Chapter 4
Maximum Likelihood Inference

In the last chapter, we explored Maximum Likelihood as a common approach for estimating model parameters given a sample from the population. In this chapter, we examine in more detail this estimate as well as the likelihood and its related functions: the log-likelihood, score function and Fisher information. We also derive properties of their asymptotic distributions, which tells us how much the functions themselves vary between samples. This is useful for deriving confidence intervals and statistical tests. We begin this chapter with a peek into the controversial history of the likelihood function. Not only does this give an insight into the often fraught process of discovery, but also makes clear the sometimes confusing difference between likelihood and probability.

Maximum Likelihood estimation was first proposed by Fisher (1922) and, although likelihood reasoning is absolutely essential to modern-day statistics, it was met with considerable resistance upon publication. In fact, as a graduate student (!), Fisher proposed the central ideas of likelihood reasoning in 1912 (see Fisher 1912), describing it as a method of maximum probability. Following the parametric distribution models of the previous chapter, he assumed that the parameter θ is unknown but follows a distribution. Therefore, the uncertainty about θ can also be represented by a distribution, whose maximum can then be found. Although this resembles the Bayesian approach, Fisher did not specify a prior distribution for the parameter, but instead suggested using only the likelihood function. That is, he considered the distribution of the data y but treated parameter θ as argument of the function, which needs to be set in relation to the data. This concept caused confusion and outrage, in part because Fisher referred to the outcome as a probability, which it is not. The discussion was settled 10 years later when Fisher finally defined the likelihood, not as a probability distribution, but as the "likelihood" function (see Fisher 1922). The name was not the only important factor. The entire concept he proposed in 1912 also had no solid theoretical justification. This was developed by Fisher over the next decade, leading to his seminal work in 1922. The major development was that, while the likelihood function itself may have no deep

meaning, its maximum most certainly does. Fisher investigated the likelihood at its maximum and showed that the resulting estimate has useful properties. We refer the curious reader to Stigler (2007), Aldrich (1997) and Edwards (1974) for a comprehensive history.

4.1 Score Function and Fisher Information

In the last chapter, we introduced the likelihood function as a representation of how much we believe that a given sample $y_1, \ldots, y_n$ was generated by a model parameterised by θ. As before, we are now interested in how this function varies with the random sampling process. Therefore, instead of assuming a concrete sample, we treat our sample as a series of n identically distributed random variables

$$Y_i \sim f(y;\theta) \quad i.i.d.$$

Here, the distribution $f(.;\theta)$ is assumed to be Fisher-regular as given in Definition 3.12. To remind the reader, Fisher-regularity primarily means that the set of possible values of Y does not depend on θ, and that the order of integration with respect to y and differentiation with respect to θ does not matter. The parameter θ can be multidimensional with $\theta \in \mathbb{R}^P$, although for notational simplicity we continue to assume that $p = 1$. Following Definition 3.2 we define the likelihood function as

$$L\big(\theta; y_1, \ldots, y_n\big) = \prod_{i=1}^{n} f\big(y_i;\theta\big).$$

The log-likelihood is given by

$$l\big(\theta; y_1, \ldots, y_n\big) = \log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f\big(y_i;\theta\big).$$

Definition 4.1 The first derivative of the log-likelihood function is called the **score function**

$$s\big(\theta; y_1, \ldots, y_n\big) = \frac{\partial l(\theta, y_1, \ldots, y_n)}{\partial \theta}. \tag{4.1.1}$$

Differentiating the score function and taking negative expectation gives the **Fisher information** already defined in (3.3.4), that is

$$I(\theta) = -E\left(\frac{\partial s(\theta; Y_1, \ldots, Y_n)}{\partial \theta}\right) = -E\left(\frac{\partial^2 l(\theta; Y_1, \ldots, Y_n)}{\partial \theta \partial \theta}\right).$$

The score function is the slope of the likelihood function and our intention to maximise the likelihood function corresponds to finding the root of the score function. That is, we find $s(\hat{\theta}) = 0$, assuming that the likelihood function is differentiable. Moreover, the Fisher information is the expected second order derivative of the likelihood function and it quantifies the curvature of the likelihood. In particular, at the maximum of the likelihood function, where the slope and thus the score function are zero, the second order derivative expresses the "peakiness" of the maximum. We will see that this in turn gives the amount of information in the data about the parameter.

We will now investigate the score function. Let us stick with a single random variable $Y \sim f(y;\theta)$ for notational simplicity, but later show the same results for a random sample. The reader should note that the outputs of the score function and Fisher information are now random as well. Let us now show that the score function has mean zero, for all possible true parameter values. Note that for each parameter value θ, we obtain that $f(\cdot;\theta)$ is a density function (or a probability function), such that the integral over all possible values of a random variable given the parameter θ is one, i.e.

$$1 = \int f(y;\theta)dy. \tag{4.1.2}$$

Differentiating both sides of Eq. (4.1.2) with respect to θ and making use of the fact that integration and differentiation are exchangeable gives

$$\begin{aligned} 0 = \frac{\partial 1}{\partial\theta} &= \int \frac{\partial f(y;\theta)}{\partial\theta}dy = \int \frac{\partial f(y;\theta)}{\partial\theta}\frac{f(y;\theta)}{f(y;\theta)}dy \\ &= \int \frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy = \int s(\theta;y)f(y;\theta)dy, \end{aligned}$$

where $s(\theta;y) = \frac{\partial}{\partial\theta}\log f(y;\theta)$. This shows that the score function has mean zero, i.e.

