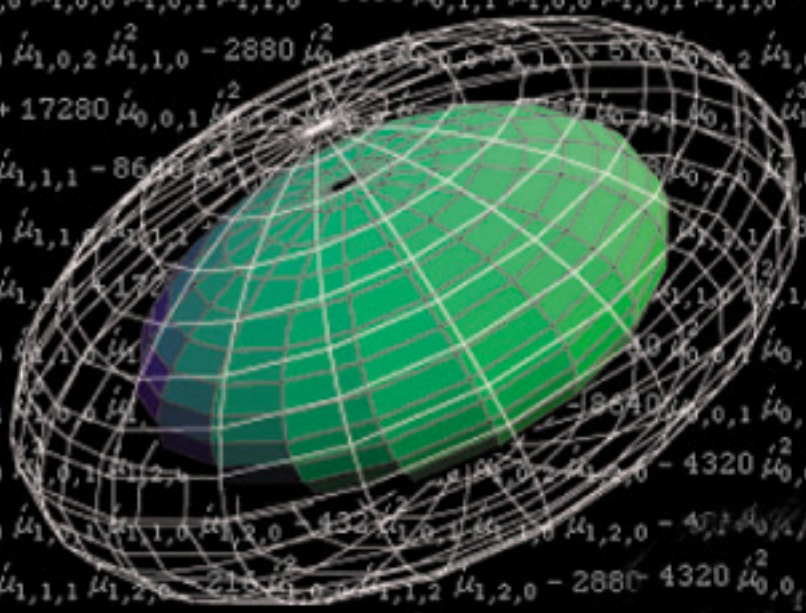
SPRINGER TEXTS IN STATISTICS

MATHEMATICAL STATISTICS

with *Mathematica*®



COLIN ROSE

MURRAY D. SMITH

Mathematical Statistics with *Mathematica*

Chapter 11 – Principles of Maximum Likelihood Estimation

Please reference this 2002 edition as:

Rose, C. and Smith, M.D. (2002)
Mathematical Statistics with Mathematica, Springer-Verlag, New York.

Latest edition

For the latest up-to-date edition, please visit: www.mathStatica.com

Chapter 11

Principles of Maximum Likelihood Estimation

11.1 Introduction

11.1 A Review

The previous chapter concentrated on obtaining unbiased estimators for parameters. The existence of unbiased estimators with minimum variance — the so-called MVUE class of estimators — required the sufficient statistics of the statistical model to be complete. Unfortunately, in practice, statistical models often falter in this respect. Therefore, parameter estimators must be found from other sources. The suitability of estimators based on large sample considerations such as consistency and limiting Normal distribution has already been addressed, as has the selection of estimators based on small sample properties dependent upon assumed loss structures. However, in both cases, the estimators that arose did so in an ad-hoc fashion. Fortunately, in the absence of complete sufficient statistics, there are other possibilities available. Of particular interest, here and in the following chapter, is the method of Maximum Likelihood (ML). ML techniques provide a way to generate parameter estimators that share some of the optimality properties, principally asymptotic ones.

§11.2 introduces the likelihood function. §11.3 defines the Maximum Likelihood Estimator (MLE) and shows how *Mathematica* can be used to determine its functional form. §11.4 discusses the statistical properties of the estimator. From the viewpoint of small sample sizes, the properties of the MLE depend very much on the particular statistical model in question. However, from a large sample perspective, the properties of the MLE are widely applicable and desirable: consistency, limiting Normal distribution and asymptotic efficiency. Desirable asymptotic properties and functional invariance (the Invariance Property) help to explain the popularity of ML in practice. §11.5 examines further the asymptotic properties of the MLE, using regularity conditions to establish these.

The statistical literature on ML methods is extensive with many texts devoting at least a chapter to the topic. The list of references that follow offers at least a sample of a range of treatments. In rough order of decreasing technical difficulty are Lehmann (1983), Amemiya (1985), Dhrymes (1970), Silvey (1975), Cox and Hinkley (1974), Stuart and Ord (1991), Gourieroux and Monfort (1995), Cramer (1986), McCabe and Tremayne (1993), Nerlove (2002), Mittelhammer (1996) and Hogg and Craig (1995). Currie (1995) gives numerical examples of computation of ML estimates using Version 2 of *Mathematica*, while Rose and Smith (2000) discuss computation under Version 4.

11.1 B SuperLog

Before embarking, we need to activate the **mathStatica** function SuperLog. This tool enhances *Mathematica*'s ability to simplify Log[Product[]] expressions. For instance, consider the following expression:

```
f = ∏_{i=1}^{n} (1 - θ)^(1 - x_i) θ^(x_i) ;        Log[f]
```

```
Log[∏_{i=1}^{n} (1 - θ)^(1-x_i) θ^(x_i)]
```

Mathematica has not simplified Log[f] at all. However, if we turn SuperLog on:

```
SuperLog[On]
```

```
– SuperLog is now On.
```

and try again:

```
Log[f]
```

```
n Log[1 - θ] + (-Log[1 - θ] + Log[θ]) ∑_{i=1}^{n} x_i
```

we obtain a significant improvement on *Mathematica*'s previous effort. SuperLog is part of the **mathStatica** suite. It modifies *Mathematica*'s Log function so that Log[Product[]] 'objects' or 'terms' get converted into sums of logarithms. At any stage, this enhancement may be removed by entering SuperLog[Off].

11.2 The Likelihood Function

In this section, we define the likelihood function and illustrate its construction in a variety of settings. To establish notation, let X denote the variable(s) of interest that has (or is assumed to have) a pdf $f(x; \theta)$ dependent upon a $(k \times 1)$ parameter $\theta \in \Theta \subset \mathbb{R}^k$ whose true value θ_0 is unknown; we assume that the functional form of f is known. Next, we let $(X_1, \ldots, X_n)$ denote a random sample of size n drawn on X. It is assumed that the pdf of the random sample $f_{1,\ldots,n}(x_1, \ldots, x_n; \theta)$ can be derived from the knowledge we have about f, and hence that the joint density depends on the unknown parameter θ. A key point is that the likelihood function is mathematically equivalent to the joint distribution of the sample. Instead of regarding it as a function of the X_i, the likelihood is interpreted as a function of θ defined over the parameter space Θ for fixed values of each $X_i = x_i$. The *likelihood* for θ is thus

$$L(\theta \mid x_1, \ldots, x_n) \equiv f_{1,\ldots,n}(x_1, \ldots, x_n; \theta). \tag{11.1}$$

Often, we will shorten the notation for the likelihood to just $L(\theta)$. Construction of the joint pdf may at first sight seem a daunting task. However, if the variables in $(X_1, \ldots, X_n)$ are

mutually independent, then the joint pdf is given by the product of the marginals,

$$f_{1,\ldots,n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{11.2}$$

which usually makes it easy to construct the joint pdf and hence the likelihood for θ.

We often need to distinguish between two forms of the likelihood for θ, namely, the likelihood function, and the observed likelihood. The *likelihood function* is defined as the likelihood for θ given the random sample prior to observation; it is given by $L(\theta \mid X_1, \ldots, X_n)$, and is a random variable. Where there is no possibility of confusion, we use 'likelihood' and 'likelihood function' interchangeably. The second form, the *observed likelihood*, is defined as the likelihood for θ evaluated for a given sample of observed data, and it is *not* random. The following examples illustrate the construction of the likelihood, and its observed counterpart.

