# Likelihood

# 2

## Contents



The term *likelihood* has been introduced by Sir Ronald A. Fisher (1890–1962). The likelihood function forms the basis of likelihood-based statistical inference.

## 2.1 Likelihood and Log-Likelihood Function

Let $X = x$ denote a realisation of a random variable or vector $X$ with probability mass or density function $f(x;\theta)$, cf. Appendix A.2. The function $f(x;\theta)$ depends on the realisation $x$ and on typically unknown parameters $\theta$, but is otherwise assumed to be known. It typically follows from the formulation of a suitable statistical model. Note that $\theta$ can be a scalar or a vector; in the latter case we will write the parameter vector $\boldsymbol{\theta}$ in boldface. The space $\mathcal{T}$ of all possible realisations of $X$ is called *sample space*, whereas the parameter $\theta$ can take values in the *parameter space* $\Theta$.


L. Held, D. Sabanés Bové, *Applied Statistical Inference*,
DOI 10.1007/978-3-642-37887-4_2, © Springer-Verlag Berlin Heidelberg 2014

The function $f(x;\theta)$ describes the distribution of the random variable $X$ for fixed parameter $\theta$. The goal of statistical inference is to infer $\theta$ from the observed datum $X = x$. Playing a central role in this task is the *likelihood function* (or simply *likelihood*)

$$L(\theta; x) = f(x;\theta), \quad \theta \in \Theta,$$

viewed as a function of $\theta$ for fixed $x$. We will often write $L(\theta)$ for the likelihood if it is clear to which observed datum $x$ the likelihood refers to.

**Definition 2.1** (Likelihood function) The *likelihood function* $L(\theta)$ is the probability mass or density function of the observed data $x$, viewed as a function of the unknown parameter $\theta$. ♦

For discrete data, the likelihood function is the probability of the observed data viewed as a function of the unknown parameter $\theta$. This definition is not directly transferable to continuous observations, where the probability of every exactly measured observed datum is strictly speaking zero. However, in reality continuous measurements are always rounded to a certain degree, and the probability of the observed datum $x$ can therefore be written as $\mathsf{Pr}(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2})$ for some small rounding interval width $\varepsilon > 0$. Here $X$ denotes the underlying true continuous measurement.

The above probability can be re-written as

$$\mathsf{Pr}\left(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2}\right) = \int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f(y;\theta)\, dy \approx \varepsilon \cdot f(x;\theta),$$

so the probability of the observed datum $x$ is approximately proportional to the density function $f(x;\theta)$ of $X$ at $x$. As we will see later, the multiplicative constant $\varepsilon$ can be ignored, and we therefore use the density function $f(x;\theta)$ as the likelihood function of a continuous datum $x$.

### 2.1.1 Maximum Likelihood Estimate

Plausible values of $\theta$ should have a relatively high likelihood. The most plausible value with maximum value of $L(\theta)$ is the *maximum likelihood estimate*.

**Definition 2.2** (Maximum likelihood estimate) The *maximum likelihood estimate* (MLE) $\hat{\theta}_{\mathrm{ML}}$ of a parameter $\theta$ is obtained through maximising the likelihood function:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta).$$

♦

In order to compute the MLE, we can safely ignore multiplicative constants in $L(\theta)$, as they have no influence on $\hat{\theta}_{\mathrm{ML}}$. To simplify notation, we therefore often only report a likelihood function $L(\theta)$ without multiplicative constants, i.e. the *likelihood kernel*.

**Definition 2.3** (Likelihood kernel) The likelihood kernel is obtained from a likelihood function by removing all multiplicative constants. We will use the symbol $L(\theta)$ both for likelihood functions and kernels. ♦

It is often numerically convenient to use the *log-likelihood function*

$$l(\theta) = \log L(\theta),$$

the natural logarithm of the likelihood function, for computation of the MLE. The logarithm is a strictly monotone function, and therefore

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} l(\theta).$$

Multiplicative constants in $L(\theta)$ turn to additive constants in $l(\theta)$, which again can often be ignored. A log-likelihood function without additive constants is called *log-likelihood kernel*. We will use the symbol $l(\theta)$ both for log-likelihood functions and kernels.

*Example 2.1* (Inference for a proportion) Let $X \sim \mathrm{Bin}(n, \pi)$ denote a binomially distributed random variable. For example, $X = x$ may represent the observed number of babies with Klinefelter's syndrome among $n$ male newborns. The number of male newborns $n$ is hence known, while the true prevalence $\pi$ of Klinefelter's syndrome among male newborns is unknown.

The corresponding likelihood function is

$$L(\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad \text{for } \pi \in (0,1)$$

with unknown parameter $\pi \in (0,1)$ and sample space $\mathcal{T} = \{0, 1, \ldots, n\}$. The multiplicative term $\binom{n}{x}$ does not depend on $\pi$ and can therefore be ignored, i.e. it is sufficient to consider the likelihood kernel $\pi^x(1-\pi)^{n-x}$. The likelihood function $L(\pi)$ is displayed in Fig. 2.1 for a sample size of $n = 10$ with $x = 2$ and $x = 0$ babies with Klinefelter's syndrome, respectively.

The log-likelihood kernel turns out to be

$$l(\pi) = x \log \pi + (n-x) \log(1-\pi)$$

with derivative

$$\frac{dl(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi}.$$

Setting this derivative to zero gives the MLE $\hat{\pi}_{\mathrm{ML}} = x/n$, the relative frequency of Klinefelter's syndrome in the sample. The MLEs are marked with a vertical line in Fig. 2.1. ■

The uniqueness of the MLE is not guaranteed, and in certain examples there may exist at least two parameter values $\hat{\theta}_1 \neq \hat{\theta}_2$ with $L(\hat{\theta}_1) = L(\hat{\theta}_2) = \arg\max_{\theta \in \Theta} L(\theta)$. In other situations, the MLE may not exist at all. The following example illustrates

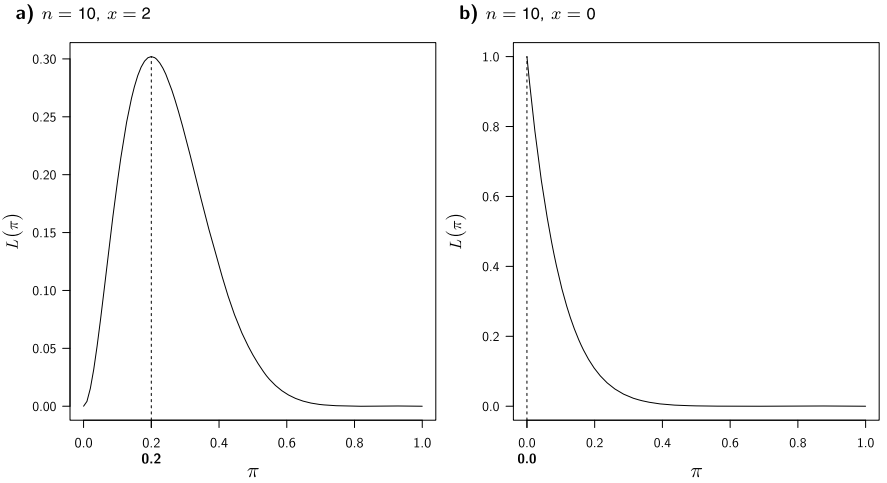


**Fig. 2.1** Likelihood function for $\pi$ in a binomial model. The MLEs are marked with a *vertical line*

that application of the capture–recapture method can result both in non-unique and non-existing MLEs.

*Example 2.2* (Capture–recapture method) As described in Sect. 1.1.3, the goal of capture–recapture methods is to estimate the number $N$ of individuals in a population. To achieve that goal, $M$ individuals are marked and randomly mixed with the total population. A sample of size $n$ without replacement is then drawn, and the number $X = x$ of marked individuals is determined. The suitable statistical model for $X$ is therefore a hypergeometric distribution

$$X \sim \text{HypGeom}(n, N, M)$$

with probability mass function

$$\Pr(X = x) = f(x; \theta = N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for $x \in \mathcal{T} = \{\max\{0, n - (N - M)\}, \ldots, \min(n, M)\}$. The likelihood function for $N$ is therefore

$$L(N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

for $N \in \Theta = \{\max(n, M + n - x), \max(n, M + n - x) + 1, \ldots\}$, where we could have ignored the multiplicative constant $\binom{M}{x}\frac{n!}{(n-x)!}$. Figure 2.2 displays this likelihood function for certain values of $x$, $n$ and $M$. Note that the unknown parameter

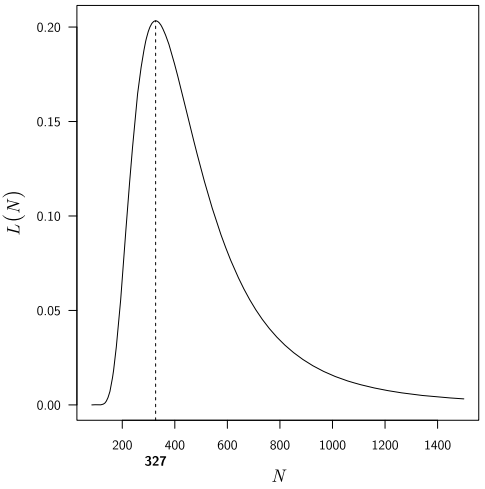


**Fig. 2.2** Likelihood function for $N$ in the capture–recapture experiment with $M = 26$, $n = 63$ and $x = 5$. The (unique) MLE is $\hat{N}_{\mathrm{ML}} = 327$

$\theta = N$ can only take integer values and is not continuous, although the figure suggests the opposite.

It is possible to show (cf. Exercise 3) that the likelihood function is maximised at $\hat{N}_{\mathrm{ML}} = \lfloor M \cdot n/x \rfloor$, where $\lfloor y \rfloor$ denotes the largest integer not greater than $y$. For example, for $M = 26, n = 63$ and $x = 5$ (cf. Fig. 2.2), we obtain $\hat{N}_{\mathrm{ML}} = \lfloor 26 \cdot 63/5 \rfloor = \lfloor 327.6 \rfloor = 327$.

However, sometimes the MLE is not unique, and the likelihood function attains the same value at $\hat{N}_{\mathrm{ML}} - 1$. For example, for $M = 13$, $n = 10$ and $x = 5$, we have $\hat{N}_{\mathrm{ML}} = 13 \cdot 10/5 = 26$, but $\hat{N}_{\mathrm{ML}} = 25$ also attains exactly the same value of $L(N)$. This can easily be verified empirically using the R-function `dhyper`, cf. Table A.1.

```
> M <- 13
> n <- 10
> x <- 5
> ml <- c(25, 26)
> (dhyper(x=x, m=M, n=ml-M, k=n))
[1] 0.311832 0.311832
```

On the other hand, the MLE will not exist for $x = 0$ because the likelihood function $L(N)$ is then monotonically increasing. ■

We often have not only one observation $x$ but a series $x_1, \ldots, x_n$ of $n$ observations from $f(x; \theta)$, usually assumed to be independent. This leads to the concept of a *random sample*.

**Definition 2.4** (Random sample) Data $x_{1:n} = (x_1, \ldots, x_n)$ are realisations of a *random sample* $X_{1:n} = (X_1, \ldots, X_n)$ of size $n$ if the random variables $X_1, \ldots, X_n$ are independent and identically distributed from some distribution with probability mass or density function $f(x; \theta)$. The number $n$ of observations is called the *sample size*. This may be denoted as $X_i \overset{\text{iid}}{\sim} f(x; \theta)$, $i = 1, \ldots, n$. ♦

The probability mass or density function of $X_{1:n}$ is

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

due to assumed independence of the components of $X_{1:n}$. The likelihood function based on a random sample can therefore be written as the product of the individual likelihood contributions $L(\theta; x_i) = f(x_i;\theta)$:

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

The log-likelihood is hence the sum of the individual log-likelihood contributions $l(\theta; x_i) = \log f(x_i;\theta)$:

$$l(\theta; x_{1:n}) = \sum_{i=1}^{n} l(\theta; x_i) = \sum_{i=1}^{n} \log f(x_i;\theta). \tag{2.1}$$

*Example 2.3* (Analysis of survival times) Let $X_{1:n}$ denote a random sample from an exponential distribution Exp($\lambda$). Then

$$L(\lambda) = \prod_{i=1}^{n} \{\lambda \exp(-\lambda x_i)\} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

is the likelihood function of $\lambda \in \mathbb{R}^+$. The log-likelihood function is therefore

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

with derivative

$$\frac{dl(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Setting the derivative to zero, we easily obtain the MLE $\hat{\lambda}_{\mathrm{ML}} = 1/\bar{x}$ where $\bar{x} = \sum_{i=1}^{n} x_i/n$ is the mean observed survival time. If our interest is instead in the theoretical mean $\mu = 1/\lambda$ of the exponential distribution, then the likelihood function takes the form

$$L(\mu) = \mu^{-n} \exp\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i\right), \quad \mu \in \mathbb{R}^+,$$

with MLE $\hat{\mu}_{\mathrm{ML}} = \bar{x}$.