$$E\Big(s\big(\theta;Y\big)\Big) = 0. \tag{4.1.3}$$

Property (4.1.3) is called the first **Bartlett identity**. Equation 4.1.3 shows now that, although its position may vary, on average there is a peak in the likelihood function at θ. Let us further differentiate both sides of Eq. (4.1.3) with respect to θ, which gives

$$\begin{aligned} 0 = \frac{\partial 0}{\partial\theta} &= \frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta}\log f(y;\theta) f(y;\theta)dy \\ &= \int \left(\frac{\partial^2}{\partial\theta\partial\theta}\log f(y;\theta)\right) f(y;\theta)dy + \int \frac{\partial \log f(y;\theta)}{\partial\theta}\,\frac{\partial f(y;\theta)}{\partial\theta}\,dy \end{aligned}$$

$$= E\left(\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y;\theta)\right) + \int \frac{\partial \log f(y;\theta)}{\partial\theta}\,\frac{\partial \log f(y;\theta)}{\partial\theta}\, f(y;\theta)dy$$

$$\Leftrightarrow E\Big(s(\theta;Y)s(\theta;Y)\Big) = E\left(-\frac{\partial^2}{\partial\theta\partial\theta}\log f(Y;\theta)\right).$$

Because $E\big(s(\theta;Y)\big) = 0$, we obtain the **second order Bartlett identity**

$$Var\,(s(Y;\theta)) = E\left(-\frac{\partial^2 \log f(Y;\theta)}{\partial\theta\partial\theta}\right). \tag{4.1.4}$$

Hence, the variance of the score function is equal to the Fisher information. The reader should also note that if $\theta \in \mathbb{R}^p$, then the score function is vector valued and second order differentiation leads to a matrix. In this case $I(\theta)$ is called the Fisher information matrix or simply the Fisher matrix. It is defined as

$$I(\theta) = -E\left(\frac{\partial^2 l(\theta;Y)}{\partial\theta\partial\theta^T}\right).$$

Given the importance of the above properties, we formulate identical results for an *i.i.d.* sample $Y_1,\ldots,Y_n$.

Property 4.1 Given the *i.i.d.* sample $Y_1,\ldots,Y_n$ from a Fisher-regular distribution, the score function

$$s\,(\theta;Y_1,\ldots,Y_n) = \sum_{i=1}^{n}\frac{\partial \log f(Y_i;\theta)}{\partial\theta}$$

has zero mean; i.e. $E\Big(s(\theta;Y_1,\ldots,Y_n)\Big) = 0$, and its variance is given by the Fisher information (3.3.4), that is

$$I(\theta) = E\left(-\frac{\partial s(\theta;Y_1,\ldots,Y_n)}{\partial\theta}\right) = \mathrm{Var}\big(s(\theta;Y_1,\ldots,Y_n)\big).$$

Example 6 Here we show the log-likelihood, score function and Fisher information for the normal distribution. Let $Y_i \sim N(\mu,\sigma^2)$ *i.i.d.* for $i = 1,\ldots,n$ where for simplicity σ^2 is assumed to be known. The log-likelihood function is given by (up to constants not depending on μ)

$$l(\mu;y_1,\ldots,y_n) = -\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

Taking the derivative with respect to μ gives the score function

$$s(\mu; y_1, \ldots, y_n) = \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma^2} = \frac{\sum_{i=1}^{n} y_i}{\sigma^2} - \frac{n}{\sigma^2}\mu,$$

which clearly gives $E(s(\mu; Y_1, \ldots, Y_n)) = 0$, as $E(Y_i) = \mu$. The Fisher information is

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

and with a bit of calculation we can see that $Var\big(s(\mu; Y_1, \ldots, Y_n)\big) = \frac{n}{\sigma^2} = I(\mu)$.

▷

The Maximum Likelihood estimate is obtained by maximising the likelihood function. Because the score function has a non-negative variance, it follows from (4.1.4) that the second order derivative is negative in expectation. This in turn guarantees a well defined optimisation problem as the likelihood function is concave and hence the maximum is unique. This property holds in expectation only and for a concrete sample we might experience non-unique maxima, i.e. local maxima. Nonetheless, we see that the Fisher information plays a fundamental role as it mirrors whether the maximisation problem is well defined or not. With the above definitions, we also define the Maximum Likelihood estimate with reference to the score function:

Definition 4.2 For a random sample $Y_1, \ldots Y_n$, the **Maximum Likelihood** (ML) estimate is defined by

$$\hat{\theta}_{ML} = \arg\max l(\theta; Y_1, \ldots, Y_n),$$

which for Fisher-regular distributions occurs when

$$s\left(\hat{\theta}_{ML}; y_1, \ldots, y_n\right) = 0.$$

For simplicity of notation and if there is no ambiguity, we sometimes drop the index ML and just write $\hat{\theta}$ for the ML estimate. Let us look at the case of the binary variable $Y_i \sim B(1, \pi)$ *i.i.d.*, $i = 1, \ldots, n$ with unknown parameter $\pi = P(Y_i = 1)$. As the number of successes $Y = \sum_{i=1}^{n} Y_i$ is a sufficient statistic, we look at Y, which has a Binomial distribution with $Y \sim B(n, \pi)$. Assume we have observed $y = 10$ with $n = 30$. The log-likelihood function is plotted in Fig. 4.1. The maximum is at $\bar{y} = \frac{10}{30}$, indicated as dashed vertical line. Now assume that we have generated the data using $\pi = 0.4$, shown as solid vertical line. Then $y = 10$ occurs with probability $P(y = 10) = \binom{30}{10} 0.4^{10} 0.6^{20} = 0.115$. In other words, if we were to draw Y again, we would get a different value of Y with a rather large probability. Consequently, the resulting log-likelihood function would be different.

