The Likelihood Function

If \( X \) is a discrete or continuous random variable with density \( p_\theta(x) \), the likelihood function, \( L(\theta) \), is defined as

\[
L(\theta) = p_\theta(x)
\]

where \( x \) is a fixed, observed data value.

**Remark 1** If \( X \) is discrete, then \( L(\theta) \) is the probability of observing \( x \) given \( \theta \). When \( X \) is continuous, then \( L(\theta) \) does not represent the probability of observing the data value \( x \) since this probability must be 0. Since data can only be observed and measure with finite precision, however, it is reasonable that \( L(\theta) \approx \epsilon p_\theta(x) \) for some \( \epsilon > 0 \) representing the precision limit. Since it turns out that \( L(\theta) \) is only meaningful up to a multiplicative constant anyway, we can ignore \( \epsilon \) and compute \( L(\theta) \) using \( p_\theta(x) \) alone.

Examples of the Likelihood Function

**Example 1** Let \( X \) be a binomial random variable with parameters \( n = 20 \) and \( \theta \). Then,

\[
L(\theta) = p_\theta(x) = \binom{20}{x} \theta^x (1 - \theta)^{20-x}, \quad 0 \leq \theta \leq 1.
\]

The following graphs show \( L(\theta) \) for various values of \( x \).
For an interesting applet illustrating the likelihood function for the binomial distribution, see [http://fisher.forestry.uga.edu/popdyn/Likelihood.html](http://fisher.forestry.uga.edu/popdyn/Likelihood.html).

**Example 2** An unknown number, say $N$, of animals inhabit a certain region. To obtain some information about the population size, ecologists often perform the following experiment. They first catch a number, $m$, of these animals and tag or mark them in some manner. The captured animals are then released back into the region. After allowing the tagged animals time to disperse throughout the region, a new catch of size, say $n$, is made. Let $X$ denote the number of marked animals in the second catch. If we assume that the number of animals in the region remains essentially constant between the times of the two captures and that each time an animal was caught it was equally likely to be any of the remaining uncaught animals, it follows that $X$ is hypergeometrically distributed and that

$$L(N) = p_N(x) = \frac{{m \choose x}{N-m \choose n-x}}{{N \choose n}}$$

where $x$ is the number of marked animals in the second catch. Suppose that $m = 50$ and $n = 40$. The following graphs show $L(N)$ for various values of $x$.

$L(N)$ for the Hypergeometric Distribution With $m = 50$ and $n = 40$

It should be the case that $m/N \approx x/n$ giving that $N \approx mn/x$. In fact, it is easily shown that $L(N)$ is maximized at $\hat{N} = [mn/x]$.

**Example 3** (based on YP, exercise 2.5, p. 49) The following data shows the heart rate (in beats/minute) of a person measured through the day.

73 75 84 76 93 79 85 80 76 78 80
Assume the data are an iid sample from $N(\theta, \sigma^2)$ where $\sigma^2$ is known as the observed sample variance $s^2$. Thus,

$$p_\theta(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right).$$

Consider the following cases: (a) only the first value $x_1 = 73$ is reported, (b) only the sample mean $\bar{x}$ is reported, (c) only the sample median $x(6)$ is reported, and (d) only $x_{(11)} = x_{\text{max}}$ is reported. (For the data above, $\bar{x} = 879/11$, $x(6) = 79$ and $x_{(11)} = 93$.) The distributions needed for each of the cases (a), (b), (c), and (d) are as follows:

- $p^a_\theta(x) = p_\theta(x),$
- $p^b_\theta(x) = \left(\frac{n}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2\sigma^2}(x-\theta)^2\right)$ (if $n = 11$),
- $p^c_\theta(x) = \frac{11!}{5!5!} [F_\theta(x)]^5[1 - F_\theta(x)]^5 p_\theta(x)$, and
- $p^d_\theta(x) = 11[F_\theta(x)]^{10} p_\theta(x)$.

$F_\theta(x)$ denotes the cdf. (Recall that the distribution of the $i^{th}$ order statistic is

$$p_{\theta,i}(x) = \frac{n!}{(i-1)!(n-i)!} [F_\theta(x)]^{(i-1)}[1 - F_\theta(x)]^{(n-i)} p_\theta(x)$$

for a distribution $p_\theta(x).$)
Combining Likelihoods

The definition of the likelihood function can be extended to accommodate the combination of likelihoods. For example, if \( \{X_1, X_2, \ldots, X_n\} \) is a random sample from a discrete or continuous random variable \( X \) with density \( p_\theta(x) \), the definition of the likelihood function, \( L(\theta) \), is now

\[
L(\theta) = p_\theta(x_1)p_\theta(x_2) \cdots p_\theta(x_n) = \prod_{i=1}^{n} p_\theta(x_i).
\]

If logarithms are taken, we obtain

\[
\log L(\theta) = \log p_\theta(x_1) + \log p_\theta(x_2) + \cdots + \log p_\theta(x_n).
\]

The function \( \log L(\theta) \) is called ‘log-likelihood’.