For pure illustration, we now consider the $n = 47$ non-censored PBC survival times from Example 1.1.8 and assume that they are exponentially distributed. We emphasise that this approach is in general not acceptable, as ignoring the censored observations will introduce bias if the distributions of censored and uncensored events differ. It is also less efficient, as a certain proportion of the available data

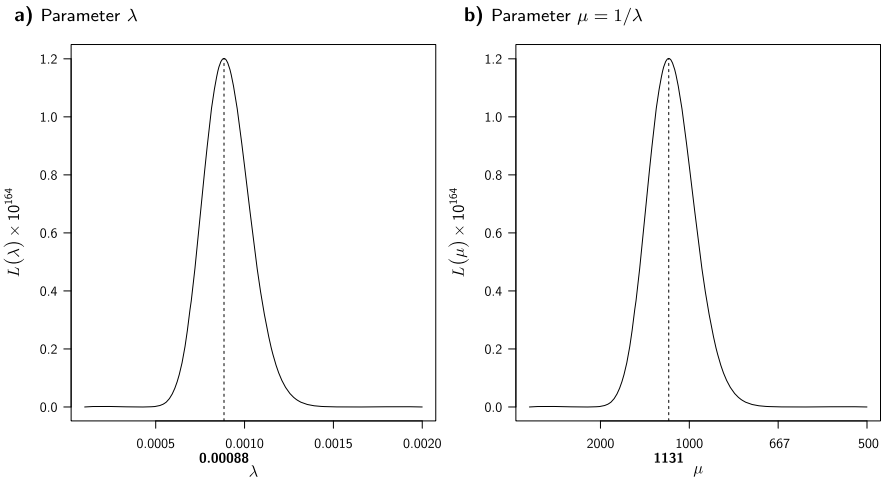


**Fig. 2.3** Likelihood function for λ (*left*) and $\mu$ (*right*) assuming independent and exponentially distributed PBC-survival times. Only uncensored observations are taken into account

is ignored. In Example 2.8 we will therefore also take into account the censored observations.

The likelihood functions for the rate parameter λ and the mean survival time $\mu = 1/\lambda$ are shown in Fig. 2.3. Note that the actual values of the likelihood functions are identical, only the scale of the $x$-axis is transformed. This illustrates that the likelihood function and in particular the MLE are *invariant* with respect to one-to-one transformations of the parameter $\theta$, see Sect. 2.1.3 for more details. It also shows that a likelihood function cannot be interpreted as a density function of a random variable. Indeed, assume that $L(\lambda)$ was an (unnormalised) density function; then the density of $\mu = 1/\lambda$ would be not equal to $L(1/\mu)$ because this change of variables would also involve the derivative of the inverse transformation, cf. Eq. (A.11) in Appendix A.2.3.

The assumption of exponentially distributed survival times may be unrealistic, and a more flexible statistical model may be warranted. Both the *gamma* and the *Weibull* distributions include the exponential distribution as a special case. The Weibull distribution $\mathrm{Wb}(\mu, \alpha)$ is described in Appendix A.5.2 and depends on two parameters $\mu$ and $\alpha$, which both are required to be positive. A random sample $X_{1:n}$ from a Weibull distribution has the density

$$f(x_{1:n}; \mu, \alpha) = \prod_{i=1}^{n} f(x_i; \mu, \alpha) = \prod_{i=1}^{n} \frac{\alpha}{\mu} \left(\frac{x_i}{\mu}\right)^{\alpha-1} \exp\left\{-\left(\frac{x_i}{\mu}\right)^{\alpha}\right\},$$

and the corresponding likelihood function can be written as

$$L(\mu, \alpha) = \frac{\alpha^n}{\mu^{n\alpha}} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left\{-\sum_{i=1}^{n} \left(\frac{x_i}{\mu}\right)^{\alpha}\right\}, \quad \mu, \alpha > 0.$$

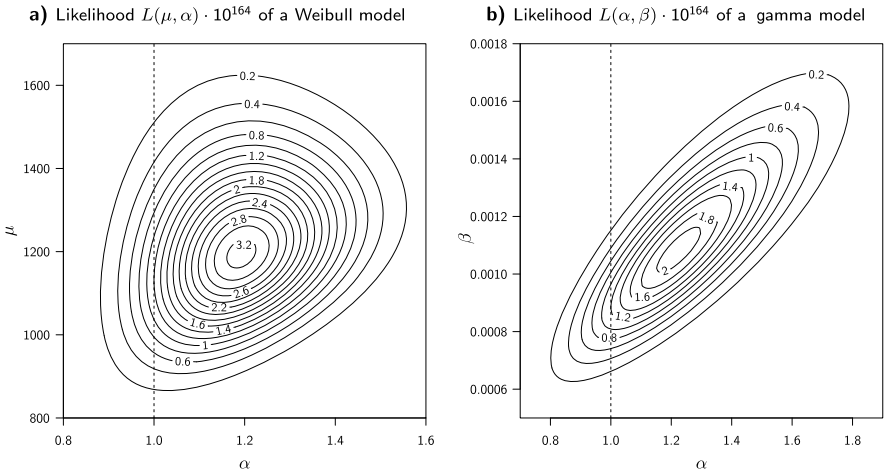


**Fig. 2.4** Flexible modeling of survival times is achieved by a Weibull or gamma model. The corresponding likelihood functions are displayed here. The *vertical line* at $\alpha = 1$ corresponds to the exponential model in both cases

For $\alpha = 1$, we obtain the exponential distribution with expectation $\mu = 1/\lambda$ as a special case.

A contour plot of the Weibull likelihood, a function of two parameters, is displayed in Fig. 2.4a. The likelihood function is maximised at $\alpha = 1.19$, $\mu = 1195$. The assumption of exponentially distributed survival times does not appear to be completely unrealistic, but the likelihood values for $\alpha = 1$ are somewhat lower. In Example 5.9 we will calculate a confidence interval for $\alpha$, which can be used to quantify the plausibility of the exponential model.

If we assume that the random sample comes from a gamma distribution $\mathrm{G}(\alpha, \beta)$, the likelihood is (cf. again Appendix A.5.2)

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \exp(-\beta x_i) = \left\{ \frac{\beta^\alpha}{\Gamma(\alpha)} \right\}^n \left( \prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left( -\beta \sum_{i=1}^{n} x_i \right).$$

The exponential distribution with parameter $\lambda = \beta$ corresponds to the special case $\alpha = 1$. Plausible values $\alpha$ and $\beta$ of the gamma likelihood function tend to lie on the diagonal in Fig. 2.4b: for larger values of $\alpha$, plausible values of $\beta$ tend to be also larger. The sample is apparently informative about the mean $\mu = \alpha/\beta$, but not so informative about the components $\alpha$ and $\beta$ of that ratio.

Alternatively, the gamma likelihood function can be *reparametrised*, and the parameters $\mu = \alpha/\beta$ and $\phi = 1/\beta$, say, could be used. The second parameter $\phi$ now represents the variance-to-mean ratio of the gamma distribution. Figure 2.5 displays the likelihood function using this new parametrisation. The dependence between the two parameters appears to be weaker than for the initial parametrisation shown in Fig. 2.4b. ■

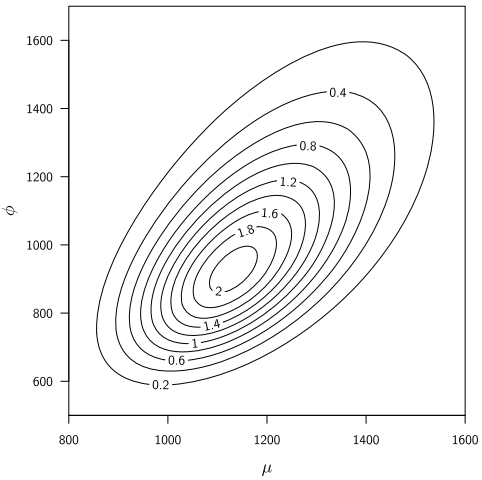


**Fig. 2.5** Likelihood $L(\mu, \phi) \cdot 10^{164}$ of the reparametrised gamma model

A slightly less restrictive definition of a random sample still requires independence, but no longer that the components $X_i$ do all have the same distribution. For example, they may still belong to the same distribution family, but with different parameters.

*Example 2.4* (Poisson model) Consider Example 1.1.6 and denote the observed and expected number of cancer cases in the $n = 56$ regions of Scotland with $x_i$ and $e_i$, respectively, $i = 1, \ldots, n$. The simplest model for such registry data assumes that the underlying relative risk $\lambda$ is the same in all regions and that the observed counts $x_i$'s constitute independent realisations from Poisson distributions with means $e_i\lambda$. The random variables $X_i$ hence belong to the same distributional family but are not identically distributed since the mean parameter $e_i\lambda$ varies from observation to observation.

The log-likelihood kernel of the relative risk $\lambda$ turns out to be

$$l(\lambda) = \sum_{i=1}^{n} x_i \log\lambda - \sum_{i=1}^{n} e_i\lambda,$$

and the MLE of $\lambda$ is

$$\hat{\lambda}_{\mathrm{ML}} = \sum_{i=1}^{n} x_i \Big/ \sum_{i=1}^{n} e_i = \bar{x}/\bar{e},$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{e} = \sum_{i=1}^{n} e_i/n$ denote the mean observed and expected number of cases, respectively. ■

### 2.1.2 Relative Likelihood

It is often useful to consider the likelihood (or log-likelihood) function relative to its value at the MLE.

**Definition 2.5** (Relative likelihood) The *relative likelihood* is

$$\tilde{L}(\theta) = \frac{L(\theta)}{L(\hat{\theta}_{\mathrm{ML}})}.$$

In particular we have $0 \leq \tilde{L}(\theta) \leq 1$ and $\tilde{L}(\hat{\theta}_{\mathrm{ML}}) = 1$. The relative likelihood is also called the normalised likelihood.

Taking the logarithm of the relative likelihood gives the *relative log-likelihood*

$$\tilde{l}(\theta) = \log \tilde{L}(\theta) = l(\theta) - l(\hat{\theta}_{\mathrm{ML}}),$$

where we have $-\infty < \tilde{l}(\theta) \leq 0$ and $\tilde{l}(\hat{\theta}_{\mathrm{ML}}) = 0$. ♦

*Example 2.5* (Inference for a proportion) All different likelihood functions are displayed for a binomial model (cf. Example 2.1) with sample size $n = 10$ and observation $x = 2$ in Fig. 2.6. Note that the change from an ordinary to a relative likelihood changes the scaling of the $y$-axis, but the shape of the likelihood function remains the same. This is also true for the log-likelihood function. ■

It is important to consider the entire likelihood function as the carrier of the information regarding $\theta$ provided by the data. This is far more informative than to consider only the MLE and to disregard the likelihood function itself. Using the values of the relative likelihood function gives us a method to derive a set of parameter values (usually an interval), which are supported by the data. For example, the following categorisation based on thresholding the relative likelihood function using the cutpoints $1/3$, $1/10$, $1/100$ and $1/1000$ has been proposed:

$$\begin{aligned}
1 \geq \tilde{L}(\theta) > \frac{1}{3} &\qquad \theta \text{ very plausible,} \\
\frac{1}{3} \geq \tilde{L}(\theta) > \frac{1}{10} &\qquad \theta \text{ plausible,} \\
\frac{1}{10} \geq \tilde{L}(\theta) > \frac{1}{100} &\qquad \theta \text{ less plausible,} \\
\frac{1}{100} \geq \tilde{L}(\theta) > \frac{1}{1000} &\qquad \theta \text{ barely plausible,} \\
\frac{1}{1000} \geq \tilde{L}(\theta) \geq 0 &\qquad \theta \text{ not plausible.}
\end{aligned}$$

However, such a *pure likelihood* approach to inference has the disadvantage that the scale and the thresholds are somewhat arbitrarily chosen. Indeed, the likelihood on its own does not allow us to quantify the support for a certain set of parameter values