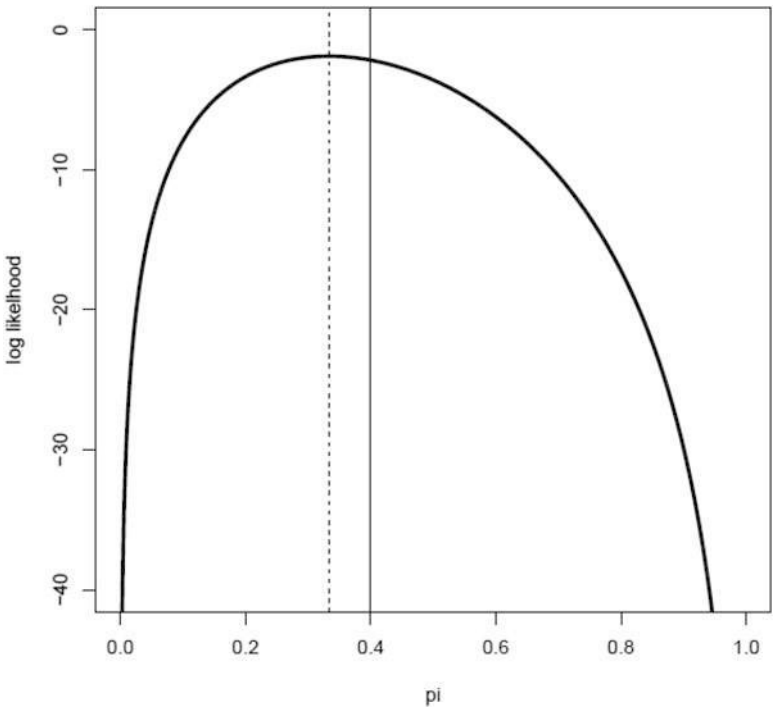

Fig. 4.1 Log-likelihood function for a Binomial model with $n = 30$ and $y = 10$

This is to say that the log-likelihood function itself must be considered as random variable. In Fig. 4.2, we plot the resulting random log-likelihood functions $l(\pi, Y)$, where the width of the lines is proportional to the probability that this log-likelihood function results. The functions are normed to take maximum value 0. The true parameter value $\pi = 0.4$ is shown with a vertical line. We see that even though the log-likelihood functions are random, their maxima are centred around the true parameter. We may also visualise the effect of increasing sample sizes. In Fig. 4.3 we plot the log-likelihood function for the binomial model for different values of n, assuming always that the arithmetic mean $\bar{y}$ is 0.4. Apparently, the larger the sample size, the more exposed is the maximum.

The next step is to quantify the variation around the true value, meaning that we aim to assess the variability of the maximum of the log-likelihood function. Note that the maximum is the root of the score function $s(\pi; Y)$ and we have already derived the variance of the score. We will make use of this result shortly.

Example 7 To demonstrate a two dimensional likelihood function, let us look once again at the normal distribution and assume that $Y_i \sim N(\mu, \sigma^2)$ $i.i.d.$, $i = 1, \ldots, n$, where we set $\mu = 0$ and $\sigma^2 = 1$. For a given sample of size $n = 10$ we obtain an arithmetic mean of $\bar{y} = -0.41$ and a standard deviation of 0.86. We plot the resulting likelihood function for both the mean value μ and the standard

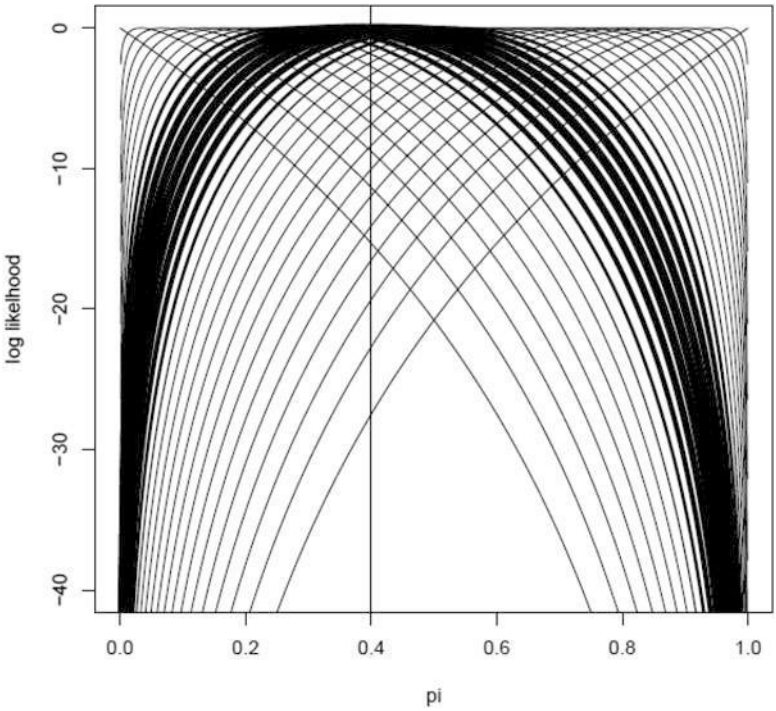

Fig. 4.2 Randomness of log-likelihood functions. Thickness is proportional to the probability that this log-likelihood function occurs

deviation σ in Fig. 4.4. This demonstrates the use of the likelihood in the case of multidimensional parameters. A horizontal line indicates the true value of μ and a vertical line that of σ. We should again bear in mind that we consider the likelihood function as random if it takes a random sample $Y_1, \ldots, Y_n$.

▷

4.2 Asymptotic Normality

Now that we understand how the score function and Fisher information vary with a sample $Y_1, \ldots, Y_n$, we can turn our attention to the Maximum Likelihood estimate itself. The ML estimate has asymptotic properties, which we will derive in the following section. These properties will be very helpful in predicting how much our estimate varies and for providing confidence intervals for the parameters.