**Example 4** In reference to Example 3, suppose now that all the data is reported. In this case,

\[
L(\theta) = \prod_{i=1}^{n} p_\theta(x_i) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right)
\]
where \( n = 11 \) and \( x_i \) denotes the \( i^{th} \) heart beat data item.

The log-likelihood is

\[
\log L(\theta) = -\frac{n}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2.
\]

If we drop the term not involving \( \theta \) (to be justified later), we obtain

\[
\log L(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2.
\]
If \( \hat{\theta} \) denotes the maximum likelihood estimate (MLE), then \( L(\theta)/L(\hat{\theta}) \) has a maximum of 1, and therefore
log $L(\theta)/L(\hat{\theta})$ has a maximum of 0. In this sense, log $L(\theta)/L(\hat{\theta})$ is normalized.

Different distributions can be combined in the same likelihood function as long as they share the same parameter and the data values are independent.

**Example 5** Again in reference to Example 3, consider the following two cases: (a) only $x(1)$ and $x(11)$ are reported and (b) only $x(1)$ and $x(2)$ are reported. Using the distributions for the appropriate order statistics, the likelihoods for each of these cases are as follows:

$$L^a(\theta) = 11 [1 - F_\theta(x(1))]^{10} p_\theta(x(1)) \cdot 11 [F_\theta(x(11))]^{10} p_\theta(x(11))$$

$$L^b(\theta) = 11 [1 - F_\theta(x(1))]^{10} p_\theta(x(1)) \cdot 110 F_\theta(x(2)) [1 - F_\theta(x(2))]^9 p_\theta(x(2))$$
$L(\theta)$ For Various Order Statistics Reported
Example 6 For determining the half-lives of radioactive isotopes, it is important to know what the background radiation is in a given detector over a period of time. The following data were obtained in a γ-ray detection experiment over 98 ten-second intervals.

Assuming a Poisson model for the data, we have

\[ p_{\lambda}(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \ldots \]

Therefore, the likelihood function is given by

\[ L(\lambda) = \frac{1}{x_1!x_2!\cdots x_n!} \left( \lambda \sum_{i=1}^{n} x_i \right)^e^{-n\lambda}, \]

and so

\[ \log L(\lambda) = (\sum_{i=1}^{n} x_i) \ln \lambda - n\lambda. \]
(Note the constant has been dropped.) For Poisson data, the MLE is $\hat{\lambda} = \bar{x}$.
The author tries to justify (very tersely) the practice of dropping multiplicative constants not involving the parameter. In Example 1,

\[ L(\theta) = \binom{20}{x} \theta^x (1-\theta)^{20-x}. \]

We can drop the binomial coefficient since it does not involve the parameter \( \theta \). Therefore, \( L(\theta) \) can be simplified to

\[ L(\theta) = \theta^x (1-\theta)^{20-x} \]

or

\[ \log L(\theta) = x \log \theta + (20-x) \log(1-\theta). \]

In Example 4,

\[ L(\theta) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right). \]

Dropping the constant \( \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \), we get that

\[ L(\theta) = \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right), \]
and therefore
\[
\log L(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2.
\]

The author shows that the likelihood ratio is *invariant* under a one-to-one continuous transformation of the data. He simply uses the *standard Jacobian method* to obtain the joint density of the transformed data. This implies that the likelihoods of the original and transformed data are related by the value of the Jacobian, and therefore the likelihood ratios for two distinct parameter values are the same for the original and transformed data. *Thus, only the ratio is really important since it is what is invariant,* and multiplicative constants can be dropped since they do not affect the ratio.

### Approximating the Likelihood Function

There are certain important quantities associated with the likelihood function. Let \( \hat{\theta} \) be the MLE.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Definition</th>
<th>Name</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S(\theta) )</td>
<td>( \frac{\partial}{\partial \theta} \log L(\theta) )</td>
<td>Score function</td>
<td>( S(\theta) = 0 )</td>
</tr>
<tr>
<td>( I(\theta) )</td>
<td>( -\frac{\partial^2}{\partial \theta^2} \log L(\theta) )</td>
<td>Curvature function</td>
<td></td>
</tr>
<tr>
<td>( I(\hat{\theta}) )</td>
<td>( -\frac{\partial^2}{\partial \theta^2} \log L(\hat{\theta}) )</td>
<td>Observed Fisher information</td>
<td>( I(\theta) \gg 0 ) indicates strong peak (less uncertainty about ( \theta ))</td>
</tr>
<tr>
<td>( se(\hat{\theta}) )</td>
<td>( \sqrt{\text{Var}(\hat{\theta})} )</td>
<td>Standard error</td>
<td>( se(\hat{\theta}) \approx I(\hat{\theta})^{-1/2} ) (for regular problems)</td>
</tr>
</tbody>
</table>