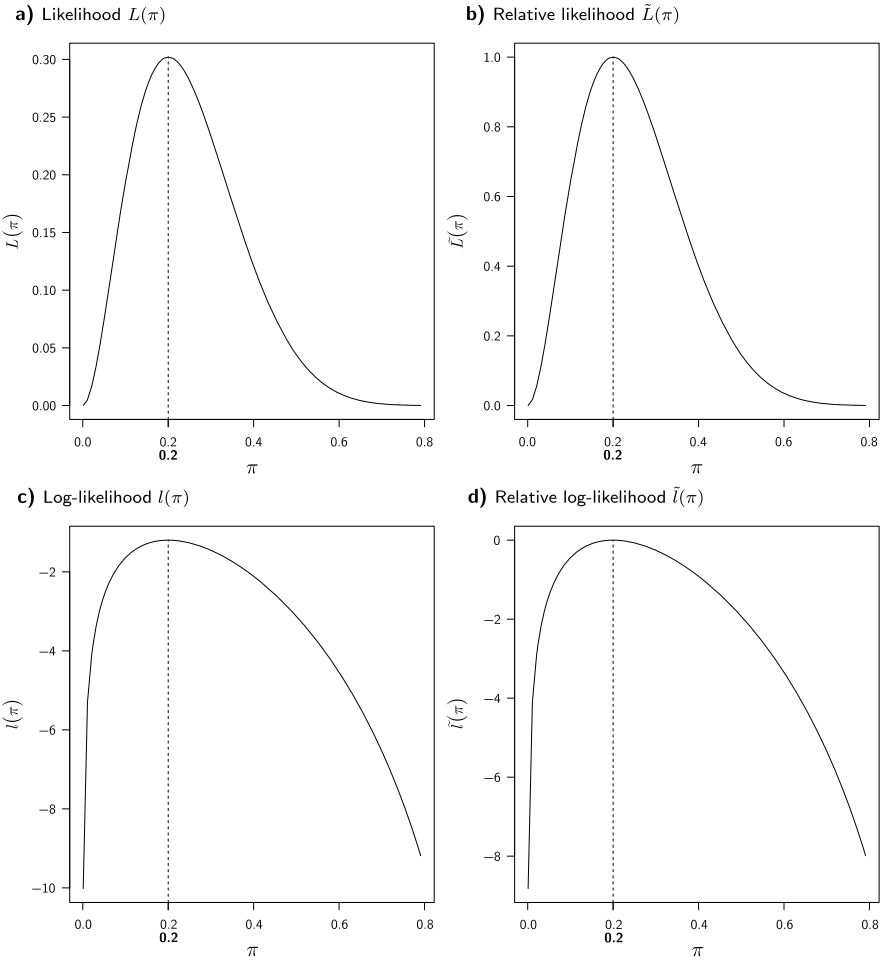


**Fig. 2.6** Various likelihood functions in a binomial model with $n = 10$ and $x = 2$

using probabilities. In Chap. 4 we will describe different approaches to *calibrate* the likelihood based on the concept of a *confidence interval*. Alternatively, a Bayesian approach can be employed, combining the likelihood with a *prior distribution* for $\theta$ and using the concept of a *credible interval*. This approach is outlined in Chap. 6.

### 2.1.3 Invariance of the Likelihood

Suppose we parametrise the distribution of $X$ not with respect to $\theta$, but with respect to a one-to-one transformation $\phi = h(\theta)$. The likelihood function $L_\phi(\phi)$ for $\phi$ and

the likelihood function $L_\theta(\theta)$ for $\theta$ are related as follows:

$$L_\theta(\theta) = L_\theta\left\{h^{-1}(\phi)\right\} = L_\phi(\phi).$$

The actual value of the likelihood will not be changed by this transformation, i.e. the likelihood is *invariant* with respect to one-to-one parameter transformations. We therefore have

$$\hat{\phi}_{\mathrm{ML}} = h(\hat{\theta}_{\mathrm{ML}})$$

for the MLEs $\hat{\phi}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{ML}}$. This is an important property of the maximum likelihood estimate:

**Invariance of the MLE**
Let $\hat{\theta}_{\mathrm{ML}}$ be the MLE of $\theta$, and let $\phi = h(\theta)$ be a one-to-one transformation of $\theta$. The MLE of $\phi$ can be obtained by inserting $\hat{\theta}_{\mathrm{ML}}$ in $h(\theta)$: $\hat{\phi}_{\mathrm{ML}} = h(\hat{\theta}_{\mathrm{ML}})$.

*Example 2.6* (Binomial model) Let $X \sim \mathrm{Bin}(n, \pi)$, so that $\hat{\pi}_{\mathrm{ML}} = x/n$. Now consider the corresponding odds parameter $\omega = \pi/(1-\pi)$. The MLE of $\omega$ is

$$\hat{\omega}_{\mathrm{ML}} = \frac{\hat{\pi}_{\mathrm{ML}}}{1-\hat{\pi}_{\mathrm{ML}}} = \frac{\frac{x}{n}}{1-\frac{x}{n}} = \frac{x}{n-x}.$$

Without knowledge of the invariance property of the likelihood function, we would have to derive the likelihood function with respect to $\omega$ and subsequently maximise it directly. We will do this now for illustrative purposes only.

The log-likelihood kernel for $\pi$ is

$$l_\pi(\pi) = x\log(\pi) + (n-x)\log(1-\pi).$$

We also have

$$\omega = h(\pi) = \frac{\pi}{1-\pi} \quad \Longleftrightarrow \quad \pi = h^{-1}(\omega) = \frac{\omega}{1+\omega} \quad \text{and} \quad 1-\pi = \frac{1}{1+\omega}$$

and therefore

$$\begin{aligned} l_\omega(\omega) &= l_\pi\left\{h^{-1}(\omega)\right\} \\ &= x\log\left(\frac{\omega}{1+\omega}\right) + (n-x)\log\left(\frac{1}{1+\omega}\right) \\ &= x\log(\omega) - n\log(1+\omega). \end{aligned}$$

The derivative with respect to $\omega$ turns out to be

$$\frac{dl_\omega(\omega)}{d\omega} = \frac{x}{\omega} - \frac{n}{1+\omega},$$

so the root $\hat{\omega}_{\mathrm{ML}}$ must fulfill $x(1+\hat{\omega}_{\mathrm{ML}}) = n\,\hat{\omega}_{\mathrm{ML}}$. We easily obtain

$$\hat{\omega}_{\mathrm{ML}} = \frac{x}{n-x}.$$

■

*Example 2.7* (Hardy–Weinberg Equilibrium) Let us now consider Example 1.1.4 and the observed frequencies $x_1 = 233$, $x_2 = 385$ and $x_3 = 129$ of the three genotypes MM, MN and NN. Assuming Hardy–Weinberg equilibrium, the multinomial log-likelihood kernel

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{3} x_i \log(\pi_i)$$

can be written with (1.1) as

$$\begin{aligned} l(\upsilon) &= x_1 \log\left(\upsilon^2\right) + x_2 \log\left\{2\upsilon(1-\upsilon)\right\} + x_3 \log\left\{(1-\upsilon)^2\right\} \\ &= 2x_1 \log(\upsilon) + \underbrace{x_2 \log(2)}_{=\text{const}} + x_2 \log(\upsilon) + x_2 \log(1-\upsilon) + 2x_3 \log(1-\upsilon) \\ &= (2x_1 + x_2)\log(\upsilon) + (x_2 + 2x_3)\log(1-\upsilon) + \text{const}. \end{aligned}$$

The log-likelihood kernel for the allele frequency $\upsilon$ is therefore $(2x_1 + x_2)\log(\upsilon) + (x_2 + 2x_3)\log(1-\upsilon)$, which can be identified as a binomial log-likelihood kernel for the success probability $\upsilon$ with $2x_1 + x_2$ successes and $x_2 + 2x_3$ failures.

The MLE of $\upsilon$ is therefore

$$\hat{\upsilon}_{\mathrm{ML}} = \frac{2x_1 + x_2}{2x_1 + 2x_2 + 2x_3} = \frac{2x_1 + x_2}{2n} = \frac{x_1 + x_2/2}{n},$$

which is exactly the proportion of A alleles in the sample. For the data above, we obtain $\hat{\upsilon}_{\mathrm{ML}} \approx 0.570$. The MLEs of $\pi_1, \pi_2$ and $\pi_3$ (assuming Hardy–Weinberg equilibrium) are therefore

$$\begin{aligned} &\hat{\pi}_1 = \hat{\upsilon}_{\mathrm{ML}}^2 \approx 0.324, \qquad \hat{\pi}_2 = 2\hat{\upsilon}_{\mathrm{ML}}(1-\hat{\upsilon}_{\mathrm{ML}}) \approx 0.490 \quad \text{and} \\ &\hat{\pi}_3 = (1-\hat{\upsilon}_{\mathrm{ML}})^2 \approx 0.185, \end{aligned}$$

using the invariance property of the likelihood. ■

In the last example, the transformation to which the MLE is invariant is not really a one-to-one transformation. A more detailed view of the situation is the following: We have the more general multinomial model with two parameters $\pi_1$, $\pi_2$ ($\pi_3$ is determined by these) and the simpler Hardy–Weinberg model with one parameter $\upsilon$. We can restrict the multinomial model to the Hardy–Weinberg model, which is hence a special case of the multinomial model. If we obtain an MLE for $\upsilon$, we can hence calculate the resulting MLEs for $\pi_1$, $\pi_2$ and also $\pi_3$. However, in the other di-

rection, i.e. by first calculating the unrestricted MLE $\hat{\boldsymbol{\pi}}_{\mathrm{ML}}$ in the multinomial model, we could not calculate a corresponding MLE $\hat{\upsilon}_{\mathrm{ML}}$ in the simpler Hardy–Weinberg model.

### 2.1.4 Generalised Likelihood

Declaring probability mass and density functions as appropriate likelihood functions is not always sufficient. In some situations this definition must be suitably generalised. A typical example is the analysis of survival data with some observations being censored.

Assume that observed survival times $x_1, \ldots, x_n$ are independent realisations from a distribution with density function $f(x;\theta)$ and corresponding distribution function $F(x;\theta) = \mathsf{Pr}(X \le x;\theta)$. The likelihood contribution of a non-censored observation $x_i$ is then (as before) $f(x_i;\theta)$. However, a censored observation will contribute the term $1 - F(x_i;\theta) = \mathsf{Pr}(X_i > x_i;\theta)$ to the likelihood since in this case we only know that the actual (unobserved) survival time is larger than $x_i$.

Compact notation can be achieved using the *censoring indicator* $\delta_i$, $i = 1, \ldots, n$, with $\delta_i = 0$ if the survival time $x_i$ is censored and $\delta_i = 1$ if it is observed. Due to independence of the observations, the likelihood can be written as

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i} \left\{1 - F(x_i;\theta)\right\}^{1-\delta_i}. \tag{2.2}$$

*Example 2.8* (Analysis of survival times) A simple statistical model to describe survival times is to assume an exponential distribution with density and distribution function

$$f(x) = \lambda \exp(-\lambda x) \quad \text{and}$$
$$F(x) = 1 - \exp(-\lambda x),$$

respectively, so $1 - F(x) = \exp(-\lambda x)$. The likelihood function (2.2) now reduces to

$$L(\lambda) = \lambda^{n\bar{\delta}} \exp(-\lambda n \bar{x}),$$

where $\bar{\delta}$ is the observed proportion of uncensored observations, and $\bar{x}$ is the mean observed survival time of all (censored and uncensored) observations. The MLE is $\hat{\lambda}_{\mathrm{ML}} = \bar{\delta}/\bar{x}$. Due to invariance of the MLE, the estimate for the mean $\mu = 1/\lambda$ is $\hat{\mu}_{\mathrm{ML}} = \bar{x}/\bar{\delta}$.

Among the $n = 94$ observations from Example 1.1.8, there are $\sum_{i=1}^{n} \delta_i = 47$ uncensored, and the total follow-up time is $\sum_{i=1}^{n} x_i = 143192$ days. The estimated rate is $\hat{\lambda}_{\mathrm{ML}} = 47/143\,192 = 32.82$ per 100 000 days, and the MLE of the expected survival time is $\hat{\mu}_{\mathrm{ML}} = 3046.6$ days. This is substantially larger than in the analysis of the uncensored observations only (cf. Example 2.3), where we have obtained the estimate $\hat{\mu}_{\mathrm{ML}} = 1130.8$ days. ■

## 2.2 Score Function and Fisher Information

The MLE of $\theta$ is obtained by maximising the (relative) likelihood function,

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \tilde{L}(\theta).$$

For numerical reasons, it is often easier to maximise the log-likelihood $l(\theta) = \log L(\theta)$ or the relative log-likelihood $\tilde{l}(\theta) = l(\theta) - l(\hat{\theta}_{\mathrm{ML}})$ (cf. Sect. 2.1), which yields the same result since

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} l(\theta) = \arg\max_{\theta \in \Theta} \tilde{l}(\theta).$$

However, the log-likelihood function $l(\theta)$ has much larger importance, besides simplifying the computation of the MLE. Especially, its first and second derivatives are important and have their own names, which are introduced in the following. For simplicity, we assume that $\theta$ is a scalar.