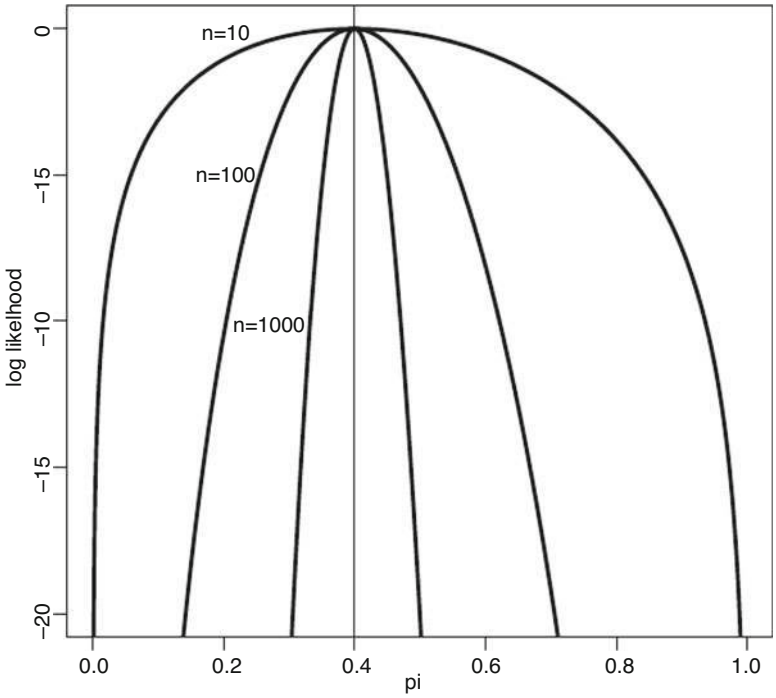

Fig. 4.3 Log-likelihood function for different sample sizes

Property 4.2 Assuming a Fisher-regular distribution with parameter θ for an $i.i.d.$ sample $Y_1, \ldots, Y_n$, the ML estimate is asymptotically normally distributed with

$$\hat{\theta}_{ML} \overset{a}{\sim} N\left(\theta, I^{-1}(\theta)\right). \tag{4.2.1}$$

In particular, this means that the ML estimate has asymptotically the smallest possible variance: the Cramer-Rao bound given in Eq. (3.3.5). Note that we also obtain that $\hat{\theta}$.

The proof of this statement is a bit lengthy and the mathematically less experienced reader may omit it. We emphasise, however, that (4.2.1) is *the* central result in Maximum Likelihood theory and proves very useful in inferring the properties of θ.

Proof In order to motivate and prove asymptotic properties of the Maximum Likelihood estimate $\hat{\theta}_{ML}$, we need to modify the notation slightly. Firstly, we take the score function $s\left(\theta; Y_1, \ldots, Y_n\right)$ as a random quantity. Secondly, we decompose the score function into the sum of the scores of each individual sample

$$s(\theta; Y_1, \ldots, Y_n) = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f(Y_i; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i; \theta) =: \sum_{i=1}^{n} s_i(\theta).$$

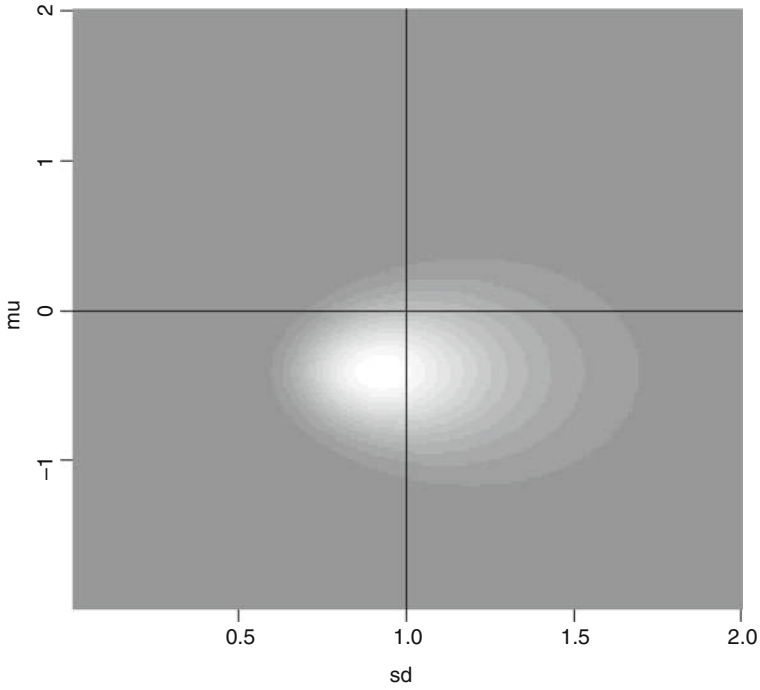

Fig. 4.4 Likelihood function for a sample of standard normally distributed variables. Lighter indicates an increased log-likelihood

Finally, to emphasise the role of the sample size, we include n as subscript and denote the score function as

$$s_{(n)}(\theta) = \sum_{i=1}^{n} s_i(\theta).$$

Note that $E\big(s_i(\theta)\big) = 0$. The Fisher information is also labelled with subscript n and similarly the sum of the Fisher Information of each individual sample

$$I_{(n)}(\theta) = \mathrm{Var}\left(s_{(n)}(\theta)\right) = \sum_{i=1}^{n} \mathrm{Var}\big(s_i(\theta)\big) =: \sum_{i=1}^{n} I_i(\theta).$$

