) with data \( y^0 \) there can be occasions when the model is viewed as only a rough approximation, and yet distributions are wanted for various likelihood characteristics when some other density \( h(y) \) is deemed as the true distribution, that is the distribution that actually produced the data \( y^0 \). This then is often explored as a case of model misspecification.

From the description above we have that \( h(y) \) is not a density in the model \( f(y; \theta) \) but there may yet be a density \( f(y; \theta^*) \) in the model that can be viewed as closest. For this we solve the score type equation

\[
\left\{ \frac{\partial}{\partial \theta} \right\} \int \ell(\theta; y)h(y)dy = \int \ell_h(y)dy = 0 \quad (1.11)
\]

obtaining say \( \theta = \theta^* \), which is a sort of maximum likelihood approximation to the true density \( h(y) \). The likelihood expansion (3.8) then uses

\[
V = V(\theta^*) = \text{var}\{\ell'(\theta^*); h(y)\}, \quad H = H(\theta) = E\{-\ell''(\theta); h(y)\}
\]

as respectively the variance of \( \ell'(\theta^*) \) and the mean of \( -\ell''(\theta^*) \) under the true distribution \( h(y) \). And then equation (3.9) using the present \( V \) and \( H \) becomes

\[
\ell(\theta) = \theta'V^{1/2}z - \theta'H\theta^2/2 + \ldots, \quad (1.12)
\]

to first order. For this we write formulas in a style suitable for the vector case; in particular we write \( V = V'^{1/2}V^{1/2} \) where \( V'^{1/2} \)
is a right square root of the $p \times p$ variance matrix $V$ and write $s = V^{1/2}z$ for the score variable with variance $V$ using $z$ as a standard Normal variable in $p$ dimensions.

We now find the maximum likelihood value for the likelihood expression (3.12)

$$\ell_{\theta}(\theta) = V^{1/2}z - H\theta = 0 \Rightarrow \hat{\theta} = H^{-1}V^{1/2}z \quad (1.13)$$

showing that $\hat{\theta}$ is first order Normally distributed with mean $\theta^*$ and variance $H^{-1}VH^{-1}$. The variance of $\hat{\theta}$ is $H^{-1}(\theta_*)V(\theta_*)H^{-1}(\theta_*)$, and can be estimated with data $y_0$ by the plug-in estimate $\hat{\text{var}} = H^{-1}(\hat{\theta}_0)V(\hat{\theta}_0)H^{-1}(\hat{\theta}_0)$ sometimes called the sandwich estimator.

The first order distribution of the maximum likelihood $\hat{\theta}$ can then be described informally as being Normal with mean $\theta^*$ and variance $H^{-1}(\hat{\theta}_0)V(\hat{\theta}_0)H^{-1}(\hat{\theta}_0)$.

Another approach towards obtaining distributional results under some density $h(y)$ not in the model is to construct using $h(y)$ a larger model in which the given model is embedded, and then to evaluate the parameter associated with $h(y)$. We pursue this in Chapter xx.

### 1.6 Problems

1. Calculate the score, the maximum likelihood value, and the information for the Bernoulli model in Example 2.1.
2. (continuation) Verify that the Bartlett identities (2.6) and (2-8) hold.
3. Calculate the score, the maximum likelihood value, and the information for the Exponential life model in Example 2.3.
4. (cont.) Verify that the Bartlett identities (2.6) and (2-8) hold.
5. Calculate the Bartlett relations for the general exponential model (2.1).
6. Use the likelihood function representation (3.10) to show that $r$ and $r^2$ have limiting standard Normal and Chi-square distributions to first order respectively.