⊕ ***Example 1:*** The Likelihood and Observed Likelihood for an Exponential Model

Let random variable $X \sim$ Exponential(θ), with pdf:

```
f = 1/θ e^(-x/θ);    domain[f] = {x, 0, ∞} && {θ > 0};
```

Let $(X_1, \ldots, X_n)$ denote a random sample of size n collected on X. Then, the *likelihood* for θ is equivalent to the joint pdf of the random sample (11.1), and as $(X_1, \ldots, X_n)$ are mutually independent, then it can be constructed as per (11.2):

```
Lθ = ∏_(i=1)^n (f /. x → x_i)
```

$$\prod_{i=1}^{n} \frac{e^{-\frac{x_i}{\theta}}}{\theta}$$

Given a random sample of size $n = 4$ on X, let us suppose that the observed data are:

```
data = {1, 2, 1, 4};
```

There are two main methods to construct the *observed likelihood* for θ:

Method 1: Substitute the data into the likelihood:

```
Lθ /. {n → Length[data], x_i_ :→ data[[i]]}
```

$$\frac{e^{-8/\theta}}{\theta^4}$$

Note the use of *delayed* replacement :→ (which is entered as :>). By contrast, *immediate* replacement → (which is entered as ->) would fail.

Method 2: Substitute the data into the density:

```
Times @@ (f /. x → data)
```

$$\frac{e^{-8/\theta}}{\theta^4}$$

Here, the immediate replacement `f /. x → data` yields a list of empirical densities $\{f(1;\theta), f(2;\theta), f(1;\theta), f(4;\theta)\}$. The observed likelihood for θ is obtained by multiplying the elements of the list together using `Times` (the `@@` is 'shorthand' for the `Apply` function). ■

⊕ ***Example 2:*** The Likelihood and Observed Likelihood for a Bernoulli Model

Now suppose that X is discrete, and, in particular, that $X \sim$ Bernoulli(θ):

```
f = θ^x (1 - θ)^(1-x);
domain[f] = {x, 0, 1} && {0 < θ < 1} && {Discrete};
```

where $0 < \theta < 1$. For $(X_1, \ldots, X_n)$, a random sample of size n drawn on X, the likelihood for θ is equivalent to the joint pmf of the random sample (11.1), and as $(X_1, \ldots, X_n)$ are mutually independent, it can be constructed as per (11.2):

```
Lθ = ∏_{i=1}^{n} (f /. x → x_i)
```

$$\prod_{i=1}^{n} (1-\theta)^{1-x_i}\,\theta^{x_i}$$

Suppose that observations were recorded as follows:

```
data = {1, 1, 0, 1, 0, 0, 1, 1, 0};
```

We again construct the observed likelihood using our two methods:

Method 1: Substitute the data into the likelihood:

```
∏_{i=1}^{n} (f /. x → x_i) /. {n → Length[data], x_i_ :→ data[[i]]}
```

$$(1-\theta)^4\,\theta^5$$

Method 2: Substitute the data into the pmf:

```
Times @@ (f /. x → data)
```

$$(1-\theta)^4\,\theta^5$$

⊕ ***Example 3:*** The Likelihood and Observed Likelihood for a Latent Variable Model

There are many instances where care is needed in deriving the likelihood. One important situation is when the variable of interest is latent (meaning that it cannot be observed), but a variable that is functionally related to it can be observed. To construct the likelihood for the parameters in a statistical model for a latent variable, we need to know the function (or the sampling scheme) that relates the observable variable to the latent variable.

Let X be the examination mark of a student in percent; thus $X = x \in [0, 100]$. Suppose that the mark is only revealed to us if the exam is passed; that is, X is disclosed provided $X \geq 50$. On the other hand, if the student fails the exam, then we receive a datum of 0 (say) and know only that $X < 50$. Thus, X is only partially observed by us and therefore it is latent. Let Y denote the observed variable, which is related to X by

$$Y = \begin{cases} X & \text{if } X \in [50, 100] \\ 0 & \text{if } X \in [0, 50). \end{cases} \tag{11.3}$$

We propose to model X with the (scaled) Beta distribution, $X \sim 100 \times \text{Beta}(a, b)$. Let $f(x; \theta)$ denote the statistical model for X:

```
f = (x/100)^(a-1) (1 - x/100)^(b-1) / (100 Beta[a, b]);

domain[f] = {x, 0, 100} && {a > 0, b > 0};
```

Although we cannot fully observe X, it is still possible to elicit information about the parameter $\theta = (a, b)$, as the relationship linking X to Y is known. Thus, given the distribution of X, we can derive the distribution of Y. The density of Y is non-standard in the sense that it has both discrete and continuous components. The discrete component of the density is a mass measured at the origin, while the continuous component of the density is equivalent to the pdf of X for values of 50 or more. By (11.3), the value of the mass at the origin is $P(Y = 0) = P(X < 50)$, which equals:

```
P0 = Prob[50, f]
```

$$\frac{\Gamma[a]\ \text{Hypergeometric2F1Regularized}[a, a+b, 1+a, -1]}{\text{Beta}[a, b]}$$

Let $(Y_1, \ldots, Y_n)$ denote a random sample of size n collected on Y (remember it is Y that is observed, not X). The likelihood for θ is, by (11.1), equivalent to the joint density of the random sample. Because of the component structure of the distribution of Y, it is convenient to introduce a quantity, n_0, defined to be the number of zeroes observed in the random sample — clearly $0 \leq n_0 \leq n$. Now, for a particular random sample $(y_1, \ldots, y_n)$, the likelihood is made up of contributions from both types of observations. For the n_0 zero observations it is

$$\prod_0 P(Y_i = 0) = (P(Y = 0))^{n_0}$$

where the product is taken over the n_0 zero observations. The contribution of the non-zero observations to the likelihood is

$$\prod_{+} f(y_i;\theta)$$

where the product is taken over the $(n - n_0)$ observations in the sample which are at least equal to 50, and f denotes the scaled Beta pdf. The likelihood is therefore

$$L(\theta) = (P(Y=0))^{n_0} \prod_{+} f(y_i;\theta). \tag{11.4}$$

To illustrate construction of the observed likelihood, we load the `CensoredMarks` data set into *Mathematica*:

```
data = ReadList["CensoredMarks.dat"];
```

There are a total of $n = 264$ observations in this data set:

```
n = Length[data]
```

```
264
```

Next, we select the marks of only those students that passed, storing them in the `PassMark` list:

```
PassMark = Select[data, (# ≥ 50) &];
```

```
n0 = n - Length[PassMark]
```

```
40
```

Calculation reveals that 40 of the 264 students must have received marks below 50, which implies a censoring (failure) rate of around 15%. As per (11.4), the observed likelihood for θ, given this data, is:

```
P0^n0 * Times @@ (f /. x → PassMark)
```

$$\frac{1}{\text{Beta}[a,b]^{264}} \big(2^{-40-202a-206b}\, 3^{-186+100a+86b}\, 5^{304-376a-376b}\, 7^{-65+31a+34b}\, 11^{-40+20a+20b}\, 13^{-25+14a+11b}\, 17^{-23+10a+13b}\, 19^{-31+17a+14b}\, 23^{-13+6a+7b}\, 29^{-20+13a+7b}\, 31^{-18+13a+5b}\, 37^{-15+4a+11b}\, 47^{-8+8b}\, 53^{-8+8a}\, 59^{-9+9a}\, 71^{-7+7a}\, 79^{-1+a}\, 1763^{-10+a+9b}\, 4087^{-6+6a}\, 6059^{-3+3a}\, \Gamma[a]^{40}\, \text{Hypergeometric2F1Regularized}[a, a+b, 1+a, -1]^{40}\big)$$

```
ClearAll[data, n, PassMark]; Unset[n0]; Unset[P0];
```

⊕ ***Example 4:*** The Likelihood and Observed Likelihood for a Time Series Model

In the previous examples, the likelihood function was easily constructed, since due to mutual independence, the joint distribution of the random sample was simply the product of the marginal distributions. In some situations, however, mutual independence amongst the sampling variables does not occur, and so the derivation of the likelihood function requires more effort. Examples include time series models, pertaining to variables collected through time that depend on their past.