The index n for the sample size is also used for the log-likelihood, i.e. we explicitly write the log-likelihood as

$$l_{(n)}(\theta) = \sum_{i=1}^{n} l_i(\theta) = \sum_{i=1}^{n} \log f(Y_i; \theta).$$

As both $s_i(\theta)$ and Y_i are $i.i.d.$, we may directly apply the central limit theorem to show that

$$s_{(n)}(\theta) \overset{a}{\sim} N\left(0, I_{(n)}(\theta)\right). \tag{4.2.2}$$

This follows easily with the two Bartlett Identities derived in Property 4.1. Finally, we also make explicit the influence of the sample size n on the Maximum Likelihood estimate and denote the ML estimate as $\hat{\theta}_{(n)}$. Property (4.2.2) is central to Maximum Likelihood estimation and it allows us to generally derive an asymptotic normality for the ML estimate. Note that the ML estimate is the root of the score function. Using Taylor series expansion around the true parameter value θ_0 leads to

$$0 = s_{(n)}(\hat{\theta}_{(n)}) = s_{(n)}(\theta_0) + \frac{\partial s_{(n)}(\theta_0)}{\partial \theta}\left(\hat{\theta}_{(n)} - \theta_0\right) + \frac{1}{2}\frac{\partial^2 s_{(n)}(\tilde{\theta})}{(\partial\theta)^2}\left(\hat{\theta}_{(n)} - \theta_0\right)^2, \tag{4.2.3}$$

where $\tilde{\theta}$ lies between θ_0 and $\hat{\theta}_{(n)}$, using the mean value theorem for differentiation. In fact, the above equation holds approximately if we set $\tilde{\theta}$ to θ_0 in the second derivative. To simplify notation we drop the parameter when we evaluate the functions at the true parameter value θ_0, i.e.

$$s_{(n)} = s_{(n)}(\theta_0),\ s'_{(n)} = \frac{\partial s_{(n)}(\theta_0)}{\partial \theta} \text{ and } s''_{(n)} = \frac{\partial^2 s_{(n)}(\theta_0)}{(\partial\theta)^2}.$$

Equation (4.2.3) reads now in approximate form as

$$0 \approx s_{(n)} + s'_{(n)}(\hat{\theta}_{(n)} - \theta_0) + \frac{1}{2}s''_{(n)}(\hat{\theta}_{(n)} - \theta_0)^2. \tag{4.2.4}$$

Because (4.2.4) is an approximate quadratic equation, we can solve it for $\hat{\theta}_{(n)} - \theta_0$ using standard calculus to get

$$\left(\hat{\theta}_{(n)} - \theta_0\right) \approx -\frac{1}{s''_{(n)}}\left(s'_{(n)} \pm \sqrt{s'^2_{(n)} - 2s_{(n)}s''_{(n)}}\right). \tag{4.2.5}$$

We will now look at the components on the right-hand side of Eq. (4.2.5) and try to simplify them with respect to their asymptotic behaviour, i.e. we investigate what happens if n tends to infinity. Taylor expansion of the square root component in (4.2.5) leads to

$$\left(s'^2_{(n)} - 2s_{(n)}s''_{(n)}\right)^{1/2} = s'_{(n)} - \frac{s_{(n)}s''_{(n)}}{s'_{(n)}} + \ldots, \tag{4.2.6}$$

where we can ignore the remaining components collected in ... as they are of negligible asymptotic order for increasing n. Taking (4.2.6) and inserting it in (4.2.5) yields

$$\hat{\theta}_{(n)} - \theta_0 = -\frac{s_{(n)}}{s'_{(n)}} \tag{4.2.7}$$

and a second solution which is not of any further interest to us here. Looking at (4.2.7), we can simplify this further by decomposing the expectation of $s'_{(n)}$ into its mean value and the remainder, where its mean is the negative Fisher information. Hence

$$s'_{(n)} = -I_{(n)} + u_{(n)},$$

where $u_{(n)} = s'_{(n)} + I_{(n)}$ is a sum of independent zero mean variables. With the results discussed in Sect. 2.2, where we looked at the sum of independent zero mean random variables, we can see that $u_{(n)}$ is of asymptotic order $\sqrt{n}$. Moreover $I_{(n)}$ is of order n, i.e. $I_{(n)}$ is proportional to n. Using Taylor series again we get

$$\left(s'_{(n)}\right)^{-1} = \left(I_{(n)} + u_{(n)}\right)^{-1} = I_{(n)}^{-1} + I_{(n)}^{-2} u_{(n)} + \ldots,$$

where components included in ... are of ignorable size as n increases. Looking at the order of the terms, we find $I_{(n)}^{-1}$ to be proportional to n^{-1} while the second component is asymptotically proportional to $n^{-2}\sqrt{n} = n^{-3/2}$ and can therefore also be ignored. Consequently, we can simplify (4.2.7) to

$$\hat{\theta}_{(n)} - \theta_0 = I_{(n)}^{-1} s_{(n)} + \ldots, \tag{4.2.8}$$

where, as above, the components collected in ... are of ignorable size. As $s_{(n)}$ is a sum of zero mean variables, the central limit theorem (4.2.2) applies which proves (4.2.1). □