**Definition 2.6** (Score function) The first derivative of the log-likelihood function

$$S(\theta) = \frac{dl(\theta)}{d\theta}$$

is called the *score function*. ♦

Computation of the MLE is typically done by solving the *score equation* $S(\theta) = 0$.

The second derivative, the curvature, of the log-likelihood function is also of central importance and has its own name.

**Definition 2.7** (Fisher information) The negative second derivative of the log-likelihood function

$$I(\theta) = -\frac{d^2 l(\theta)}{d\theta^2} = -\frac{dS(\theta)}{d\theta}$$

is called the *Fisher information*. The value of the Fisher information at the MLE $\hat{\theta}_{\mathrm{ML}}$, i.e. $I(\hat{\theta}_{\mathrm{ML}})$, is the *observed Fisher information*. ♦

Note that the MLE $\hat{\theta}_{\mathrm{ML}}$ is a function of the observed data, which explains the terminology "observed" Fisher information for $I(\hat{\theta}_{\mathrm{ML}})$.

*Example 2.9* (Normal model) Suppose we have realisations $x_{1:n}$ of a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown mean $\mu$ and known vari-

ance $\sigma^2$. The log-likelihood kernel and score function are then

$$l(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \text{and}$$

$$S(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu),$$

respectively. The solution of the score equation $S(\mu) = 0$ is the MLE $\hat{\mu}_{\mathrm{ML}} = \bar{x}$. Taking another derivative gives the Fisher information

$$I(\mu) = \frac{n}{\sigma^2},$$

which does not depend on $\mu$ and so is equal to the observed Fisher information $I(\hat{\mu}_{\mathrm{ML}})$, no matter what the actual value of $\hat{\mu}_{\mathrm{ML}}$ is.

Suppose we switch the roles of the two parameters and treat $\mu$ as known and $\sigma^2$ as unknown. We now obtain

$$\hat{\sigma}^2_{\mathrm{ML}} = \sum_{i=1}^{n}(x_i - \mu)^2/n$$

with Fisher information

$$I\left(\sigma^2\right) = \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\sigma^4}.$$

The Fisher information of $\sigma^2$ now really depends on its argument $\sigma^2$. The observed Fisher information turns out to be

$$I\left(\hat{\sigma}^2_{\mathrm{ML}}\right) = \frac{n}{2\hat{\sigma}^4_{\mathrm{ML}}}.$$

■

It is instructive at this stage to adopt a *frequentist* point of view and to consider the MLE $\hat{\mu}_{\mathrm{ML}} = \bar{x}$ from Example 2.9 as a random variable, i.e. $\hat{\mu}_{\mathrm{ML}} = \bar{X}$ is now a function of the random sample $X_{1:n}$. We can then easily compute $\mathrm{Var}(\hat{\mu}_{\mathrm{ML}}) = \mathrm{Var}(\bar{X}) = \sigma^2/n$ and note that

$$\mathrm{Var}(\hat{\mu}_{\mathrm{ML}}) = \frac{1}{I(\hat{\mu}_{\mathrm{ML}})}$$

holds. In Sect. 4.2.3 we will see that this equality is approximately valid for other statistical models. Indeed, under certain regularity conditions, the variance $\mathrm{Var}(\hat{\theta}_{\mathrm{ML}})$ of the MLE turns out to be approximately equal to the inverse observed Fisher information $1/I(\hat{\theta}_{\mathrm{ML}})$, and the accuracy of this approximation improves with increasing sample size $n$. Example 2.9 is a special case, where this equality holds exactly for any sample size.

*Example 2.10* (Binomial model) The score function of a binomial observation $X = x$ with $X \sim \text{Bin}(n, \pi)$ is

$$S(\pi) = \frac{dl(\pi)}{d\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi}$$

and has been derived already in Example 2.1. Taking the derivative of $S(\pi)$ gives the Fisher information

$$\begin{aligned} I(\pi) &= -\frac{d^2 l(\pi)}{d\pi^2} = -\frac{dS(\pi)}{d\pi} \\ &= \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2} \\ &= n\left\{\frac{x/n}{\pi^2} + \frac{(n-x)/n}{(1-\pi)^2}\right\}. \end{aligned}$$

Plugging in the MLE $\hat{\pi}_{\text{ML}} = x/n$, we finally obtain the observed Fisher information

$$I(\hat{\pi}_{\text{ML}}) = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})}.$$

This result is plausible if we take a frequentist point of view and consider the MLE as a random variable. Then

$$\text{Var}(\hat{\pi}_{\text{ML}}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(X) = \frac{1}{n^2} n\pi(1-\pi) = \frac{\pi(1-\pi)}{n},$$

so the variance of $\hat{\pi}_{\text{ML}}$ has the same form as the inverse observed Fisher information; the only difference is that the MLE $\hat{\pi}_{\text{ML}}$ is replaced by the true (and unknown) parameter $\pi$. The inverse observed Fisher information is hence an estimate of the variance of the MLE. ■

How does the observed Fisher information change if we reparametrise our statistical model? Here is the answer to this question.

**Result 2.1** (Observed Fisher information after reparametrisation) *Let $I_\theta(\hat{\theta}_{\text{ML}})$ denote the observed Fisher information of a scalar parameter $\theta$ and suppose that $\phi = h(\theta)$ is a one-to-one transformation of $\theta$. The observed Fisher information $I_\phi(\hat{\phi}_{\text{ML}})$ of $\phi$ is then*

$$I_\phi(\hat{\phi}_{\text{ML}}) = I_\theta(\hat{\theta}_{\text{ML}}) \left\{\frac{dh^{-1}(\hat{\phi}_{\text{ML}})}{d\phi}\right\}^2 = I_\theta(\hat{\theta}_{\text{ML}}) \left\{\frac{dh(\hat{\theta}_{\text{ML}})}{d\theta}\right\}^{-2}. \tag{2.3}$$

*Proof* The transformation $h$ is assumed to be one-to-one, so $\theta = h^{-1}(\phi)$ and $l_\phi(\phi) = l_\theta\{h^{-1}(\phi)\}$. Application of the chain rule gives

$$\begin{aligned} S_\phi(\phi) &= \frac{dl_\phi(\phi)}{d\phi} = \frac{dl_\theta\{h^{-1}(\phi)\}}{d\phi} \\ &= \frac{dl_\theta(\theta)}{d\theta} \cdot \frac{dh^{-1}(\phi)}{d\phi} \\ &= S_\theta(\theta) \cdot \frac{dh^{-1}(\phi)}{d\phi}. \end{aligned} \tag{2.4}$$

The second derivative of $l_\phi(\phi)$ can be computed using the product and chain rules:

$$\begin{aligned} I_\phi(\phi) &= -\frac{dS_\phi(\phi)}{d\phi} = -\frac{d}{d\phi}\left\{ S_\theta(\theta) \cdot \frac{dh^{-1}(\phi)}{d\phi} \right\} \\ &= -\frac{dS_\theta(\theta)}{d\phi} \cdot \frac{dh^{-1}(\phi)}{d\phi} - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2} \\ &= -\frac{dS_\theta(\theta)}{d\theta} \cdot \left\{ \frac{dh^{-1}(\phi)}{d\phi} \right\}^2 - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2} \\ &= I_\theta(\theta) \left\{ \frac{dh^{-1}(\phi)}{d\phi} \right\}^2 - S_\theta(\theta) \cdot \frac{d^2h^{-1}(\phi)}{d\phi^2}. \end{aligned}$$

Evaluating $I_\phi(\phi)$ at the MLE $\phi = \hat{\phi}_{\text{ML}}$ (so $\theta = \hat{\theta}_{\text{ML}}$) leads to the first equation in (2.3) (note that $S_\theta(\hat{\theta}_{\text{ML}}) = 0$). The second equation follows with

$$\frac{dh^{-1}(\phi)}{d\phi} = \left\{ \frac{dh(\theta)}{d\theta} \right\}^{-1} \quad \text{for } \frac{dh(\theta)}{d\theta} \neq 0. \tag{2.5}$$

□

*Example 2.11* (Binomial model) In Example 2.6 we saw that the MLE of the odds $\omega = \pi/(1-\pi)$ is $\hat{\omega}_{\text{ML}} = x/(n-x)$. What is the corresponding observed Fisher information? First, we compute the derivative of $h(\pi) = \pi/(1-\pi)$, which is

$$\frac{dh(\pi)}{d\pi} = \frac{1}{(1-\pi)^2}.$$

Using the observed Fisher information of $\pi$ derived in Example 2.10, we obtain

$$\begin{aligned} I_\omega(\hat{\omega}_{\text{ML}}) &= I_\pi(\hat{\pi}_{\text{ML}}) \left\{ \frac{dh(\hat{\pi}_{\text{ML}})}{d\pi} \right\}^{-2} = \frac{n}{\hat{\pi}_{\text{ML}}(1-\hat{\pi}_{\text{ML}})} \cdot (1-\hat{\pi}_{\text{ML}})^4 \\ &= n \cdot \frac{(1-\hat{\pi}_{\text{ML}})^3}{\hat{\pi}_{\text{ML}}} = \frac{(n-x)^3}{nx}. \end{aligned}$$

As a function of $x$ for fixed $n$, the observed Fisher information $I_\omega(\hat{\omega}_{\text{ML}})$ is monotonically decreasing (the numerator is monotonically decreasing, and the denominator is monotonically increasing). In other words, the observed Fisher information increases with decreasing MLE $\hat{\omega}_{\text{ML}}$.

The observed Fisher information of the log odds $\phi = \log(\omega)$ can be similarly computed, and we obtain

$$I_\phi(\hat{\phi}_{\text{ML}}) = I_\omega(\hat{\omega}_{\text{ML}})\left(\frac{1}{\hat{\omega}_{\text{ML}}}\right)^{-2} = \frac{(n-x)^3}{nx} \cdot \frac{x^2}{(n-x)^2} = \frac{x(n-x)}{n}.$$

Note that $I_\phi(\hat{\phi}_{\text{ML}})$ does not change if we redefine successes as failures and vice versa. This is also the case for the observed Fisher information $I_\pi(\hat{\pi}_{\text{ML}})$ but not for $I_\omega(\hat{\omega}_{\text{ML}})$. ■

## 2.3 Numerical Computation of the Maximum Likelihood Estimate

Explicit formulas for the MLE and the observed Fisher information can typically only be derived in simple models. In more complex models, numerical techniques have to be applied to compute maximum and curvature of the log-likelihood function. We first describe the application of general purpose optimisation algorithms to this setting and will discuss the Expectation-Maximisation (EM) algorithm in Sect. 2.3.2.

### 2.3.1 Numerical Optimisation

Application of the *Newton–Raphson algorithm* (cf. Appendix C.1.3) requires the first two derivatives of the function to be maximised, so for maximising the log-likelihood function, we need the score function and the Fisher information. Iterative application of the equation

$$\theta^{(t+1)} = \theta^{(t)} + \frac{S(\theta^{(t)})}{I(\theta^{(t)})}$$

gives after convergence (i.e. $\theta^{(t+1)} = \theta^{(t)}$) the MLE $\hat{\theta}_{\text{ML}}$. As a by-product, the observed Fisher information $I(\hat{\theta}_{\text{ML}})$ can also be extracted.

To apply the Newton–Raphson algorithm in `R`, the function `optim` can conveniently be used, see Appendix C.1.3 for details. We need to pass the log-likelihood function as an argument to `optim`. Explicitly passing the score function into `optim` typically accelerates convergence. If the derivative is not available, it can sometimes be computed symbolically using the `R` function `deriv`. Generally no derivatives need to be passed to `optim` because it can approximate them numerically. Particularly useful is the option `hessian = TRUE`, in which case `optim` will also return the negative observed Fisher information.

*Example 2.12* (Screening for colon cancer) The goal of Example 1.1.5 is to estimate the false negative fraction of a screening test, which consists of six consecutive medical examinations. Let $\pi$ denote the probability for a positive test result of the $i$th diseased individual and denote by $X_i$ the number of positive test results among the six examinations. We start by assuming that individual test results are independent and that $\pi$ does not vary from patient to patient (two rather unrealistic assumptions, as we will see later), so that $X_i$ is binomially distributed: $X_i \sim \text{Bin}(N = 6, \pi)$. However, due to the study design, we will not observe a patient with $X_i = 0$ positive tests. We therefore need to use the *truncated binomial distribution* as the appropriate statistical model. The corresponding log-likelihood can be derived by considering

$$\Pr(X_i = k \mid X_i > 0) = \frac{\Pr(X_i = k)}{\Pr(X_i > 0)}, \quad k = 1, \ldots, 6, \tag{2.6}$$

and turns out to be (cf. Example C.1 in Appendix C)

$$l(\pi) = \sum_{k=1}^{N} Z_k \big\{ k \log(\pi) + (N-k) \log(1-\pi) \big\} - n \log\big\{ 1 - (1-\pi)^N \big\}. \tag{2.7}$$

Here $Z_k$ denotes the number of patients with $k$ positive test results, and $n = \sum_{k=1}^{N} Z_k = 196$ is the total number of diseased patients with at least one positive test result.