Consider a random walk with drift model

$$X_t = \mu + X_{t-1} + U_t$$

with initial condition $X_0 = 0$. The drift is given by the constant $\mu \in \mathbb{R}$, while the disturbances U_t are assumed to be independently Normally distributed with zero mean and common variance $\sigma^2 \in \mathbb{R}_+$; that is, $U_t \sim N(0, \sigma^2)$, for all $t = 1, \ldots, T$, and $E[U_t\, U_s] = 0$ for all $t \neq s$.

We wish to construct the likelihood for parameter $\theta = (\mu, \sigma^2)$. One approach is to use conditioning arguments. We begin by considering the joint distribution of the sample $(X_1, \ldots, X_T)$. This cannot be written as the product of the marginals (*cf.* (11.2)) as X_t depends on $X_{t-1}, \ldots, X_0$, for all $t = 1, \ldots, T$. However, in light of this dependence, suppose instead that we decompose the joint distribution of the entire sample into the distribution of X_T conditional on all previous variables, multiplied by the joint distribution of all the conditioning variables:

$$\begin{aligned} f_{1,\ldots,T}(x_1, \ldots, x_T; \theta) \;=\; & f_{T\,|\,1,\ldots,T-1}(x_T \;|\; x_1, \ldots, x_{T-1}; \theta) \\ & \times f_{1,\ldots,T-1}(x_1, \ldots, x_{T-1}; \theta) \end{aligned} \tag{11.5}$$

where $f_{T\,|\,1,\ldots,T-1}$ denotes the distribution of X_T conditional on $X_1 = x_1, \ldots, X_{T-1} = x_{T-1}$, and $f_{1,\ldots,T-1}$ denotes the joint distribution of $(X_1, \ldots, X_{T-1})$. From the form of the random walk model, it is clear that when fixing any X_t, all previous X_s $(s < t)$ must also be fixed. This enables us to simplify the notation, for the conditional pdf on the right-hand side of (11.5) may be written as

$$f_{T|\,1,\ldots,T-1}(x_T \;|\; x_1, \ldots, x_{T-1}; \theta) \;=\; f_{T\,|\,T-1}(x_T \;|\; x_{T-1}; \theta). \tag{11.6}$$

From the assumptions on the disturbances, it follows that

$$X_T \;\big|\; (X_{T-1} = x_{T-1}) \;\sim\; N(\mu + x_{T-1}, \sigma^2) \tag{11.7}$$

which makes it is easy to write down the conditional density given in (11.6). Consider now the joint distribution of $(X_1, \ldots, X_{T-1})$ on the right-hand side of (11.5). Here, again, the same idea is used to decompose the joint distribution of the remaining variables: the appropriate equations are (11.5) and (11.6) but with T replaced by $T - 1$. By recursion,

$$\begin{aligned} f_{1,\ldots,T}(x_1, \ldots, x_T; \theta) \;=\; & f_{T\,|\,T-1}(x_T \;|\; x_{T-1}; \theta) \times f_{T-1\,|\,T-2}(x_{T-1} \;|\; x_{T-2}; \theta) \times \cdots \\ & \times f_{2\,|\,1}(x_2 \;|\; x_1; \theta) \times f_{1\,|\,0}(x_1 \;|\; (X_0 = 0); \theta) \end{aligned}$$

$$= \prod_{t=1}^{T} f_{t|t-1}(x_t \mid x_{t-1}; \theta) \tag{11.8}$$

where each of the conditional densities in (11.8) is equivalent to (11.6) for $t = 2, \ldots, T$, and $f_{1|0}$ is the pdf of a $N(\mu, \sigma^2)$ distribution because of the assumption $X_0 = 0$. By (11.1), (11.8) is equivalent to the likelihood for θ.

To enter this likelihood into *Mathematica*, we begin by entering the time t conditional pdf given in (11.7):

```
f = 1/(σ Sqrt[2 π]) Exp[-(x_t - μ - x_t-1)^2/(2 σ^2)];
```

Let us suppose we have data $\{x_1, \ldots, x_6\} = \{1, 2, 4, 2, -3, -2\}$:

```
xdata = {1, 2, 4, 2, -3, -2};
```

To obtain the observed likelihood, we use a modified form of *Method 1* that accounts for the initial condition $x_0 = 0$:

```
xlis = Thread[x_Range[Length[xdata]]];
xrules = Join[{x_0 → 0}, Thread[xlis → xdata]]
```

```
{x0 → 0, x1 → 1, x2 → 2, x3 → 4, x4 → 2, x5 → -3, x6 → -2}
```

Then, the observed likelihood for $\theta = (\mu, \sigma^2)$ is obtained by substituting in the observational rules:

```
obsLθ = Product_{t=1}^{6} f /. xrules // Simplify
```

$$\frac{e^{-\frac{18+2\mu+3\mu^2}{\sigma^2}}}{8\pi^3\sigma^6}$$

Figure 1 plots the observed likelihood against values of μ and σ^2. Evidently, `obsLθ` is maximised in the neighbourhood of $(\mu, \sigma^2) = (0, 6)$.

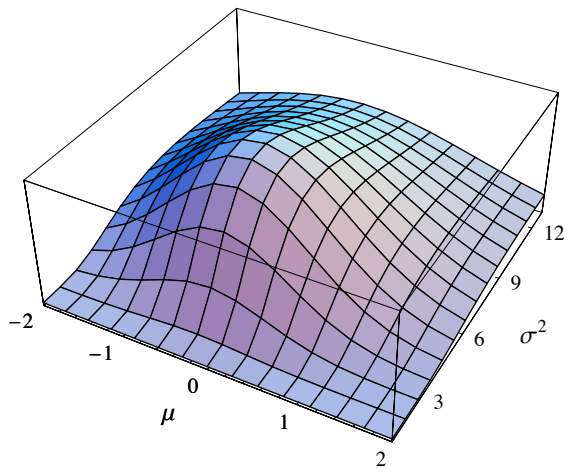


Fig. 1: Observed likelihood for μ and σ^2

11.3 Maximum Likelihood Estimation

Maximum likelihood parameter estimation is based on choosing values for θ so as to maximise the likelihood function. That is, the MLE of θ, denoted $\hat{\theta}$, is the solution to the optimisation problem:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta \mid X_1 = x_1, \ldots, X_n = x_n). \tag{11.9}$$

Thus, $\hat{\theta}$ is the value of the argument of the likelihood, selected from anywhere in the parameter space, that maximises the value of the likelihood after we have been given the sample. In other words, we seek the particular value of θ, namely, $\hat{\theta}$, which makes it most likely to have observed the sample that we actually have. We may view the solution to (11.9) in two ways depending on whether the objective function is the *likelihood function* or the *observed likelihood function*. If the objective is the likelihood, then (11.9) defines the ML *estimator*, $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$; since this is a function of the random sample, $\hat{\theta}$ is a random variable. If the objective is the observed likelihood, then (11.9) defines the ML *estimate*, $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$, where $(x_1, \ldots, x_n)$ denotes observed data; in this case $\hat{\theta}$ is a point estimate.