Occasionally, we need or want to transform a parameter so that $\gamma = g(\theta)$ is the transformed value and $g(.)$ is an invertible and differentiable transformation, such that $\theta = g^{-1}(\gamma)$. The ML estimate $\hat{\gamma}_{ML}$ is **transformation invariant**, which means that it can simply be calculated as $\hat{\gamma}_{ML} = g(\hat{\theta}_{ML})$. This is easily seen, because the log-likelihood with parameterisation γ is defined by $l\left(g^{-1}(\gamma)\right)$. Hence

$$\frac{\partial l\left(g^{-1}(\gamma)\right)}{\partial \gamma} = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \theta}{\partial \gamma} \frac{\partial l}{\partial \theta}$$

such that for $\hat{\gamma}_{ML} = g(\hat{\theta}_{ML}) \Leftrightarrow \hat{\theta}_{ML} = g^{-1}(\hat{\gamma}_{ML})$ it follows

$$\frac{\partial l\left(g^{-1}(\hat{\gamma}_{ML})\right)}{\partial \gamma} = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \underbrace{\frac{\partial l(\hat{\theta}_{ML})}{\partial \theta}}_{=0} = 0.$$

The root of the original score function is also the root with the new transformed parameter. Transformation of parameters often requires the calculation of the Fisher information for the transformed parameter. This is fortunately straightforward. If we use the parameterisation with γ instead of θ, we get for the second order derivative

$$\begin{aligned}\frac{\partial^2 l\left(g^{-1}(\gamma)\right)}{\partial \gamma \partial \gamma} &= \frac{\partial}{\partial \gamma}\left(\frac{\partial g^{-1}(\gamma)}{\partial \gamma}\frac{\partial l(\theta)}{\partial \theta}\right) \\ &= \frac{\partial^2 g^{-1}(\gamma)}{(\partial \gamma)^2}\frac{\partial l(\theta)}{\partial \theta} + \frac{\partial g^{-1}(\gamma)}{\partial \gamma}\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta}\frac{\partial g^{-1}(\gamma)}{\partial \gamma}.\end{aligned}$$

The first component has mean zero when calculated at the true parameter θ_0, because the expectation of the score function $\partial l(\theta_0)/\partial \theta$ vanishes. Hence for the Fisher information we get

$$I_\gamma(\gamma) = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} I(\theta) \frac{\partial g^{-1}(\gamma)}{\partial \gamma} = \frac{\partial \theta}{\partial \gamma} I(\theta) \frac{\partial \theta}{\partial \gamma}, \tag{4.2.9}$$

where $I_\gamma(.)$ refers to the Fisher information using parameter γ. This result allows us to derive asymptotic normality for transformed ML estimates, a property sometimes called the **delta rule**.

Property 4.3 Let $\gamma = g(\theta)$ be a transformed parameter with an invertible and differentiable function g and let $\hat{\theta}$ be the Maximum Likelihood estimate with Fisher information $I(\theta)$. Then

$$\hat{\gamma}_{ML} - \gamma_0 \overset{a}{\sim} N\left(0, \frac{\partial \gamma}{\partial \theta} I^{-1}(\theta_0) \frac{\partial \gamma}{\partial \theta}\right).$$

In other words, transformation of parameters does not require the recalculation of the ML estimate or its variance. Instead, one can simply make use of the estimate of the untransformed parameter. However, this result is asymptotic and the quality of the approximation could be distorted by the transformation. We finish off this section with a number of examples that demonstrate the usefulness of this property.

Example 8 This example demonstrates the use of the delta rule to transform the parameter of a binomial distribution. Assume that Y is binomially distributed with $Y \sim B(n, \pi)$. The ML estimate is given by

$$\hat{\pi}_{ML} = Y/n \overset{a}{\sim} N(\pi, \underbrace{\frac{\pi(1-\pi)}{n}}_{I^{-1}(\pi)}).$$

If we now use the log odds as the transformed parameter we get

$$\begin{aligned} \vartheta = g(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) \\ \frac{\partial g(\pi)}{\partial \pi} &= \frac{1-\pi}{\pi} \cdot \frac{(1-\pi)+\pi}{(1-\pi)^2} = \frac{1}{\pi(1-\pi)} \\ \hat{\vartheta}_{ML} &= \log\left(\frac{Y/n}{1-Y/n}\right) \overset{a}{\sim} N\left(\log\left(\frac{\pi}{1-\pi}\right), \frac{1}{n(\pi(1-\pi))}\right). \end{aligned}$$

▷

Example 9 We are now able to prove an interesting result in parameter estimation, namely that the Maximum Likelihood and method of moments estimates are the same for exponential family distributions. Assume an exponential family distribution

$$f(y; \theta) = \exp\{t^T(y)\theta - \kappa(\theta)\}.$$

The log-likelihood function then is given by

$$l(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} t^T(y_i)\theta - n\kappa(\theta)$$

and the score function is vector valued and is given by

$$\begin{aligned} s(\theta; y_1, \ldots, y_n) &= \sum_{i=1}^{n} t(y_i) - nE\big(t(Y)\big) \\ \Rightarrow I(\theta) &= Var(s(\theta, Y_1, \ldots, Y_n)) = n \cdot Var(t(Y_i)). \end{aligned}$$

where we used the fact that $\partial\kappa(\theta) = E(t(Y))$. Therefore the Fisher information given by

$$I(\theta) = Var(s(\theta, Y_1, \ldots, Y_n)) = n \cdot Var(t(Y_i)),$$

Note also that

$$\partial^2 \kappa(\theta)/\partial\theta\partial\theta^T = Var\big(t(Y)\big).$$

Hence, the ML estimate $\hat{\theta}$ fulfils

$$\sum_{i=1}^{n} t(y_i) = nE_{\hat{\theta}}\big(t(Y)\big),$$

where the expectation is calculated with parameter $\hat{\theta}$. Consequently, we can interpret the ML estimate as a method of moments estimate.