Computation of the MLE is now most conveniently done with numerical techniques. To do so, we write an `R` function `log.likelihood`, which returns the log-likelihood kernel of the unknown probability (`pi`) for a given vector of counts (`data`) and maximise, it with the `optim` function.

```
> ## Truncated binomial log-likelihood function
> ## pi: the parameter, the probability of a positive test result
> ## data: vector with counts Z_1, ..., Z_N
> log.likelihood <- function(pi, data)
  {
      n <- sum(data)
      k <- length(data)
      vec <- seq_len(k)
      result <- sum(data * (vec * log(pi) + (k-vec) * log(1-pi))) -
             n * log(1 - (1-pi)^k)

  }
> data <- c(37, 22, 25, 29, 34, 49)
> eps <- 1e-10
> result <- optim(0.5, log.likelihood, data = data,
                  method = "L-BFGS-B", lower = eps, upper = 1-eps,
                  control = list(fnscale = -1), hessian = TRUE)
> (ml <- result$par)
[1] 0.6240838
```

The MLE turns out to be $\hat{\pi}_{\text{ML}} = 0.6241$, and the observed Fisher information $I(\hat{\pi}_{\text{ML}}) = 4885.3$ is computed via

```
> (observed.fisher <- - result$hessian)
         [,1]
[1,] 4885.251
```

Invariance of the MLE immediately gives the MLE of the false negative fraction $\xi = \mathsf{Pr}(X_i = 0)$ via

$$\hat{\xi}_{\mathrm{ML}} = (1 - \hat{\pi}_{\mathrm{ML}})^N = 0.0028.$$

A naive estimate of the number of undetected cancer cases $Z_0$ can be obtained by solving $\hat{Z}_0/(196 + \hat{Z}_0) = \hat{\xi}_{\mathrm{ML}}$ for $\hat{Z}_0$:

$$\hat{Z}_0 = 196 \cdot \frac{\hat{\xi}_{\mathrm{ML}}}{1 - \hat{\xi}_{\mathrm{ML}}} = 0.55. \tag{2.8}$$

It turns out that this estimate can be justified as a maximum likelihood estimate. To see this, note that the probability to detect a cancer case in one particular application of the six-stage screening test is $1 - \xi$. The number of samples until the first cancer case is detected therefore follows a geometric distribution with success probability $1 - \xi$, cf. Table A.1 in Appendix A.5.1. Thus, if $n$ is the observed number of detected cancer cases, the total number of cancer cases $Z_0 + n$ follows a negative binomial distribution with parameters $n$ and $1 - \xi$ (again cf. Appendix A.5.1):

$$Z_0 + n \sim \mathrm{NBin}(n, 1 - \xi),$$

so the expectation of $Z_0 + n$ is $\mathsf{E}(Z_0 + n) = n/(1 - \xi)$. Our interest is in the expectation of $Z_0$,

$$\mathsf{E}(Z_0) = \frac{n}{1 - \xi} - n = n \cdot \frac{\xi}{1 - \xi}. \tag{2.9}$$

The MLE (2.8) of $\mathsf{E}(Z_0)$ can now easily be obtained by replacing $\xi$ with $\hat{\xi}_{\mathrm{ML}}$.

A closer inspection of Table 1.2 makes 0.55 expected undetected cancer cases rather implausible and raises questions about the appropriateness of the binomial model. Indeed, a naive extrapolation of the observed frequencies $Z_k$, $k = 1, \ldots, 6$, leads to a considerably larger values of $Z_0$. The fact that the binomial model does not fit the data well can also be seen from the frequencies in Table 1.2 with a "bathtub" shape with extreme values $k = 1$ and $k = 5, 6$ more likely than the intermediate values $k = 2, 3, 4$. Such a form is impossible with a (truncated) binomial distribution. This can also be seen from the expected frequencies under the binomial model with $\hat{\pi}_{\mathrm{ML}} = 0.6241$, which are 5.5, 22.9, 50.8, 63.2, 42.0 and 11.6 for $k = 1, \ldots, 6$. The difference to the observed frequencies $Z_k$ is quite striking. In Example 5.18 we will apply a goodness-of-fit test, which quantifies the evidence against this model. Earlier in Chap. 5 we will introduce an alternative model with two parameters, which describes the observed frequencies much better. ■

### 2.3.2 The EM Algorithm

The *Expectation-Maximisation (EM) algorithm* is an iterative algorithm to compute MLEs. In certain situations it is particularly easy to apply, as the following example illustrates.

*Example 2.13* (Screening for colon cancer) We reconsider Example 1.1.5, where the number $Z_k$ of 196 cancer patients with $k \geq 1$ positive test results among six cancer colon screening tests has been recorded. Due to the design of the study, we have no information on the number $Z_0$ of patients with solely negative test results (cf. Table 1.2). Numerical techniques allow us to fit a truncated binomial distribution to the observed data $Z_1, \ldots, Z_6$, cf. Example 2.12.

However, the EM algorithm could also be used to compute the MLEs. The idea is that an explicit and simple formula for the MLE of $\pi$ would be available if the number $Z_0$ was known as well:

$$\hat{\pi} = \frac{\sum_{k=0}^{6} k \cdot Z_k}{6 \cdot \sum_{k=0}^{6} Z_k}. \tag{2.10}$$

Indeed, in this case we are back in the untruncated binomial case with $\sum_{k=0}^{6} k \cdot Z_k$ positive tests among $6 \cdot \sum_{k=0}^{6} Z_k$ tests. However, $Z_0$ is unknown, but if $\pi$ and hence $\xi = (1-\pi)^6$ are known, $Z_0$ can be estimated by the expectation of a negative binomial distribution (cf. Eq. (2.9) at the end of Example 2.12):

$$\hat{Z}_0 = \mathsf{E}(Z_0) = n \cdot \frac{\xi}{1-\xi}, \tag{2.11}$$

where $n = 196$ and $\xi = (1-\pi)^6$. The EM algorithm now iteratively computes (2.10) and (2.11) and replaces the terms $Z_0$ and $\pi$ on the right-hand sides with their current estimates $\hat{Z}_0$ and $\hat{\pi}$, respectively. Implementation of the algorithm is shown in the following R-code:

```
> ## data set
> fulldata <- c(NA, 37, 22, 25, 29, 34, 49)
> k <- 0:6
> n <- sum(fulldata[-1])
> ## impute start value for Z0 (first element)
> ## and initialise some different old value
> fulldata[1] <- 10
> Z0old <- 9
> ## the EM algorithm
> while(abs(Z0old - fulldata[1]) >= 1e-7)
  {
      Z0old <- fulldata[1]
      pi <- sum(fulldata * k) / sum(fulldata) / 6
      xi <- (1-pi)^6
      fulldata[1] <- n * xi / (1-xi)
  }
```

**Table 2.1** Values of the EM algorithm until convergence (Difference between old and new estimate $\hat{\pi}$ smaller than $10^{-7}$)

| Iteration | $\hat{\pi}$ | $\hat{Z}_0$ |
|---|---|---|
| 1 | 0.5954693 | 0.8627256 |
| 2 | 0.6231076 | 0.5633868 |
| 3 | 0.6240565 | 0.5549056 |
| 4 | 0.6240835 | 0.5546665 |
| 5 | 0.6240842 | 0.5546597 |
| 6 | 0.6240842 | 0.5546596 |

This method quickly converges, as illustrated in Table 2.1, with starting value $Z_0 = 10$. Note that the estimate $\hat{\pi}$ from Table 2.1 is numerically equal to the MLE (cf. Example 2.12) already after 5 iterations. As a by-product, we also obtain the estimate $\hat{Z}_0 = 0.55$. ■

A general derivation of the EM algorithm distinguishes *observed* data $X$ ($Z_1, \ldots, Z_6$ in Example 2.13) and *unobserved* data $U$ ($Z_0$ in Example 2.13). The joint probability mass or density function of the *complete* data $X$ and $U$ can always be written as (cf. Eq. (A.7) in Appendix A.2)

$$f(x, u) = f(u \mid x) f(x),$$

so the corresponding log-likelihood functions of $\theta$ obey the relationship

$$l(\theta; x, u) = l(\theta; u \mid x) + l(\theta; x). \tag{2.12}$$

Note that the log-likelihood functions cannot be written in the simple form $l(\theta)$ as they are based on different data: $l(\theta; x, u)$ is the complete data log-likelihood, while $l(\theta; x)$ is the observed data log-likelihood. Now $u$ is unobserved, so we replace it in (2.12) by the random variable $U$:

$$l(\theta; x, U) = l(\theta; U \mid x) + l(\theta; x)$$

and consider the expectation of this equation with respect to $f(u \mid x; \theta')$ (it will soon become clear why we distinguish $\theta$ and $\theta'$):

$$\underbrace{\mathsf{E}\{l(\theta; x, U); \theta'\}}_{=: Q(\theta;\theta')} = \underbrace{\mathsf{E}\{l(\theta; U \mid x); \theta'\}}_{=: C(\theta;\theta')} + \underbrace{l(\theta; x)}_{=l(\theta)}. \tag{2.13}$$

Note that the last term does not change, as it does not depend on $U$. So if

$$Q(\theta; \theta') \geq Q(\theta'; \theta'), \tag{2.14}$$

we have with (2.13):

$$l(\theta) - l(\theta') \geq C(\theta'; \theta') - C(\theta; \theta') \geq 0, \tag{2.15}$$

where the last inequality follows from the information inequality (cf. Appendix A.3.8).

This leads to the EM algorithm with starting value $\theta'$:

1. *Expectation (E-step):* Compute $Q(\theta;\theta')$.
2. *Maximisation (M-step):* Maximise $Q(\theta;\theta')$ with respect to $\theta$ to obtain $\theta''$.
3. Now iterate Steps 1 and 2 (i.e. set $\theta' = \theta''$ and apply Step 1), until the values of $\theta$ converge. A possible stopping criterion is $|\theta' - \theta''| < \varepsilon$ for some small $\varepsilon > 0$.

Equation (2.15) implies that every iteration of the EM algorithm increases the log-likelihood. This follows from the fact that—through maximisation—$Q(\theta'';\theta')$ is larger than $Q(\theta;\theta')$ for all $\theta$, so in particular (2.14) holds (with $\theta = \theta'$), and therefore $l(\theta'') \geq l(\theta')$. In contrast, the Newton–Raphson algorithm is not guaranteed to increase the log-likelihood in every iteration.

*Example 2.14* (Example 2.13 continued) The joint probability mass function of observed data $X = (Z_1, \ldots, Z_6)$ and unobserved data $U = Z_0$ is multinomially distributed (cf. Appendix A.5.3) with size parameter equal to $n + Z_0$ and probability vector $\boldsymbol{p}$ with entries

$$p_k = \binom{6}{k} \pi^k (1-\pi)^{6-k},$$

$k = 0, \ldots, 6$, which we denote by

$$(Z_0, Z_1, \ldots, Z_6)^\top \sim \mathrm{M}_7(n + Z_0, \boldsymbol{p}).$$

The corresponding log-likelihood

$$l(\pi) = \sum_{k=0}^{6} Z_k \log(p_k) \tag{2.16}$$

is linear in the unobserved data $Z_0$. Therefore, the E-step of the EM algorithm is achieved if we replace $Z_0$ with its expectation conditional on the observed data $X = (Z_1, \ldots, Z_6)$ and on the unknown parameter $\pi'$. From Example 2.12 we know that this expectation is $\mathsf{E}(Z_0 \mid Z_1 = z_1, \ldots, Z_6 = z_6; \pi') = n \cdot \xi'/(1-\xi')$ with $n = z_1 + \cdots + z_6$ and $\xi' = (1-\pi')^6$. This completes the E-step.

The M-step now maximises (2.16) with $Z_0 = n \cdot \xi'/(1-\xi')$ with respect to $\pi$. We have argued earlier that the complete data MLE is (2.10). This estimate can formally be derived from the complete data log-likelihood kernel

$$l(\pi) = \sum_{k=0}^{6} Z_k \big\{ (6-k)\log(1-\pi) + k \log(\pi) \big\},$$

easily obtained from (2.16), which leads to the complete data score function

$$S(\pi) = \left\{ \sum_{k=0}^{6} k \cdot Z_k/\pi - 6 \sum_{k=0}^{6} Z_k \right\} / (1-\pi).$$

It is easy to show that the root of this equation is (2.10). ■

One can also show that the EM algorithm always converges to a local or global maximum, or at least to a saddlepoint of the log-likelihood. However, the convergence can be quite slow; typically, more iterations are required than for the Newton–Raphson algorithm. Another disadvantage is that the algorithm does not automatically give the observed Fisher information. Of course, this can be calculated after convergence if the second derivative of the log-likelihood $l(\theta; x)$ of the observed data $x$ is available.