The solution to (11.9) is invariant to any monotonic increasing transformation of the objective. Since the natural logarithm is a monotonic transformation, it follows that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta) \tag{11.10}$$

which we shall use, from now on, as the definition of the estimator (estimate). The natural logarithm of the likelihood, $\log L(\theta)$, is called the *log-likelihood function*. A weaker definition of the MLE, but one that, in practice, is often equivalent to (11.10) is

$$\hat{\theta} = \arg\max_{\tilde{\theta} \in \tilde{\Theta}} \log L(\tilde{\theta}) \tag{11.11}$$

where $\tilde{\Theta}$ denotes a finite, non-null set whose elements $\tilde{\theta}$ satisfy the conditions

$$\frac{\partial}{\partial \theta} \log L(\tilde{\theta}) = 0 \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} \log L(\tilde{\theta}) < 0. \tag{11.12}$$

The two parts of (11.12) express, respectively, the *first-* and *second-order conditions* familiar from basic calculus for determining local maxima of a function.[1] Generally speaking, we shall determine MLE through (11.12), although *Example 7* below relies on (11.10) alone. One further piece of notation is the so-called *score* (or 'efficient score' in some texts), defined as the gradient of the log-likelihood,

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta).$$

For example, the first-order condition is simply $S(\tilde{\theta}) = 0$.

```
Clear[n];
```

⊕ ***Example 5:*** The MLE for the Exponential Parameter

Let $X \sim$ Exponential(θ), where parameter $\theta \in \mathbb{R}_+$. Here is its pdf:

```
f = 1/θ e^(-x/θ) ;     domain[f] = {x, 0, ∞} && {θ > 0};
```

For a random sample of size n drawn on X, the log-likelihood function is:

```
logLθ = Log[ ∏_{i=1}^{n} (f /. x → x_i) ]
```

$$-\frac{n\,\theta\,\text{Log}[\theta] + \sum_{i=1}^{n} x_i}{\theta}$$

Of course, this will only work if `SuperLog` has been activated (see §11.1 B). The score function is the gradient of the log-likelihood with respect to θ:

```
score = Grad[logLθ, θ]
```

$$\frac{-n\,\theta + \sum_{i=1}^{n} x_i}{\theta^2}$$

where we have applied **mathStatica**'s `Grad` function. Setting the score to zero and solving for θ corresponds to the first-order condition given in (11.12). We find:

```
solθ = Solve[score == 0, θ]
```

$$\left\{\left\{\theta \to \frac{\sum_{i=1}^{n} x_i}{n}\right\}\right\}$$

The unique solution, `solθ`, appears in the form of a replacement rule and corresponds to the sample mean. The nature of the solution is not yet clear; that is, does the sample mean correspond to a local minimum, local maximum, or saddle point of the log-likelihood? A check of the second-order condition, evaluated at `solθ`:

```
Hessian[logLθ, θ] /. Flatten[solθ]
```

$$-\frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2}$$

… reveals that the Hessian is strictly negative at the sample mean and therefore the log-likelihood is maximised at the sample mean. Hence, the MLE of θ is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Note that `Hessian[`f`,` x`]` is a **mathStatica** function. ■

⊕ ***Example 6:*** The MLE for the Normal Parameters

Let $X \sim N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$, with pdf $f(x; \mu, \sigma^2)$:

```
f = 1/(σ Sqrt[2 π]) Exp[-(x - μ)^2/(2 σ^2)];        domain[f] = {x, -∞, ∞};
```

For a random sample of size n drawn on X, the log-likelihood for parameter $\theta = (\mu, \sigma)$ is:[2]

```
logLθ = Log[∏_{i=1}^{n} (f /. x → x_i)]
```

$$-\frac{n\,(\mu^2 + \sigma^2\,\text{Log}[2\,\pi] + 2\,\sigma^2\,\text{Log}[\sigma]) - 2\,\mu\,\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}{2\,\sigma^2}$$

The score vector $S(\theta) = S(\mu, \sigma)$ is given by:

```
score = Grad[logLθ, {μ, σ}]
```

$$\left\{\frac{-n\,\mu + \sum_{i=1}^{n} x_i}{\sigma^2},\ \frac{n\,\mu^2 - n\,\sigma^2 - 2\,\mu\,\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}{\sigma^3}\right\}$$

Mathematica's `Solve` command is quite flexible in allowing various forms of the first-order conditions to be entered; for example, `{score[[1]] == 0, score[[2]] == 0}` or `score == {0, 0}`, or `score == 0`. Setting the score to zero and solving yields:

```
solθ = Solve[score == 0, {μ, σ}]
```

$$\left\{\left\{\sigma \to -\frac{\sqrt{-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \sum_{i=1}^{n} x_i^2}}{\sqrt{n}},\ \mu \to \frac{\sum_{i=1}^{n} x_i}{n}\right\},\right.$$

$$\left.\left\{\sigma \to \frac{\sqrt{-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \sum_{i=1}^{n} x_i^2}}{\sqrt{n}},\ \mu \to \frac{\sum_{i=1}^{n} x_i}{n}\right\}\right\}$$

Clearly, the negative-valued solution for σ lies outside the parameter space and is therefore invalid; thus, the only permissible solution to the first-order conditions is:

```
solθ = solθ[[2]]
```

$$\left\{\sigma \to \frac{\sqrt{-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \sum_{i=1}^{n} x_i^2}}{\sqrt{n}},\ \mu \to \frac{\sum_{i=1}^{n} x_i}{n}\right\}$$

Then $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ is the MLE of θ, where $\hat{\mu}$ and $\hat{\sigma}$ are the formulae given in `solθ` (we check second-order conditions below). The functional form given by *Mathematica* for $\hat{\sigma}$

may appear unfamiliar. However, if we utilise the following identity for the sum of squared deviations about the sample mean,

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, then

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2}\,.$$

By the Invariance Property (see §11.4 E), the MLE of σ^2 is

$$(\hat{\sigma})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

which is the 2nd sample central moment.

The second-order conditions may, for example, be checked by examining the eigenvalues of the Hessian matrix evaluated at $\hat{\theta}$:

```
Eigenvalues[Hessian[logLθ, {μ, σ}] /. solθ] // Simplify
```

$$\left\{ \frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2}\,,\ \frac{2\,n^3}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2} \right\}$$

Given the identity for the sum of squared deviations, the eigenvalues of the Hessian are $-n\,\hat{\sigma}^{-2}$ and $-2\,n\,\hat{\sigma}^{-2}$, which clearly are negative. Thus, the Hessian is negative definite at $\hat{\theta}$ and therefore the log-likelihood is maximised at $\hat{\theta}$. ■

⊕ ***Example 7:*** The MLE for the Pareto Parameters

Let $X \sim$ Pareto(α, β), where parameters $\alpha \in \mathbb{R}_+$ and $\beta \in \mathbb{R}_+$. The pdf of X is given by:

```
f = α β^α x^-(α + 1);    domain[f] = {x, β, ∞} && {α > 0, β > 0};
```

Since $X \geq \beta$, there exists dependence between the parameter and sample spaces. Given a random sample of size n collected on X, the log-likelihood for $\theta = (\alpha, \beta)$ is:

```
logLθ = Log[∏_{i=1}^{n} (f /. x → x_i)]
```

$$n\,(\text{Log}[\alpha] + \alpha\,\text{Log}[\beta]) - (1 + \alpha) \sum_{i=1}^{n} \text{Log}[x_i]$$

The score vector is given by:

```
score = Grad[logLθ, {α, β}]
```

$$\left\{n\left(\frac{1}{\alpha} + \text{Log}[\beta]\right) - \sum_{i=1}^{n} \text{Log}[x_i],\ \frac{n\,\alpha}{\beta}\right\}$$

If we attempt to solve the first-order conditions in the usual way:

```
Solve[score == 0, {α, β}]
```

```
{}
```

… we see that `Solve` cannot find a solution to the equations. However, if we focus on solving just the first of the first-order conditions, we find:[3]

```
solα = Solve[score[[1]] == 0, α]
```

$$\left\{\left\{\alpha \to -\frac{n}{n\,\text{Log}[\beta] - \sum_{i=1}^{n} \text{Log}[x_i]}\right\}\right\}$$

This time a solution is provided, albeit in terms of β; that is, $\hat{\alpha} = \hat{\alpha}(\beta)$. We now take this solution and substitute it back into the log-likelihood:

```
logLθ /. Flatten[solα] // Simplify
```

$$n\left(-1 + \text{Log}\left[\frac{n}{-n\,\text{Log}[\beta] + \sum_{i=1}^{n} \text{Log}[x_i]}\right]\right) - \sum_{i=1}^{n} \text{Log}[x_i]$$

This function is known as the *concentrated log-likelihood*. It corresponds to $\log L(\hat{\alpha}(\beta), \beta)$. Since it no longer involves α, we can maximise it with respect to β. Let $\hat{\beta}$ denote the solution to this optimisation problem. This solution can then be substituted back to recover $\hat{\alpha} = \hat{\alpha}(\hat{\beta})$; then $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}\right)$ would be the MLE of θ. In general, when the first-order conditions can be solved uniquely for some subset of parameters in θ, then those solutions can be substituted back into the log-likelihood to yield the concentrated log-likelihood. The concentrated log-likelihood is then maximised with respect to the remaining parameters, usually using numerical techniques.