▷

Example 10 This example demonstrates the calculation of the asymptotic distribution of the ML estimate for the mean and variance of normally distributed random variables. Let $Y_i \sim N(\mu, \sigma^2)$ *i.i.d.*. We denote with $y = (y_1, \ldots, y_n)$ the sample, such that the log-likelihood is given by

$$l(\mu, \sigma^2; y) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i - \mu)^2.$$

This leads to the two dimensional score function

$$\frac{\partial l(\mu, \sigma^2; y)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) \stackrel{!}{=} 0 \tag{4.2.10}$$

$$\frac{\partial l(\mu, \sigma^2; x)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu)^2 \stackrel{!}{=} 0 \tag{4.2.11}$$

$$\Leftrightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 \stackrel{!}{=} 0.$$

The system of equations can be solved by solving (4.2.10), which gives $\widehat{\mu} = \bar{y} = \sum_{i=1}^{n} y_i/n$. This can then be inserted into (4.2.11), giving

$$\frac{n}{\sigma} = \frac{1}{\sigma^3}\sum(y_i - \bar{y})^2 \qquad \Leftrightarrow \qquad \hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2.$$

Note that the estimation of σ^2 is biased, but for n going to infinity we have asymptotic unbiasedness. With the asymptotic results derived for the ML estimate, we get

$$\sqrt{n}\begin{pmatrix} \bar{Y} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \stackrel{a}{\sim} N(0, \Sigma)$$

with

$$\Sigma = I^{-1}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}^{-1} = \frac{1}{n} \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

Note that the covariance of the two estimators $\hat{\mu}$ and $\hat{\sigma}^2$ is 0, which means that the estimators of the parameters μ and σ^2 are independent.

▷

4.3 Numerical Calculation of ML Estimate

The calculation of the ML estimate corresponds to finding the root of the score function $s(\theta; y) \stackrel{!}{=} 0$. Apart from very simple cases, an analytical solution is not available and numerical methods are needed to solve the score function. A convenient and commonly used method in this respect is Fisher scoring. This is a statistical version of the classical Newton-Raphson procedure. Note that in first order approximation we have

$$0 = s(\hat{\theta}_{ML}; y) \approx s(\theta_0; y) + \frac{\partial s(\theta_0; y)}{\partial \theta}(\hat{\theta}_{ML} - \theta_0).$$

Rewriting this gives

$$\hat{\theta}_{ML} = \theta_0 - \left(\frac{\partial s(\theta_0; y)}{\partial \theta} \right)^{-1} s(\theta_0; y).$$

In many models, the derivative of the score is rather complicated and in order to simplify it, we replace $\partial s(\theta; y)/\partial \theta$ with its expectation:

$$\hat{\theta}_{ML} \approx \theta_0 + I^{-1}(\theta_0) s(\theta_0; y).$$

A simple iteration scheme follows naturally from this formula. Starting with an initial estimate $\theta_{(0)}$ and setting $t = 0$ we iterate as follows:

1. Calculate $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)}) s(\theta_{(t)}; y)$
2. Repeat step 1 until $||\theta_{(t+1)} - \theta_{(t)}|| < d$
3. Set $\hat{\theta}_{ML} = \hat{\theta}_{(t+1)}$

As with any Newton-Raphson procedure, the process may fail if the starting value $\theta_{(0)}$ is too far away from the target value $\hat{\theta}_{ML}$. In this case, it can help to work with a reduced step size. Hence, one can add some $0 < \delta < 1$:

$$\theta_{(t+1)} = \theta_{(t)} + \delta I^{-1}(\theta_{(t)}) s(\theta_{(t)}; y).$$

This step size can even be chosen adaptively based on the current step t, i.e. $\delta(t)$, which can even be traced as far back as Robbins and Monro (1951).

In recent years, models have become more complex and sometimes the calculation of neither the score function nor the Fisher matrix is possible analytically. In this case, simulation based methods can help. This can be applied to exponential family distributions as follows. Let

$$f(y; \theta) = \exp\{t^T(y)\theta - \kappa(\theta)\}$$

such that

$$s(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} t(y_i) - nE\big(t(Y)\big).$$

Hence, the Maximum Likelihood estimate is defined by

$$\frac{1}{n}\sum_{i=1}^{n} t(y_i) = E_{\hat{\theta}_{ML}}\big(t(Y)\big).$$

The idea is now to simulate Y from $f(y; \theta)$. If this is possible, we may simulate for a given $\theta_{(t)}$

$$Y_j^* \sim f(y; \theta_{(t)}) \qquad j = 1, \ldots, N.$$

With these simulations, we can now estimate the mean and the variance of $t(Y)$ with

$$\widehat{E_{\theta_{(t)}}(t(Y))} = \frac{1}{n}\sum_{j=1}^{N} t(y_j^*)$$

$$\widehat{Var_{\theta_{(t)}}(t(Y))} = \frac{1}{N}\sum_{j=1}^{N}\left(t(y_j^*) - \widehat{E_{\theta_{(t)}}(t(Y))}\right)\left(t(y_j^*) - \widehat{E_{\theta_{(t)}}(t(Y))}\right).$$

This allows now to replace the iteration step during Fisher scoring with

$$\theta_{(t+1)} = \theta_{(t)} + \widehat{Var_{\theta_{(t)}}(t(Y))}^{-1} \widehat{E_{\theta_{(t)}}(t(Y))}.$$

Although it sounds like a formidable effort to simulate the score function and the Fisher information, with increasing computing power these operations are becoming bearable. More details regarding simulation based calculation of the ML estimate are provided in Geyer (1992).