## 2.4 Quadratic Approximation of the Log-Likelihood Function

An important approximation of the log-likelihood function is based on a quadratic function. To do so, we apply a Taylor approximation of second order (cf. Appendix B.2.3) around the MLE $\hat{\theta}_{\mathrm{ML}}$:

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}_{\mathrm{ML}}) + \frac{dl(\hat{\theta}_{\mathrm{ML}})}{d\theta}(\theta - \hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\frac{d^2 l(\hat{\theta}_{\mathrm{ML}})}{d\theta^2}(\theta - \hat{\theta}_{\mathrm{ML}})^2 \\ &= l(\hat{\theta}_{\mathrm{ML}}) + S(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}}) - \frac{1}{2} \cdot I(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}})^2. \end{aligned}$$

Due to $S(\hat{\theta}_{\mathrm{ML}}) = 0$, the quadratic approximation of the relative log-likelihood is

$$\tilde{l}(\theta) = l(\theta) - l(\hat{\theta}_{\mathrm{ML}}) \approx -\frac{1}{2} \cdot I(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}})^2. \tag{2.17}$$

*Example 2.15* (Poisson model) Assume that we have one observation $x = 11$ from a Poisson distribution $\mathrm{Po}(e\lambda)$ with known offset $e = 3.04$ and unknown parameter $\lambda$. The MLE of $\lambda$ is $\hat{\lambda}_{\mathrm{ML}} = x/e = 3.62$, cf. Example 2.4. The observed Fisher information turns out to be $I(\hat{\lambda}_{\mathrm{ML}}) = x/\hat{\lambda}^2_{\mathrm{ML}}$, so that the quadratic approximation of the relative log-likelihood is

$$\tilde{l}(\lambda) \approx -\frac{1}{2}\frac{x}{\hat{\lambda}^2_{\mathrm{ML}}}(\lambda - \hat{\lambda}_{\mathrm{ML}})^2.$$

Figure 2.7 displays $\tilde{l}(\lambda)$ and its quadratic approximation. ■

*Example 2.16* (Normal model) Let $X_{1:n}$ denote a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown mean $\mu$ and known variance $\sigma^2$. We know from Example 2.9 that

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

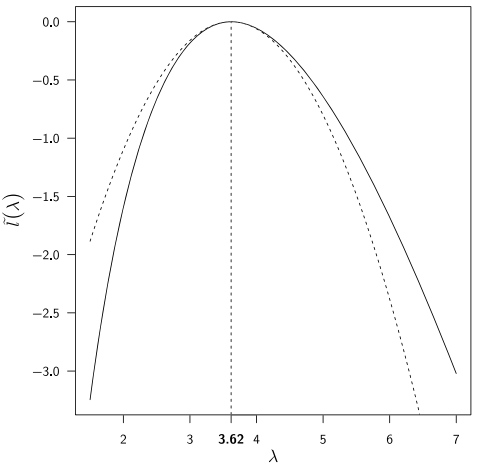


**Fig. 2.7** Relative log-likelihood $\tilde{l}(\lambda)$ and its quadratic approximation (*dashed line*) for a single observation $x = 11$ from a Poisson distribution with mean $e\lambda$ and known offset $e = 3.04$

$$l(\hat{\mu}_{\mathrm{ML}}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{and hence}$$

$$\tilde{l}(\mu) = l(\mu) - l(\hat{\mu}_{\mathrm{ML}}) = -\frac{n}{2\sigma^2}(\bar{x} - \mu)^2,$$

but we also have

$$-\frac{1}{2} \cdot I(\hat{\mu}_{\mathrm{ML}})(\mu - \hat{\mu}_{\mathrm{ML}})^2 = -\frac{n}{2\sigma^2}(\mu - \bar{x})^2.$$

Both sides of Eq. (2.17) are hence identical, so the quadratic approximation is here exact. ■

Under certain regularity conditions, which we will not discuss here, it can be shown that a quadratic approximation of the log-likelihood improves with increasing sample size. The following example illustrates this phenomenon in the binomial model.

*Example 2.17* (Binomial model) Figure 2.8 displays the relative log-likelihood of the success probability $\pi$ in a binomial model with sample size $n = 10, 50, 200, 1000$. The observed datum $x$ has been fixed at $x = 8, 40, 160, 800$ such that the MLE of $\pi$ is $\hat{\pi}_{\mathrm{ML}} = 0.8$ in all four cases. We see that the quadratic approximation of the relative log-likelihood improves with increasing sample size $n$. The two functions are nearly indistinguishable for $n = 1000$. ■

The advantage of the quadratic approximation of the relative log-likelihood lies in the fact that we only need to know the MLE $\hat{\theta}_{\mathrm{ML}}$ and the observed Fisher information $I(\hat{\theta}_{\mathrm{ML}})$, no matter what the actual log-likelihood looks like. However, in certain pathological cases the approximation may remain poor even if the sample size increases.

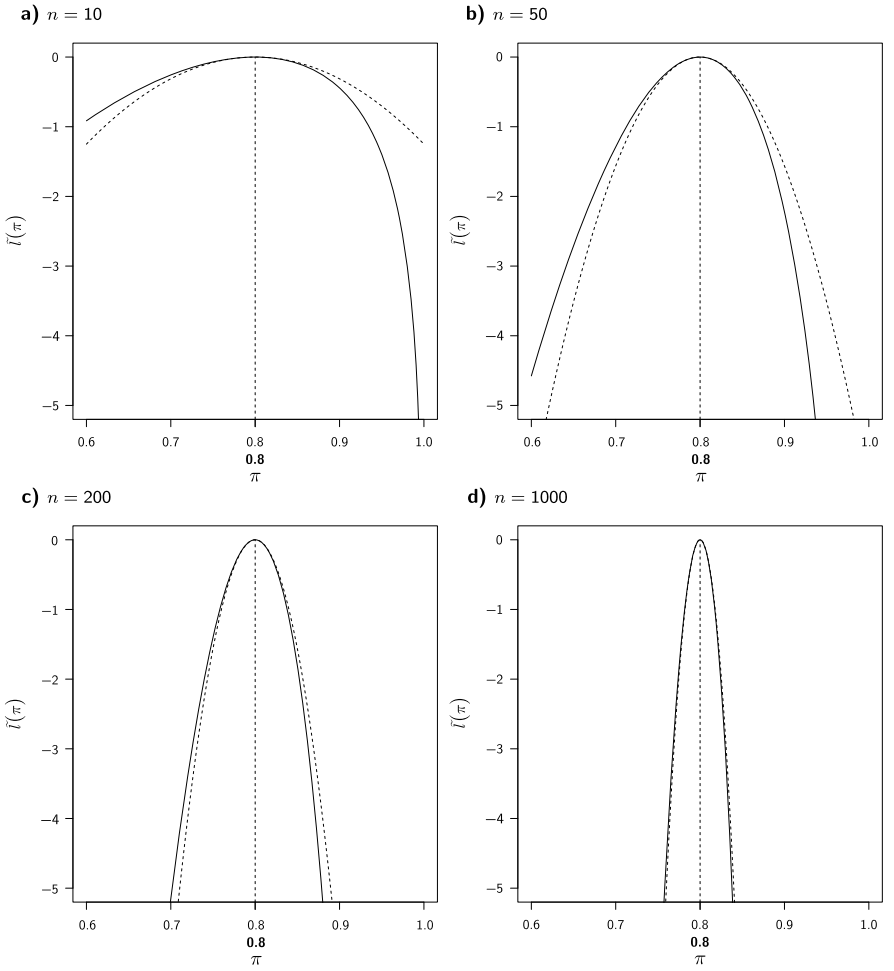


**Fig. 2.8** Quadratic approximation (*dashed line*) of the relative log-likelihood (*solid line*) of the success probability $\pi$ in a binomial model

*Example 2.18* (Uniform model) Let $X_{1:n}$ denote a random sample from a continuous uniform distribution $\mathrm{U}(0, \theta)$ with unknown upper limit $\theta \in \mathbb{R}^+$. The density function of the uniform distribution is

$$f(x; \theta) = \frac{1}{\theta} \mathsf{I}_{[0,\theta)}(x)$$

with *indicator function* $\mathsf{I}_A(x)$ equal to one if $x \in A$ and zero otherwise. The likelihood function of $\theta$ is

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} f(x_i; \theta) = \theta^{-n} & \text{for } \theta \geq \max_i(x_i), \\ 0 & \text{otherwise,} \end{cases}$$

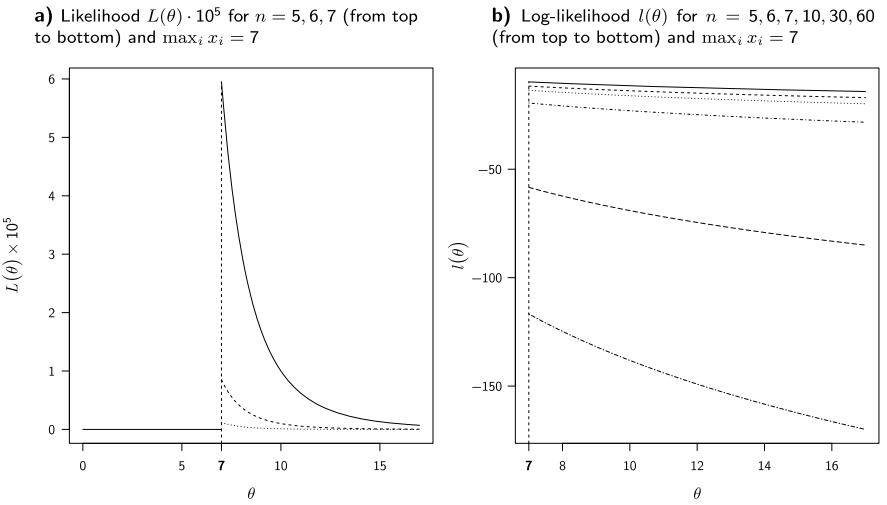


**Fig. 2.9** Likelihood and log-likelihood function for a random sample of different size $n$ from a uniform distribution with unknown upper limit $\theta$. Quadratic approximation of the log-likelihood is impossible even for large $n$

with MLE $\hat{\theta}_{\text{ML}} = \max_i(x_i)$, cf. Fig. 2.9a.

The derivatives of the log-likelihood function

$$l(\theta) = -n \log(\theta) \quad \text{for } \theta > \max_i(x_i)$$

are

$$S(\hat{\theta}_{\text{ML}}) = \frac{dl(\hat{\theta}_{\text{ML}})}{d\theta} \neq 0 \quad \text{and} \quad -I(\hat{\theta}_{\text{ML}}) = \frac{d^2 l(\hat{\theta}_{\text{ML}})}{d\theta^2} = \frac{n}{\hat{\theta}_{\text{ML}}^2} > 0,$$

so the log-likelihood $l(\theta)$ is not concave but convex, with negative (!) observed Fisher information, cf. Fig. 2.9b. It is obvious from Fig. 2.9b that a quadratic approximation to $l(\theta)$ will remain poor even if the sample size $n$ increases. The reason for the irregular behaviour of the likelihood function is that the support of the uniform distribution depends on the unknown parameter $\theta$. ■

## 2.5 Sufficiency

Under certain regularity conditions, a likelihood function can be well characterised by the MLE and the observed Fisher information. However, Example 2.18 illustrates that this is not always the case. An alternative characterisation of likelihood functions is in terms of *sufficient statistics*, a concept which we will introduce in the following. We will restrict our attention to random samples, but the description could be easily generalised if required.

**Definition 2.8** (Statistic) Let $x_{1:n}$ denote the realisation of a random sample $X_{1:n}$ from a distribution with probability mass of density function $f(x;\theta)$. Any function $T = h(X_{1:n})$ of $X_{1:n}$ with realisation $t = h(x_{1:n})$ is called a *statistic*. ♦

For example, the mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ is a statistic. Also the maximum $\max_i(X_i)$ and the range $\max_i(X_i) - \min_i(X_i)$ are statistics.