For our example, maximising the concentrated log-likelihood using standard calculus will not work. This is because the parameter space depends on the sample space. However, by inspection, it is apparent that the concentrated log-likelihood is increasing in β. Therefore, we should select β as large as possible. Now, since each $X_i \geq \beta$, we can choose β no larger than the smallest observation. Hence, the MLE for β is

$$\hat{\beta} = \min(X_1, X_2, \ldots, X_n)$$

which is the smallest order statistic. Replacing β in $\hat{\alpha}(\beta)$ with $\hat{\beta}$ yields the MLE for α,

$$\hat{\alpha} = n \Big/ \sum_{i=1}^{n} \log\left(\frac{X_i}{\min(X_1, X_2, \ldots, X_n)}\right). \qquad \blacksquare$$

11.4 Properties of the ML Estimator

11.4 A Introduction

This section considers the small and large sample statistical properties of the MLE. Typically, small sample properties of a MLE are determined on a case-by-case basis. Finding the distribution of the estimator is the most important — its pdf and/or cdf, mgf or cf — for from this we can determine the moments of the estimator and construct confidence intervals about point estimates, and so on. Unlike, say, the MVUE class of estimator, whose properties are supported by a set of elegant theorems, the MLE has only limited small sample properties. Generally though, the MLE has the 'property' of being biased. The MLE properties are listed in Table 1.

Sufficiency	The MLE is a function of sufficient statistics.
Efficiency	If an estimator is BUE, then it is equivalent to the MLE, provided that the MLE is the unique solution to the first-order condition that maximises the log-likelihood function.
Asymptotic	Under certain regularity conditions, the MLE is *consistent*; it has a *limiting Normal distribution* when suitably scaled; and it is *asymptotically efficient*.
Invariance	If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Table 1: General properties of ML estimators

For proofs of these properties see, amongst others, Stuart and Ord (1991). The *Invariance* property is particularly important for estimation and it will be extensively exploited in the following chapter. Under fairly general conditions, the *Asymptotic* properties of the MLE are quite desirable; it is the attractiveness of its large sample properties which has contributed to the popularity of this estimator in practice. Even if the functional form of the MLE is not known (*i.e.* the solution to (11.12) can only be obtained by numerical methods), one can assert asymptotic properties by checking regularity conditions; in such situations, it is popular to use simulation techniques to determine small sample properties.

In §11.4 B, we examine the small sample properties of the MLE. Then, in §11.4 C, some of the estimators asymptotic properties are derived. In §11.4 D, further asymptotic properties of the MLE are revealed as a result of the model being shown to satisfy certain regularity conditions. Finally, in §11.4 E, the invariance property is illustrated. We begin with *Example 8*, which describes the model and derives the MLE.

⊕ ***Example 8:*** The MLE of θ

Let the continuous random variable X have pdf $f(x; \theta)$:

```
f = θ x^(θ-1);   domain[f] = {x, 0, 1} && {θ > 0};
```

where parameter $\theta \in \mathbb{R}_+$. The distribution of X can be viewed as either a special case of the Beta distribution (*i.e.* Beta(θ, 1)), or as a special case of the Power Function distribution (*i.e.* PowerFunction(θ, 1)). Assuming `SuperLog` has been activated (see §11.1 B), the log-likelihood for θ is derived with:

```
logLθ = Log[∏_{i=1}^{n} (f /. x → x_i)]
```

$$\texttt{n Log[}\theta\texttt{] + (-1 +}\theta\texttt{)} \sum_{\texttt{i=1}}^{\texttt{n}} \texttt{Log[x}_\texttt{i}\texttt{]}$$

In this example, the MLE of θ is the unique solution to the first-order condition:

```
solθ = Solve[Grad[logLθ, θ] == 0, θ]
```

$$\left\{\left\{\theta \to -\frac{\texttt{n}}{\sum_{\texttt{i=1}}^{\texttt{n}} \texttt{Log[x}_\texttt{i}\texttt{]}}\right\}\right\}$$

… because the log-likelihood is globally concave with respect to θ; that is, the Hessian is negative-valued at all points in the parameter space:

```
Hessian[logLθ, θ]
```

$$-\frac{\texttt{n}}{\theta^2}$$

Thus, the MLE of θ is

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log(X_i)}. \blacksquare \tag{11.13}$$

11.4 B Small Sample Properties

The sufficiency and efficiency properties listed in Table 1 pertain to the small sample performance of the MLE. The first property (sufficiency; see §10.4), is desirable because sufficient statistics retain all statistical information about parameters, and therefore so too must the MLE. Despite this, the MLE does not always use this information in an optimal fashion, for generally the MLE is a biased estimator.[4] Consequently, the second property (efficiency; see §10.3), should be seen as a special situation in which the MLE is unbiased and its variance attains the Cramér–Rao Lower Bound.

⊕ ***Example 9:*** Sufficiency, Efficiency and $\hat{\theta}$

Consider again the model given in *Example 8*, with pdf $f(x; \theta)$:

```
f = θ x^(θ-1);   domain[f] = {x, 0, 1} && {θ > 0};
```

The first property claims that there should exist a functional relationship between a sufficient statistic for θ and the MLE $\hat{\theta}$, given in (11.13). This can be shown by identifying a sufficient statistic for θ. Following the procedure given in §10.4, we apply **mathStatica**'s `Sufficient` function to find:

```
Sufficient[f]
```

$$\theta^{n} \prod_{i=1}^{n} x_i^{-1+\theta}$$

Then, by the Factorisation Criterion, the statistic $S = \prod_{i=1}^{n} X_i$ is sufficient for θ. We therefore have

$$\hat{\theta} = -\frac{n}{\log(S)}$$

and so the MLE is indeed a function of a sufficient statistic for θ.

The second property states that the MLE is the BUE provided the latter exists, and provided the MLE is the unique solution to the first-order conditions. Unfortunately, even though it was demonstrated in *Example 8* that $\hat{\theta}$ uniquely solved the first-order conditions, there is no BUE in this case. Nevertheless, the MVUE of θ does exist (since S is a complete sufficient statistic for θ) and it is given by

$$\tilde{\theta} = -\frac{n-1}{\log(S)}.$$

It is easy to see that the MLE $\hat{\theta}$ and the MVUE $\tilde{\theta}$ are related by a simple scaling transformation, $\tilde{\theta} = \frac{n-1}{n}\hat{\theta}$. In light of this, it follows immediately that the MLE must be biased upwards. ■

⊕ ***Example 10:*** The Distribution of $\hat{\theta}$

Consider again the model given in *Example 8*, with pdf $f(x; \theta)$:

```
f = θ x^(θ-1);   domain[f] = {x, 0, 1} && {θ > 0};
```

In this example, we derive the (small sample) distribution of the MLE

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log(X_i)}$$

by applying the MGF Theorem (see §2.4 D). We begin by deriving the mgf of

$$\overline{\log X} = -\frac{1}{n}\sum_{i=1}^{n} \log(X_i)$$

and then matching it to the mgf of a known distribution. In this way, we obtain the distribution of $\overline{\log X}$. The final step involves transforming from $\overline{\log X}$ to $\hat{\theta}$.