**Example 7** Consider the Poisson data in Example 6.

\[
\frac{\partial}{\partial \theta} \log L(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n
\]

\[
-\frac{\partial^2}{\partial \theta^2} \log L(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda^2}
\]

Solving the equation
\[
\frac{\partial}{\partial \theta} \log L(\lambda) = 0
\]
yields \( \hat{\lambda} = \bar{x} \). *The Fisher information is \( I(\hat{\lambda}) = n/\hat{\lambda} \) and the standard error is \( se(\hat{\lambda}) \approx \sqrt{\hat{\lambda}/n} \). ■

By Taylor’s Theorem, we can expand \( \log L(\theta) \) around the MLE \( \hat{\theta} \) to obtain an approximation of \( \log L(\theta) \). Specifically, since \( S(\theta) = 0 \),

\[
\log L(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2 = \log L(\hat{\theta}) - \frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2
\]

and so

\[
\log L(\theta)/L(\hat{\theta}) \approx -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2.
\]

If \( \log L(\theta)/L(\hat{\theta}) \) is ‘well-approximated’ by the quadratic \((-1/2) I(\hat{\theta})(\theta - \hat{\theta})^2\), then the model is said to be regular. (The author only gives this somewhat intuitive definition of regularity.)

**Example 8** Suppose that only the first row of data is considered in Example 6.

\[
58 \ 50 \ 57 \ 58 \ 64 \ 63 \ 54 \ 64 \ 59 \ 41 \ 43 \ 56 \ 60 \ 50
\]
The Fisher information is $I(\theta) = \frac{28}{111}$.

**The Approximation in the Normal Case**

In Example 4, it was shown that for an iid sample from $N(\theta, \sigma^2)$, we have that

$$\log L(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2.$$
It is easy to show that $\hat{\theta} = \bar{x}$ and that $I(\hat{\theta}) = n/\sigma^2$. Thus, $se(\hat{\theta}) = \sigma/\sqrt{n} = I(\hat{\theta})^{-1/2}$. (See Example 2.9, p. 32.) Moreover,

$$\log L(\theta) - \log L(\hat{\theta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\theta})^2$$

$$= \frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} (x_i - \theta)^2 - \sum_{i=1}^{n} (x_i - \hat{\theta})^2 \right]$$

$$= \frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} (x_i^2 - 2x_i\theta + \theta^2) - \sum_{i=1}^{n} (x_i^2 - 2x_i\hat{\theta} + \hat{\theta}^2) \right]$$

$$= \frac{1}{2\sigma^2} \left[ -2n\theta^2 + n\hat{\theta}^2 + 2n\hat{\theta}\theta - n\theta^2 \right]$$

$$= -\frac{n}{2\sigma^2} \left[ 2\theta^2 - \hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2 \right]$$

$$= -\frac{n}{2\sigma^2} \left[ \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2 \right]$$

$$= -\frac{n}{2\sigma^2} (\theta - \hat{\theta})^2$$

$$= -\frac{1}{2} I(\hat{\theta})(\theta - \hat{\theta})^2.$$

Therefore, the approximation is exact in the normal case! Of course we have that $\hat{\theta}$ is $N(\theta, \sigma^2/n)$. So, if the log-likelihood of nonnormal data is approximated well by the quadratic $-1/2I(\hat{\theta})(\theta - \hat{\theta})^2$, then its MLE $\hat{\theta}$ should be approximately normal as well. In particular,

$$se(\hat{\theta}) \approx I(\hat{\theta})^{-1/2}.$$

**Approximating the Score Function**

Since the log-likelihood can be approximated, its derivative, $S(\theta)$, can also be approximated as

$$S(\theta) \approx -I(\hat{\theta})(\theta - \hat{\theta}).$$

In the normal case, the approximation is exact as can be easily verified. Therefore, in the normal case,

$$S(\theta) = -I(\hat{\theta})(\theta - \hat{\theta}).$$

Thus,

$$-S(\theta)I(\hat{\theta})^{-1/2} = I(\theta)^{1/2}(\theta - \hat{\theta}).$$

(Notice that $I(\hat{\theta})^{1/2}(\theta - \hat{\theta})$ is $N(0, 1)$. For nonnormal data, $-S(\theta)I(\hat{\theta})^{-1/2} \approx I(\hat{\theta})^{1/2}(\theta - \hat{\theta})$, and so we can appraise the quality of the quadratic approximation by plotting $-S(\theta)I(\hat{\theta})^{-1/2}$ against $I(\hat{\theta})^{1/2}(\theta - \hat{\theta})$.

**Example 9** Again consider the first row of data is considered in Example 6. The Fisher information is $I(\hat{\theta}) = 28/111$.

$$S(\lambda) = \sum_{i=1}^{n} \frac{x_i}{\lambda} - n.$$
$S(\lambda)$ and Its Linear Approximation