**Definition 2.9** (Sufficiency) A statistic $T = h(X_{1:n})$ is *sufficient* for $\theta$ if the conditional distribution of $X_{1:n}$ given $T = t$ is independent of $\theta$, i.e. if

$$f(x_{1:n} \mid T = t)$$

does not depend on $\theta$. ♦

*Example 2.19* (Poisson model) Let $x_{1:n}$ denote the realisation of a random sample $X_{1:n}$ from a Po($\lambda$) distribution with unknown rate parameter $\lambda$. The statistic $T = X_1 + \cdots + X_n$ is sufficient for $\lambda$ since the conditional distribution of $X_{1:n} \mid T = t$ is multinomial with parameters not depending on $\lambda$. Indeed, first note that $f(t \mid X_{1:n} = x_{1:n}) = 1$ if $t = x_1 + \cdots + x_n$ and 0 elsewhere. We also know from Appendix A.5.1 that $T \sim \text{Po}(n\lambda)$, and therefore we have

$$\begin{aligned} f(x_{1:n} \mid t) &= \frac{f(t \mid x_{1:n}) f(x_{1:n})}{f(t)} \\ &= \frac{f(x_{1:n})}{f(t)} \\ &= \frac{\prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right\}}{\frac{(n\lambda)^t}{t!} \exp(-n\lambda)} \\ &= \frac{t!}{\prod_{i=1}^{n} x_i!} \left( \frac{1}{n} \right)^t, \end{aligned}$$

which can easily be identified as the probability mass function of a multinomial distribution with size parameter $t = x_1 + \cdots + x_n$ and all probabilities equal to $1/n$, compare Appendix A.5.3. ■

A sufficient statistic $T$ contains all relevant information from the sample $X_{1:n}$ with respect to $\theta$. To show that a certain statistic is sufficient, the following result is helpful.

**Result 2.2** (Factorisation theorem) *Let $f(x_{1:n};\theta)$ denote the probability mass or density function of the random sample $X_{1:n}$. A statistic $T = h(X_{1:n})$ with realisation $t = h(x_{1:n})$ is sufficient for $\theta$ if and only if there exist functions $g_1(t;\theta)$ and $g_2(x_{1:n})$ such that for all possible realisations $x_{1:n}$ and all possible parameter values $\theta \in \Theta$,*

$$f(x_{1:n};\theta) = g_1(t;\theta) \cdot g_2(x_{1:n}). \tag{2.18}$$

*Note that $g_1(t;\theta)$ depends on the argument $x_{1:n}$ only through $t = h(x_{1:n})$, but also depends on $\theta$. The second term $g_2(x_{1:n})$ must not depend on $\theta$.*

A proof of this result can be found in Casella and Berger (2001, p. 276). As a function of $\theta$, we can easily identify $g_1(t;\theta)$ as the likelihood kernel, cf. Definition 2.3. The second term $g_2(x_{1:n})$ is the corresponding multiplicative constant.

*Example 2.20* (Poisson model) We already know from Example 2.19 that $T = h(X_{1:n}) = X_1 + \cdots + X_n$ is sufficient for $\lambda$, so the factorisation (2.18) must hold. This is indeed the case:

$$\begin{aligned} f(x_{1:n};\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right\} \\ &= \underbrace{\lambda^t \exp(-n\lambda)}_{g_1\{t;\lambda\}} \underbrace{\prod_{i=1}^{n} \frac{1}{x_i!}}_{g_2(x_{1:n})}. \end{aligned}$$

■

*Example 2.21* (Normal model) Let $x_{1:n}$ denote a realisation of a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with known variance $\sigma^2$. We now show that the *sample mean* $\bar{X} = \sum_{i=1}^{n} X_i/n$ is sufficient for $\mu$. First, note that

$$\begin{aligned} f(x_{1:n};\mu) &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \cdot \frac{(x_i-\mu)^2}{\sigma^2} \right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \cdot \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} \right\}. \end{aligned}$$

Now

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2,$$

and we can therefore factorise $f(x_{1:n};\mu)$ as follows:

$$f(x_{1:n};\mu) = \underbrace{\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \cdot \frac{\sum(x_i-\bar{x})^2}{\sigma^2} \right\}}_{g_2(x_{1:n})} \cdot \underbrace{\exp\left\{ -\frac{1}{2} \cdot \frac{n(\bar{x}-\mu)^2}{\sigma^2} \right\}}_{g_1(t;\mu) \text{ with } t=\bar{x}}.$$

Result 2.2 now ensures that the sample mean $\bar{X}$ is sufficient for $\mu$. Note that, for example, also $n\bar{X} = \sum_{i=1}^{n} X_i$ is sufficient for $\mu$.

Suppose now that also $\sigma^2$ is unknown, i.e. $\boldsymbol{\theta} = (\mu, \sigma^2)$, and assume that $n \geq 2$. It is easy to show that now $T = (\bar{X}, S^2)$, where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is the *sample variance*, is sufficient for $\boldsymbol{\theta}$. Another sufficient statistic for $\boldsymbol{\theta}$ is $\tilde{T} = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$. ■

*Example 2.22* (Blood alcohol concentration) If we are prepared to assume that the transformation factor is normally distributed, knowledge of $n$, $\bar{x}$ and $s^2$ (or $s$) in each group (cf. Table 1.3) is sufficient to formulate the likelihood function. It is not necessary to know the actual observations. ■

**Definition 2.10** (Likelihood ratio) The quantity

$$\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \frac{L(\theta_1; x_{1:n})}{L(\theta_2; x_{1:n})} = \frac{\tilde{L}(\theta_1; x_{1:n})}{\tilde{L}(\theta_2; x_{1:n})}$$

is the *likelihood ratio* of one parameter value $\theta_1$ relative to another parameter value $\theta_2$ with respect to the realisation $x_{1:n}$ of a random sample $X_{1:n}$. ♦

Note that likelihood ratios between any two parameter values $\theta_1$ and $\theta_2$ can be calculated from the relative likelihood function $\tilde{L}(\theta; x_{1:n})$. Note also that

$$\tilde{L}(\theta; x_{1:n}) = \Lambda_{x_{1:n}}(\theta, \hat{\theta}_{\mathrm{ML}})$$

because $\tilde{L}(\hat{\theta}_{\mathrm{ML}}; x_{1:n}) = 1$, so the relative likelihood function can be recovered from the likelihood ratio.

**Result 2.3** *A statistic $T = h(X_{1:n})$ is sufficient for $\theta$ if and only if for any pair $x_{1:n}$ and $\tilde{x}_{1:n}$ such that $h(x_{1:n}) = h(\tilde{x}_{1:n})$,*

$$\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \Lambda_{\tilde{x}_{1:n}}(\theta_1, \theta_2) \tag{2.19}$$

*for all $\theta_1, \theta_2 \in \Theta$.*

*Proof* We show the equivalence of the factorisation (2.18) and Eq. (2.19). Suppose that (2.18) holds. Then

$$\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \frac{g\{h(x_{1:n}); \theta_1\} \cdot h(x_{1:n})}{g\{h(x_{1:n}); \theta_2\} \cdot h(x_{1:n})} = \frac{g\{h(x_{1:n}); \theta_1\}}{g\{h(x_{1:n}); \theta_2\}},$$

so if $h(x_{1:n}) = h(\tilde{x}_{1:n})$, we have $\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \Lambda_{\tilde{x}_{1:n}}(\theta_1, \theta_2)$ for all $\theta_1$ and $\theta_2$.

Conversely, suppose that (2.19) holds if $h(x_{1:n}) = h(\tilde{x}_{1:n})$, so $\Lambda_{x_{1:n}}(\theta_1, \theta_2)$ is a function (say $g^*$) of $h(x_{1:n})$, $\theta_1$ and $\theta_2$ only. Let us now fix $\theta_2 = \theta_0$. With $\theta = \theta_1$ we obtain

$$\frac{f(x_{1:n};\theta)}{f(x_{1:n};\theta_0)} = \Lambda_{x_{1:n}}(\theta,\theta_0) = g^*\big\{h(x_{1:n}),\theta,\theta_0\big\},$$

so (2.18) holds:

$$f(x_{1:n};\theta) = \underbrace{g^*\big\{h(x_{1:n}),\theta,\theta_0\big\}}_{g_1\{h(x_{1:n});\theta\}} \underbrace{f(x_{1:n};\theta_0)}_{g_2(x_{1:n})}.$$

□

This result establishes an important relationship between a sufficient statistic $T$ and the likelihood function: If $T = h(X_{1:n})$ is a sufficient statistic and $h(x_{1:n}) = h(\tilde{x}_{1:n})$, then $x_{1:n}$ and $\tilde{x}_{1:n}$ define the same likelihood ratio.

The following result establishes another important property of the likelihood function. We distinguish in the following the likelihood $L(\theta; x_{1:n})$ with respect to a realisation $x_{1:n}$ of a random sample $X_{1:n}$ and the likelihood $L(\theta; t)$ with respect to the corresponding realisation $t$ of a sufficient statistic $T = h(X_{1:n})$ of the same random sample $X_{1:n}$.

**Result 2.4** *Let $L(\theta; x_{1:n})$ denote the likelihood function with respect to a realisation $x_{1:n}$ of a random sample $X_{1:n}$. Let $L(\theta; t)$ denote the likelihood with respect to the realisation $t = h(x_{1:n})$ of a sufficient statistic $T = h(X_{1:n})$ for $\theta$. For all possible realisations $x_{1:n}$, the ratio*

$$\frac{L(\theta; x_{1:n})}{L(\theta; t)}$$

*will then not depend on $\theta$, i.e. the two likelihood functions are (up to a proportionality constant) identical.*

*Proof* To show Result 2.4, first note that $f(t \mid x_{1:n}) = 1$ if $t = h(x_{1:n})$ and 0 otherwise, so $f(x_{1:n}, t) = f(x_{1:n}) f(t \mid x_{1:n}) = f(x_{1:n})$ if $t = h(x_{1:n})$. For $t = h(x_{1:n})$, the likelihood function can therefore be written as

$$\begin{aligned} L(\theta; x_{1:n}) &= f(x_{1:n};\theta) = f(x_{1:n}, t;\theta) = f(x_{1:n} \mid t;\theta) f(t;\theta) \\ &= f(x_{1:n} \mid t;\theta) L(\theta; t). \end{aligned}$$

Now $T$ is sufficient for $\theta$, so $f(x_{1:n} \mid t;\theta) = f(x_{1:n} \mid t)$ does not depend on $\theta$. Therefore, $L(\theta; x_{1:n}) \propto L(\theta; t)$. The sign "$\propto$", in words "proportional to", means that there is a constant $C > 0$ (not depending on $\theta$) such that $L(\theta; x_{1:n}) = C \cdot L(\theta; t)$. □

*Example 2.23* (Binomial model) Let $X_{1:n}$ denote a random sample from a Bernoulli distribution $\mathrm{B}(\pi)$ with unknown parameter $\pi \in (0, 1)$. The likelihood function based on the realisation $x_{1:n}$ equals

$$L(\pi; x_{1:n}) = f(x_{1:n}; \pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^t(1-\pi)^{n-t},$$

where $t = \sum_{i=1}^{n} x_i$. Obviously, $T = h(x_{1:n}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\pi$. Now $T$ follows the binomial distribution $\mathrm{Bin}(n, \pi)$, so the likelihood function with respect to its realisation $t$ is

$$L(\pi, t) = \binom{n}{t} \pi^t(1-\pi)^{n-t}.$$

As Result 2.4 states, the likelihood functions with respect to $x_{1:n}$ and $t$ are identical up to the multiplicative constant $\binom{n}{t}$. ■

Example 2.23 has shown that regarding the information about the proportion $\pi$, the whole random sample $X_{1:n}$ can be compressed into the total number of successes $T = \sum_{i=1}^{n} X_i$ without any loss of information. This will be important in Chap. 4, where we consider asymptotic properties of ML estimation, i.e. properties of certain statistics for sample size $n \to \infty$. Then we can consider a single binomial random variable $X \sim \mathrm{Bin}(n, \pi)$ because it implicitly contains the whole information of $n$ independent Bernoulli random variables with respect to $\pi$. We can also approximate the binomial distribution $\mathrm{Bin}(n, \pi)$ by a Poisson distribution $\mathrm{Po}(n\pi)$ when $\pi$ is small compared to $n$. Therefore, we can consider a single Poisson random variable $\mathrm{Po}(e\lambda)$ and assume that the asymptotic properties of derived statistics are a good approximation of their finite sample properties. We will often use this Poisson model parametrisation with expected number of cases $e = n \cdot p$ and relative risk $\lambda = \pi/p$, using a reference probability $p$, see e.g. Example 2.4.

### 2.5.1 Minimal Sufficiency

We have seen in the previous section that sufficient statistics are not unique. In particular, the original sample $X_{1:n}$ is always sufficient due to Result 2.2:

$$f(x_{1:n}; \theta) = \underbrace{f(x_{1:n}; \theta)}_{=g_1\{h(x_{1:n})=x_{1:n};\theta\}} \cdot \underbrace{1}_{=g_2(x_{1:n})}.$$

The concept of minimal sufficiency ensures that a sufficient statistic cannot be reduced further.