By the MGF Theorem, the mgf of $\overline{\log X}$ is:

```
Expect[ℯ^(t Log[x]), f]^n /. t → -t/n
```

– This further assumes that: {t + Θ > 0}

$$\left(\frac{\Theta}{-\frac{t}{n}+\Theta}\right)^n$$

This expression matches the mgf of a Gamma(n, $\frac{1}{n\theta}$) distribution.[5] Hence, $\overline{\log X} \sim$ Gamma(n, $\frac{1}{n\theta}$). Then, since $\hat{\theta} = 1/\overline{\log X}$, it follows that $\hat{\theta}$ has an Inverse Gamma distribution with parameters n and $\frac{1}{n\theta}$. That is,

$$\hat{\theta} \sim \text{InverseGamma}(n, \tfrac{1}{n\theta}).$$

The pdf of $\hat{\theta}$, say $f_{\hat{\theta}}$, can be entered from **mathStatica**'s *Continuous* palette:

```
f_θ̂ = θ̂^-(a+1) ℯ^-(1/(b θ̂)) / (Γ[a] b^a) /. {a → n, b → 1/(n Θ)};
domain[f_θ̂] = {θ̂, 0, ∞} && {n > 0, n ∈ Integers, Θ > 0};
```

We now determine the mean (although we have already deduced its nature through the relation between $\hat{\theta}$ and $\tilde{\theta}$ given in *Example 9*) and the variance of the MLE:

```
Expect[θ̂, f_θ̂]
```

– This further assumes that: {n > 1}

$$\frac{n\Theta}{-1+n}$$

```
Var[θ̂, f_θ̂] // FullSimplify
```

– This further assumes that: {n > 2}

$$\frac{n^2\Theta^2}{(-2+n)(-1+n)^2}$$

11.4 C Asymptotic Properties

Recall that estimators may possess large sample properties such as asymptotic unbiasedness, consistency, asymptotic efficiency, be limit Normally distributed when suitably scaled, and so on. These properties are also relevant to ML estimators. Like the small sample properties, large sample properties can be examined on a case-by-case basis. Analysis might proceed by applying the appropriate Central Limit Theorem and Law of Large Numbers.

⊕ ***Example 11:*** Asymptotic Unbiasedness and Consistency of $\hat{\theta}$

Consider the model of *Example 8*, with pdf $f(x; \theta)$:

```
f = θ x^(θ-1);   domain[f] = {x, 0, 1} && {θ > 0};
```

Since we have already shown $E[\hat{\theta}] = \frac{n\theta}{n-1}$ in *Example 10*, it is particularly easy to establish whether or not $\hat{\theta}$ is asymptotically unbiased for θ:

```
Limit[n θ / (n - 1), n → ∞]
```

```
θ
```

As the mean of $\hat{\theta}$ tends to θ as n increases, we say that $\hat{\theta}$ is asymptotically unbiased for θ. Here we have defined asymptotic unbiasedness such that $\lim_{n\to\infty} E[\hat{\theta}] = \theta$. Note that there are other definitions of asymptotic unbiasedness in use in the literature. For example, an estimator may be termed asymptotically unbiased if the mean of its asymptotic distribution is θ. In most cases, such as the present one, this second definition will coincide with the first so that there is no ambiguity.

We can also establish whether or not $\hat{\theta}$ is a consistent estimator of θ by using Khinchine's Weak Law of Large Numbers (see §8.5 C), and the Continuous Mapping Theorem. Consider

$$\overline{\log X} = \frac{1}{n}\sum_{i=1}^{n}(-\log(X_i))$$

which is in the form of a sample mean. Each variable in the sum is mutually independent, identically distributed, with mean

```
Expect[-Log[x], f]
```

$$\frac{1}{\theta}$$

Therefore, by Khinchine's Theorem, $\overline{\log X} \xrightarrow{p} \theta^{-1}$. As $\hat{\theta} = 1/(\overline{\log X})$, $\hat{\theta} \xrightarrow{p} \theta$ by the Continuous Mapping Theorem.[6] Therefore, the MLE $\hat{\theta}$ is a consistent estimator of θ.

The next asymptotic property concerns the limiting distribution of $\sqrt{n}\left(\hat{\theta} - \theta\right)$. Unfortunately, in this case, it is *not* possible to derive the limiting distribution using the asymptotic theory presented so far. If we apply Lindeberg–Lévy's version of the Central Limit Theorem (see §8.4) to $-\sum_{i=1}^{n}\log(X_i)$, we can only get as far as stating,

$$\frac{\sum_{i=1}^{n}(-\log(X_i)) - n\theta^{-1}}{\theta^{-1}\sqrt{n}} = \sqrt{n}\left(\frac{\theta}{\hat{\theta}} - 1\right) \xrightarrow{d} Z \sim N(0, 1).$$

To proceed any further, we must establish whether or not certain regularity conditions are satisfied by the distribution of X.[7] ■

11.4 D Regularity Conditions

To derive (some of) the asymptotic properties of $\hat{\theta}$, we used the fact that we knew the estimator's functional form, just as we did when determining its small sample properties. Alas, the functional form of the MLE is often unknown; how then are we to determine the asymptotic properties of the MLE? Fortunately, there exist sets of regularity conditions that, if satisfied, permit us to make relatively straightforward statements about the asymptotic properties of the MLE. Those stated here apply if the random sample is a collection of mutually independent, identically distributed random variables, if the parameter θ is a scalar, and if there is a unique solution to the first-order condition that globally maximises the log-likelihood function. This ideal setting fits our particular case.

Let θ_0 denote the 'true value' of θ, let i_0 denote the Fisher Information on θ evaluated at $\theta = \theta_0$, and let n denote the sample size. Under the previously mentioned conditions, the MLE has the following asymptotic properties,

consistency	$\hat{\theta} \xrightarrow{p} \theta_0$
limit Normal distribution	$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0, i_0^{-1})$
asymptotic efficiency	relative to all other consistent uniformly limiting Normal estimators

Table 2: Asymptotic properties of the MLE, given regularity conditions

under the following *regularity conditions*:

1. The parameter space Θ is an open interval of the real line within which θ_0 lies.
2. The probability distributions defined by any two different values of θ are distinct.
3. For any finite n, the first three derivatives of the log-likelihood function with respect to θ exist in an open neighbourhood of θ_0.
4. In an open neighbourhood of θ_0, the information identity for Fisher Information holds:

$$i_0 = E\Big[\Big(\frac{\partial}{\partial\theta}\log f(X;\theta_0)\Big)^2\Big] = -E\Big[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta_0)\Big].$$

   Moreover, i_0 is finite and positive.

5. In an open neighbourhood of θ_0:

   (i) $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log L(\theta_0) \xrightarrow{d} N(0, i_0)$

   (ii) $-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log L(\theta_0) \xrightarrow{p} i_0$

   (iii) For some constant $M < \infty$, $\frac{1}{n}\left|\frac{\partial^3}{\partial\theta^3}\log L(\theta_0)\right| \xrightarrow{p} M$.

For discussion about the role of regularity conditions in determining asymptotic properties of estimators such as the MLE, see, for example, Cox and Hinkley (1974), Amemiya (1985) and McCabe and Tremayne (1993).

⊕ ***Example 12:*** Satisfying Regularity Conditions

The model of *Example 8*, with pdf $f(x; \theta_0)$, is given by:

```
f = θ0 x^(θ0 - 1);   domain[f] = {x, 0, 1} && {θ0 > 0};
```

Note that the parameter of the distribution is given at its true value θ_0.