4.4 Likelihood-Ratio

So far we have looked at the asymptotic properties of the ML estimate but the properties of the log-likelihood function itself are also very useful. We therefore define the likelihood-ratio as

$$lr(\hat{\theta}; \theta) = l(\hat{\theta}) - l(\theta) = \log \frac{L(\hat{\theta})}{L(\theta)} \geq 0.$$

The likelihood-ratio is positive and it takes value 0 if $\theta = \hat{\theta}$. If we consider the estimate $\hat{\theta}$ as a random variable, the likelihood-ratio itself becomes a random variable, for which we can derive some asymptotic properties. With Taylor series expansion we get

$$l(\theta) \approx l(\hat{\theta}) + \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta}}_{=0}(\theta - \hat{\theta}) + \frac{1}{2}\frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta}(\theta - \hat{\theta})^2. \tag{4.4.1}$$

Using Eq. (4.2.7), or similarly (4.2.8), we can approximate $\hat{\theta} - \theta$ with $I^{-1}(\theta)s(\theta; Y_1, \ldots, Y_n)$. We can similarly approximate the second order derivative in (4.4.1) with $-I(\theta)$, such that (4.4.1) simplifies to

$$l(\theta_0) \approx l(\hat{\theta}) - \frac{1}{2}\frac{s^2(\theta_0; Y_1, \ldots, Y_n)}{I(\theta_0)}. \tag{4.4.2}$$

We have shown the asymptotic normality of $s(\theta_0; Y_1, \ldots, Y_n)$ in (4.2.2). Note that $Var(s(\theta_0; Y_1, \ldots, Y_n)) = I(\theta_0)$, such that the latter component in (4.4.2) is equal to the square of the score divided by its standard deviation. In other words, the last component in (4.4.2) is asymptotically equal to Z^2 with $Z \sim N(0, 1)$ being standard normally distributed. Hence, we asymptotically get a chi-squared distribution for the likelihood-ratio.

Property 4.4 (Likelihood-Ratio) The likelihood-ratio for a Fisher-regular distribution converges for sample size n increasing to a chi-squared distribution $\mathcal{X}_1^2$, that is

$$2\{l(\hat{\theta}_0) - l(\theta)\} \overset{a}{\sim} \mathcal{X}_1^2. \tag{4.4.3}$$

Subscript 1 in (4.4.3) refers to the degrees of freedom of the chi-squared distribution and, in fact, if $\theta \in \mathbb{R}^p$, we find that the likelihood-ratio converges to a chi-squared distribution with p degrees of freedom. The likelihood-ratio has proven itself to be quite powerful in statistical testing, which will be discussed in the next chapter.

4.5 Exercises

Exercise 1
In a clinical study for a certain disease, n patients are treated with a new drug while another n patients are treated with a placebo. Let $Y_1 \sim Bin(n, p_1)$ the number of diseased patients in the drug group and $Y_0 \sim Bin(n, p_0)$ the number of diseased patients in the placebo group. We assume that the groups are independent. An interesting measure is the *relative risk*

$$RR = \frac{p_1}{p_0} \in \mathbb{R}_+ .$$

Consider a family of estimates

$$\widehat{RR}_\theta = \frac{\hat{p}_1 + \theta}{\hat{p}_0 + \theta}, \theta > 0 \text{ and } \hat{p}_1 = Y_1/n,\ \hat{p}_0 = Y_0/n.$$

1. Derive the asymptotic distribution of $\log(\widehat{RR})_\theta$. (Note: assume that $\hat{p}_0$ and $\hat{p}_1$ are independent and asymptotically normally distributed and apply the delta rule.)
2. Calculate the asymptotic mean value and the variance of the estimate $\log(\widehat{RR})_\theta$.
3. Derive an asymptotic 95% confidence interval for RR.

Exercise 2 (Use `R` Statistical Software)
The toxicity of a chemical substance is tested by exposing it to beetles in different concentrations. The data are given in the file `ch4exerc2.csv`. The following table shows the results:

Experiment	Concentration	NumberExposed	NumberDied
1	1.70	60	5
2	1.72	60	12
3	1.75	62	19
4	1.79	60	29
5	1.80	60	51
6	1.84	60	54
7	1.86	65	62
8	1.89	65	65

`numberExposed` is the number of beetles exposed to the corresponding concentration, `numberDied` is the number of beetles that died at that concentration.

1. Three different models are used to estimate the influence of toxicity on the probability that a beetle dies given a certain concentration x. Let Y be the binary response with $Y = 1$ if a beetle dies at concentration x with $\pi(x) = P(Y = 1|x)$ and $Y = 0$ otherwise.

   (a) The probability is linked to the concentration x through the logistic function

   $$\pi(x) = \frac{\exp(\alpha_1 + \beta_1 x)}{1 + \exp(\alpha_1 + \beta_1 x)}.$$

   (b) The probability is linked through the probit function

   $$\pi(x) = \Phi(\alpha_2 + \beta_2 x) \ ,$$

   where $\Phi(\cdot)$ is the distribution function of the normal distribution.

   (c) The probability is linked through the complementary log-log function

   $$\pi(x) = 1 - \exp[-\exp(\alpha_3 + \beta_3 x)].$$

   For all three models, determine the likelihood and log-likelihood given the above data. Find the Maximum Likelihood estimates of the parameters (α_j, β_j), $j = 1, 2, 3$ using a generic optimisation function in R, e.g. the function `optim`.
2. Alternatively, a Fisher Scoring or Newton algorithm can be used for maximising the likelihood. Develop a suitable algorithm for the three models and derive the score function as well as the expected and observed Fisher information.
3. Using the Maximum Likelihood estimates $(\hat{\alpha}_j, \hat{\beta}_j)$, $j = 1, 2, 3$, calculate the expected proportion of dead beetles for each concentration. Compare the results with the raw proportions (`numberDied`/`numberExposed`) of the data and visualise the results in an appropriate plot.
4. Think about how to determine which of the three models has the best fit to the observed proportions.