**Definition 2.11** (Minimal sufficiency) A sufficient statistic $T = h(X_{1:n})$ for $\theta$ is called *minimal sufficient*, if, for every possible realisation $x_{1:n}$ of $X_{1:n}$, $t = h(x_{1:n})$ can be written as a transformation of the realisation $\tilde{t}$ of any other sufficient statistic $\tilde{T} = \tilde{h}(X_{1:n})$. ♦

The following result describes the relationship between two minimal sufficient statistics.

**Result 2.5** *If $T$ and $\tilde{T}$ are minimal sufficient statistics, then there exists a one-to-one function $g$ such that $\tilde{T} = g(T)$ and $T = g^{-1}(\tilde{T})$.*

Loosely speaking, a minimal sufficient statistic is unique up to any one-to-one transformation. For example, if $T$ is minimal sufficient, then $T/2$ will also be minimal sufficient, but $|T|$ will not be minimal sufficient if $T$ can take values that differ only in sign.

**Result 2.6** *A necessary and sufficient criterion for a statistic $T(x_{1:n})$ to be minimal sufficient is that $h(x_{1:n}) = h(\tilde{x}_{1:n})$ if and only if*

$$\Lambda_{x_{1:n}}(\theta_1, \theta_2) = \Lambda_{\tilde{x}_{1:n}}(\theta_1, \theta_2)$$

*for all $\theta_1, \theta_2$.*

This is an extension of Result 2.3, where the sufficiency is characterised. For sufficiency, only the implication of equal likelihood ratios from equal statistics is needed. For minimal sufficiency, in addition, the implication of equal statistics from equal likelihood ratios is required. A proof of this result can be found in Young and Smith (2005, p. 92).

Result 2.6 means that any minimal sufficient statistic creates the same partition of the sample space as the likelihood ratio function. A more exact formulation is based on equivalence classes, where two different observations $x_{1:n}$ and $\tilde{x}_{1:n}$ are equivalent if they lead to the same likelihood ratio function. The likelihood ratio function partitions the sample space into the same equivalence classes as any minimal sufficient statistic. Therefore, the likelihood ratio is a one-to-one function of a minimal sufficient statistic and hence also minimal sufficient. Since we have described above that the likelihood ratio and the likelihood function can be recovered from each other, they are one-to-one transformations of each other. Hence, also the likelihood is minimal sufficient:

**Minimal sufficiency of the likelihood**
The likelihood function $L(\theta)$ is minimal sufficient for $\theta$.

This implies that the likelihood function contains the whole information of the data with respect to $\theta$. Any further reduction will result in information loss.

*Example 2.24* (Normal model) Let $x_{1:n}$ denote a realisation of a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with known variance $\sigma^2$. The mean $h(X_{1:n}) = \bar{X}$ is minimal sufficient for $\mu$, whereas $\tilde{T}(X_{1:n}) = (\bar{X}, S^2)$ is sufficient but not minimal sufficient for $\mu$. ■

### 2.5.2 The Likelihood Principle

Are there general principles how to infer information from data? In the previous section we have seen that sufficient statistics contain the complete information of a sample with respect to an unknown parameter. It is thus natural to state the *sufficiency principle*:

**Sufficiency principle**
Consider a random sample $X_{1:n}$ from a distribution with probability mass or density function $f(x;\theta)$ and unknown parameter $\theta$. Assume that $T = h(X_{1:n})$ is a sufficient statistic for $\theta$. If $h(x_{1:n}) = h(\tilde{x}_{1:n})$ for two realisations of $X_{1:n}$, then inference for $\theta$ should be the same whether $x_{1:n}$ or $\tilde{x}_{1:n}$ has been observed.

The likelihood function is also sufficient, so we immediately have the *likelihood principle*:

**Likelihood principle**
If realisations $x_{1:n}$ and $\tilde{x}_{1:n}$ from a random sample $X_{1:n}$ with probability mass or density function $f(x;\theta)$ have proportional likelihood functions, i.e. $L(\theta; x_{1:n}) \propto L(\theta; \tilde{x}_{1:n})$ for all $\theta$, then inference for $\theta$ should be the same, whether $x_{1:n}$ or $\tilde{x}_{1:n}$ is observed.

This principle is also called the *weak likelihood principle* to distinguish it from the *strong likelihood principle*:

**Strong likelihood principle**
Suppose $x_{1:n_1}$ is a realisation from a random sample $X_{1:n_1}$ with probability mass or density function $f_1(x;\theta)$. Let $\tilde{x}_{1:n_2}$ denote a realisation from a random sample $\tilde{X}_{1:n_2}$ with probability mass or density function $f_2(x;\theta)$, not necessarily identical to $f_1(x;\theta)$. If the corresponding two likelihood functions are proportional, i.e. $L_1(\theta; x_{1:n_1}) \propto L_2(\theta; \tilde{x}_{1:n_2})$, then inference for $\theta$ should be the same, whether $x_{1:n_1}$ from $f_1(x;\theta)$ or $\tilde{x}_{1:n_2}$ from $f_2(x;\theta)$ has been observed.

To illustrate the strong likelihood principle, we consider a classical example.

*Example 2.25* (Binomial and negative binomial model) A binomial model $X \sim \text{Bin}(m, \pi)$ is appropriate if a fixed number $m$ of independent and identical Bernoulli experiments is conducted. The random variable $X$ denotes the number of successes

of the event considered with success probability $\pi$. The likelihood function of $n_1 = 1$ realisation $x$ is then

$$L_1(\pi) = \binom{m}{x} \pi^x (1-\pi)^{m-x}.$$

As discussed in Sect. 1.1.1, alternatively we can imagine a design where we conduct independent Bernoulli experiments until we have a total of $x$ successes. In this inverse binomial model $x$ is fixed, but now the number of required samples $m$ is a realisation of a random variable $M$, say. Of interest is again the parameter $\pi$, the unknown probability of the event of interest, with likelihood function

$$L_2(\pi) = \binom{m-1}{x-1} \pi^x (1-\pi)^{m-x},$$

derived from the realisation $m$ of a random sample of size $n_2 = 1$ from the negative binomial distribution $M \sim \mathrm{NBin}(x, \pi)$. The likelihood functions $L_1(\theta)$ and $L_2(\theta)$ are (up to different multiplicative constants) identical if $m$ and $x$ are the same. The strong likelihood principle requires that statistical inference for $\theta$ must be the same, whether or not the data have arisen from the binomial or the negative binomial model. ■

## 2.6 Exercises

1. Examine the likelihood function in the following examples.
   (a) In a study of a fungus that infects wheat, 250 wheat seeds are disseminated after contaminating them with the fungus. The research question is how large the probability $\theta$ is that an infected seed can germinate. Due to technical problems, the exact number of germinated seeds cannot be evaluated, but we know only that less than 25 seeds have germinated. Write down the likelihood function for $\theta$ based on the information available from the experiment.
   (b) Let $X_{1:n}$ be a random sample from an $\mathrm{N}(\theta, 1)$ distribution. However, only the largest value of the sample, $Y = \max(X_1, \ldots, X_n)$, is known. Show that the density of $Y$ is

$$f(y) = n\big\{\Phi(y-\theta)\big\}^{n-1} \varphi(y-\theta), \quad y \in \mathbb{R},$$

   where $\Phi(\cdot)$ is the distribution function, and $\varphi(\cdot)$ is the density function of the standard normal distribution $\mathrm{N}(0, 1)$. Derive the distribution function of $Y$ and the likelihood function $L(\theta)$.
   (c) Let $X_{1:3}$ denote a random sample of size $n = 3$ from a Cauchy $\mathrm{C}(\theta, 1)$ distribution, cf. Appendix A.5.2. Here $\theta \in \mathbb{R}$ denotes the location parameter of the Cauchy distribution with density

$$f(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}, \quad x \in \mathbb{R}.$$

Derive the likelihood function for $\theta$.

(d) Using R, produce a plot of the likelihood functions:
   i. $L(\theta)$ in 1(a).
   ii. $L(\theta)$ in 1(b) if the observed sample is $x = (1.5, 0.25, 3.75, 3.0, 2.5)$.
   iii. $L(\theta)$ in 1(c) if the observed sample is $x = (0, 5, 9)$.

2. A first-order autoregressive process $X_0, X_1, \ldots, X_n$ is specified by the conditional distribution

$$X_i \mid X_{i-1} = x_{i-1}, \ldots, X_0 = x_0 \sim \mathrm{N}(\alpha \cdot x_{i-1}, 1), \quad i = 1, 2, \ldots, n,$$

and some initial distribution for $X_0$. This is a popular model for time series data.

(a) Consider the observation $X_0 = x_0$ as fixed. Show that the log-likelihood kernel for a realisation $x_1, \ldots, x_n$ can be written as

$$l(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2.$$

(b) Derive the score equation for $\alpha$, compute $\hat{\alpha}_{\mathrm{ML}}$ and verify that it is really the maximum of $l(\alpha)$.

(c) Create a plot of $l(\alpha)$ and compute $\hat{\alpha}_{\mathrm{ML}}$ for the following sample:

$$(x_0, \ldots, x_6) = (-0.560, -0.510, 1.304, 0.722, 0.490, 1.960, 1.441).$$

3. Show that in Example 2.2 the likelihood function $L(N)$ is maximised at $\hat{N} = \lfloor M \cdot n/x \rfloor$, where $\lfloor x \rfloor$ is the largest integer that is smaller than $x$. To this end, analyse the monotonic behaviour of the ratio $L(N)/L(N-1)$. In which cases is the MLE not unique? Give a numeric example.

4. Derive the MLE of $\pi$ for an observation $x$ from a geometric Geom($\pi$) distribution. What is the MLE of $\pi$ based on a realisation $x_{1:n}$ of a random sample from this distribution?

5. A sample of 197 animals has been analysed regarding a specific phenotype. The number of animals with phenotypes AB, Ab, aB and ab, respectively, turned out to be

$$\boldsymbol{x} = (x_1, x_2, x_3, x_4)^\top = (125, 18, 20, 34)^\top.$$

A genetic model now assumes that the counts are realisations of a multinomially distributed multivariate random variable $\boldsymbol{X} \sim \mathrm{M}_4(n, \boldsymbol{\pi})$ with $n = 197$ and probabilities $\pi_1 = (2+\phi)/4$, $\pi_2 = \pi_3 = (1-\phi)/4$ and $\pi_4 = \phi/4$ (Rao 1973, p. 368).

(a) What is the parameter space of $\phi$? See Table A.3 in Appendix A for details on the multinomial distribution and the parameter space of $\boldsymbol{\pi}$.

(b) Show that the likelihood kernel function for $\phi$, based on the observation $\boldsymbol{x}$, has the form

$$L(\phi) = (2+\phi)^{m_1}(1-\phi)^{m_2}\phi^{m_3}$$

and derive expressions for $m_1$, $m_2$ and $m_3$ depending on $\boldsymbol{x}$.

(c) Derive an explicit formula for the MLE $\hat{\phi}_{\mathrm{ML}}$, depending on $m_1$, $m_2$ and $m_3$. Compute the MLE given the data given above.

(d) What is the MLE of $\theta = \sqrt{\phi}$?

6. Show that $h(X) = \max_i(X_i)$ is sufficient for $\theta$ in Example 2.18.

7. (a) Let $X_{1:n}$ be a random sample from a distribution with density

$$f(x_i;\theta) = \begin{cases} \exp(i\theta - x_i), & x_i \geq i\theta, \\ 0, & x_i < i\theta, \end{cases}$$

for $X_i$, $i = 1, \ldots, n$. Show that $T = \min_i(X_i/i)$ is a sufficient statistic for $\theta$.

(b) Let $X_{1:n}$ denote a random sample from a distribution with density

$$f(x;\theta) = \exp\{-(x-\theta)\}, \quad \theta < x < \infty, \ -\infty < \theta < \infty.$$

Derive a minimal sufficient statistic for $\theta$.

8. Let $T = h(X_{1:n})$ be a sufficient statistic for $\theta$, $g(\cdot)$ a one-to-one function, and $\tilde{T} = \tilde{h}(X_{1:n}) = g\{h(X_{1:n})\}$. Show that $\tilde{T}$ is sufficient for $\theta$.

9. Let $X_1$ and $X_2$ denote two independent exponentially Exp($\lambda$) distributed random variables with parameter $\lambda > 0$. Show that $h(X_1, X_2) = X_1 + X_2$ is sufficient for $\lambda$.

## 2.7 Bibliographic Notes

A good introduction to likelihood methods is given by Pawitan (2001). More ambitious and rigorous is the presentation in Davison (2003). The pure likelihood approach is described in Edwards (1992) and Royall (1997).

http://www.springer.com/978-3-642-37886-7

Applied Statistical Inference
Likelihood and Bayes
Held, L.; Sabanés Bové, D.
2014, XIII, 376 p. 71 illus., Softcover
ISBN: 978-3-642-37886-7