The first regularity condition is satisfied as the parameter space $\Theta = \{\theta : \theta \in \mathbb{R}_+\}$ is an open interval of the real line, within which we assume θ_0 lies. The second condition pertains to parameter identification and is satisfied in our single-parameter case. For the third condition, the first three derivatives of the log-likelihood function evaluated at θ_0 are:

```
Table[D[Log[∏_(i=1)^n (f /. x → x_i)], {θ0, j}], {j, 3}]
```

$$\left\{\frac{n}{\theta_0} + \sum_{i=1}^{n} \text{Log}[x_i],\ -\frac{n}{\theta_0^2},\ \frac{2\,n}{\theta_0^3}\right\}$$

and each exists within a neighbourhood about θ_0 (wherever that might be). Next, the information identity is satisfied:

```
FisherInformation[θ0, f, Method → 1] ==
 FisherInformation[θ0, f, Method → 2]
```

```
True
```

Moreover, the Fisher Information i_0 is equal to:

```
FisherInformation[θ0, f]
```

$$\frac{1}{\theta_0^2}$$

which is finite, so the fourth condition is satisfied. From the derivatives of the log-likelihood function, we can establish that the fifth condition is satisfied. For 5(i),

$$\frac{1}{\sqrt{n}}\,\frac{\partial}{\partial\theta}\log L(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\log X_i + \theta_0^{-1})$$

which, by the Lindeberg–Lévy version of the Central Limit Theorem, is $N(0, i_0)$ in the limit, as each term in the summand has mean and variance:

```
Expect[Log[x] + 1/θ0, f]
```

```
0
```

```
Var[Log[x] + 1/θ0, f]
```

$$\frac{1}{\theta_0^2}$$

For 5(ii),

$$-\frac{1}{n}\,\frac{\partial^2}{\partial\theta^2}\log L(\theta_0) \;=\; \theta_0^{-2} \;=\; i_0$$

for every n, including in the limit. For 5(iii),

$$\frac{1}{n}\left|\,\frac{\partial^3}{\partial\theta^3}\log L(\theta_0)\,\right| = 2\,\theta_0^{-3}$$

is non-stochastic and finite for every n, including in the limit. In conclusion, each regularity condition is satisfied. Thus, $\hat{\theta}$ is consistent for θ_0, $\sqrt{n}\,\hat{\theta}$ has a limit Normal distribution, in particular, $\sqrt{n}\,(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,\,\theta_0^2)$, and $\hat{\theta}$ is asymptotically efficient. These results enable us, for example, to construct the estimator's asymptotic distribution: $\hat{\theta} \overset{a}{\sim} N(\theta_0,\,\theta_0^2/n)$, which may be contrasted against the estimator's exact distribution $\hat{\theta} \sim \text{InverseGamma}\left(n,\,\frac{1}{n\,\theta_0}\right)$ found in §11.4 B. ■

11.4 E Invariance Property

Throughout this section, our example has concentrated on estimation of θ. But suppose another parameter λ, related functionally to θ, is also of interest. Given what we already know about $\hat{\theta}$, it is usually possible to obtain the MLE of λ and to establish its statistical properties by the Invariance Property (see Table 1), provided we know the functional form that links λ to θ.

Consider a multi-parameter setting in which θ is a $(k\times 1)$ vector and λ is a $(j\times 1)$ vector, where $j \le k$. The link from θ to λ is through a vector function g; that is, $\lambda = g(\theta)$, where g is assumed known. The parameters are such that $\theta \in \Theta$ and $\lambda \in \Lambda$, with the particular true values once again indicated by a 0 subscript. The parameter spaces are $\Theta \subset \mathbb{R}^k$ and $\Lambda \subset \mathbb{R}^j$, so that $g : \Theta \to \Lambda$. Moreover, we assume that g is a continuous function of θ, and that the $(j\times k)$ matrix of partial derivatives

$$G(\theta) = \frac{\partial g(\theta)}{\partial \theta^{\mathrm{T}}}$$

has finite elements and is of full row rank; that is, $\text{rank}(G(\theta)) = j$, for all $\theta \in \Theta$.

Of particular use is the case when $j = k$, for then the dimensions of θ and λ are the same and $G(\theta)$ becomes a square matrix having full rank (which means that the inverse function g^{-1} must exist). In this case, the parameter λ is said to represent a *re-parameterisation* of θ. There are a number of examples of re-parameterisation in the next chapter, the idea there being to transform a constrained optimisation problem in θ (occurring when Θ is a proper subset of $\mathbb{R}^k$) into an unconstrained optimisation problem in λ (re-parameterisation achieves $\Lambda = \mathbb{R}^k$).

The key results of the Invariance Property apply to the MLE of $g(\theta)$ and to its asymptotic properties. First, if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $\lambda = g(\theta)$. This is an extremely useful property for it means that if we already know $\hat{\theta}$, then we do *not* need to find the MLE of λ_0 by maximising the log-likelihood $\log L(\lambda)$. Second, if $\hat{\theta}$ is consistent, and has a limiting Normal distribution when suitably scaled, and is asymptotically efficient, then so too is $\hat{\lambda} = g(\hat{\theta})$. That is, if

$$\hat{\theta} \xrightarrow{p} \theta_0 \tag{11.14}$$

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\vec{0},\, i_0^{-1}\right) \tag{11.15}$$

then

$$g(\hat{\theta}) \xrightarrow{p} g(\theta_0) \tag{11.16}$$

$$\sqrt{n}\left(g(\hat{\theta}) - g(\theta_0)\right) \xrightarrow{d} N\left(\vec{0},\, G(\theta_0) \times i_0^{-1} \times G(\theta_0)^{\mathrm{T}}\right). \tag{11.17}$$

The small sample properties of $\hat{\lambda}$ generally cannot be deduced from those of $\hat{\theta}$, but must be examined on a case-by-case basis. To see this, a simple example suffices. Let $\lambda = g(\theta) = \theta^2$, and suppose that the MLE $\hat{\theta}$ is unbiased. By the Invariance Property, the MLE of λ is $\hat{\lambda} = \hat{\theta}^2$; however, it is *not* necessarily true that $\hat{\lambda}$ is unbiased for λ, for in general $E[\hat{\theta}^2] \neq \left(E[\hat{\theta}]\right)^2$.

⊕ ***Example 13:*** The Invariance Property

The model of *Example 8*, with pdf $f(x; \theta)$, is given by:

```
f = θ x^(θ-1);   domain[f] = {x, 0, 1} && {θ > 0};
```

Consider the parameter $\lambda = E[X]$:

```
λ = Expect[x, f]
```

$$\frac{\theta}{1+\theta}$$

Clearly, parameter $\lambda \in \Lambda = (0, 1)$ is a function of θ; $\lambda = g(\theta) = \theta/(1+\theta)$, with true value $\lambda_0 = g(\theta_0)$. To estimate λ_0, one possibility is to re-parameterise the pdf of X from θ to λ and repeat the same ML estimation procedures from the very beginning. But we can do better by applying the Invariance Property, for we already have the functional form of $\hat{\theta}$ (see (11.13)) as well as its asymptotic properties. The MLE of λ_0 is given by

$$\hat{\lambda} = \frac{\hat{\theta}}{1+\hat{\theta}} = \frac{n}{n - \sum_{i=1}^{n} \log(X_i)}.$$

Since g is continuously differentiable with respect to θ, it follows from (11.17) that the limiting distribution of $\hat{\lambda}$ is

$$\sqrt{n}\left(\hat{\lambda} - \lambda_0\right) \xrightarrow{d} N\Big(0, \left(\frac{\partial}{\partial\theta}\, g(\theta_0)\right)^2 \Big/ i_0\Big).$$

In particular, the variance of the limiting distribution of $\sqrt{n}\left(\hat{\lambda} - \lambda_0\right)$ in terms of θ_0, is given by:

```
Grad[λ, θ]^2
------------------------------ /. θ → θ0
FisherInformation[θ, f]
```

$$\frac{\theta_0^2}{(1+\theta_0)^4}$$

The asymptotic distribution of the MLE of λ_0 is therefore

$$\hat{\lambda} \overset{a}{\sim} N\Big(\lambda_0, \frac{\theta_0^2}{n(1+\theta_0)^4}\Big). \qquad \blacksquare$$

11.5 Asymptotic Properties: Extensions

The asymptotic properties of the MLE — consistency, a limiting Normal distribution when suitably scaled, and asymptotic efficiency — generally hold in a variety of circumstances far weaker than those considered in §11.4. In fact, there exists a range of regularity conditions designed to cater for a variety of settings involving various combinations of non-independent and/or non-identically distributed samples, parameter θ a vector, multiple local optima, and so on. In this section, we consider two departures from the setup in §11.4 D. Texts that discuss proofs of asymptotic properties of the MLE and regularity conditions include Amemiya (1985), Cox and Hinkley (1974), Dhrymes (1970), Lehmann (1983), McCabe and Tremayne (1993) and Mittelhammer (1996).

11.5 A More Than One Parameter

Suppose we now allow parameter θ to be k-dimensional, but otherwise keep the statisticial setup described in §11.4 unaltered; namely, the random sample consists of mutually independent and identically distributed random variables, and there is a unique solution to the first-order condition — a system of k equations — that maximises the log-likelihood function. Then, it seems reasonable to expect that regularity conditions 1, 4 and 5 given in §11.4 D need only be extended to account for the higher dimensionality of θ:

1a. The k-dimensional parameter space Θ must be of finite dimension as sample size n increases, it must be an open subset of $\mathbb{R}^k$, and it must contain the true value θ_0 within its interior.

4a. In an open neighbourhood of θ_0, the information identity for Fisher Information (a $(k \times k)$ symmetric matrix) holds. That is:

$$\begin{aligned} i_0 &= E\Big[\Big(\frac{\partial}{\partial\theta}\log f(X;\theta_0)\Big)\Big(\frac{\partial}{\partial\theta}\log f(X;\theta_0)\Big)^{\mathrm{T}}\Big] \\ &= -E\Big[\frac{\partial^2}{\partial\theta\,\partial\theta^{\mathrm{T}}}\log f(X;\theta_0)\Big]. \end{aligned}$$

Moreover, every element of i_0 is finite, and i_0 is positive definite.

5a. In an open neighbourhood of θ_0:

(i) $\frac{1}{\sqrt{n}}\,\frac{\partial}{\partial\theta}\log L(\theta_0) \xrightarrow{d} N\big(\vec{0},\, i_0\big)$

(ii) $-\frac{1}{n}\,\frac{\partial^2}{\partial\theta\,\partial\theta^{\mathrm{T}}}\log L(\theta_0) \xrightarrow{p} i_0$

(iii) Let indexes $u, v, w \in \{1, \ldots, k\}$ pick out elements of θ. For constants $M_{u,v,w} < \infty$, $\frac{1}{n}\Big|\frac{\partial^3}{\partial\theta_u\,\partial\theta_v\,\partial\theta_w}\log L(\theta_0)\Big| \xrightarrow{p} M_{u,v,w}$.

If these conditions hold, as well as conditions 2 and 3, then the MLE $\hat{\theta}$ is a consistent estimator of θ, $\sqrt{n}\,\big(\hat{\theta}-\theta_0\big)\xrightarrow{d} N\big(\vec{0},\, i_0^{-1}\big)$, and $\hat{\theta}$ is asymptotically efficient (*cf.* Table 2).

⊕ ***Example 14:*** The Asymptotic Distribution of $\hat{\theta}$: $X \sim$ Normal

Let $X \sim N(\mu_0, \sigma_0^2)$, with pdf $f(x; \mu_0, \sigma_0^2)$:

```
f = 1/(σ0 Sqrt[2 π]) Exp[-(x - μ0)^2/(2 σ0^2)];
domain[f] = {x, -∞, ∞} && {μ0 ∈ Reals, σ0 > 0};
```

In this case, the parameter $\theta = (\mu, \sigma^2)$ is two-dimensional ($k = 2$), with true value $\theta_0 = (\mu_0, \sigma_0^2)$. In *Example 6*, where $(X_1, \ldots, X_n)$ denoted a size n random sample drawn on X, the MLE of θ was derived as

$$\hat{\theta} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \end{pmatrix}.$$

The regularity conditions 1a, 2, 3, 4a, 5a hold in this case. The dimension k is fixed at 2 for all n, the parameter space $\Theta = \{\theta = (\mu, \sigma^2) : \mu \in \mathbb{R},\, \sigma^2 \in \mathbb{R}_+\}$ is an open subset of $\mathbb{R}^2$ within which we assume θ_0 lies, and the information identity holds:

```
FisherInformation[{μ0, σ0^2}, f, Method → 1] ==
 FisherInformation[{μ0, σ0^2}, f, Method → 2]
```

```
True
```

The Fisher Information matrix i_0 is equal to:

```
i0 = FisherInformation[{μ0, σ0^2}, f]
```

$$\begin{pmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\,\sigma_0^4} \end{pmatrix}$$

and it has finite elements and is positive definite. The asymptotic conditions 5a are satisfied too. We demonstrate 5a(i), leaving verification of 5a(ii) and 5a(iii) to the reader. For 5a(i), we require the derivatives of the log-likelihood function with respect to the elements of θ. Here is the log-likelihood:

```
logLθ = Log[∏_{i=1}^{n} (f /. x → xi)]
```

$$-\frac{n\,\mu_0^2 + n\,(\text{Log}[2\,\pi] + 2\,\text{Log}[\sigma_0])\,\sigma_0^2 - 2\,\mu_0\,\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}{2\,\sigma_0^2}$$

and here are the derivatives:

```
Grad[logLθ, {μ0, σ0^2}]
```

$$\left\{\frac{-n\,\mu_0 + \sum_{i=1}^{n} x_i}{\sigma_0^2},\ \frac{n\,\mu_0^2 - n\,\sigma_0^2 - 2\,\mu_0\,\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}{2\,\sigma_0^4}\right\}$$

For the first element, we have for 5a(i),

$$\frac{1}{\sqrt{n}}\,\frac{\partial}{\partial\mu}\log L(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu_0}{\sigma_0^2}$$

which, by the Lindeberg–Lévy version of the Central Limit Theorem, is $N(0, \sigma_0^{-2})$ in the limit, as each term in the summand has mean and variance:

```
Expect[(x - μ0)/σ0^2, f]
```

$$0$$

```
Var[(x - μ0)/σ0^2, f]
```

$$\frac{1}{\sigma_0^2}$$

Similarly, for the derivative with respect to σ^2,

$$\frac{1}{\sqrt{n}}\,\frac{\partial}{\partial\sigma^2}\log L(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{2\,\sigma_0^2}\left(\left(\frac{X_i - \mu_0}{\sigma_0}\right)^2 - 1\right)$$

which is $N(0, \frac{1}{2}\sigma_0^{-4})$ in the limit, as each term in the summand has mean and variance:

```
Expect[ 1/(2 σ0^2) ( ((x - μ0)/σ0)^2 - 1 ) , f]
```

```
0
```

```
Var[ 1/(2 σ0^2) ( ((x - μ0)/σ0)^2 - 1 ) , f]
```

$$\frac{1}{2\,\sigma_0^4}$$

Finally then, as $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are independent (see *Example 27* of Chapter 4):

$$\frac{1}{\sqrt{n}}\,\frac{\partial}{\partial\theta}\log L(\theta_0) \xrightarrow{d} N\left(\vec{0},\, i_0\right).$$

As all regularity conditions hold, $\sqrt{n}\,\left(\hat{\theta}-\theta_0\right)\xrightarrow{d} N\left(\vec{0},\, i_0^{-1}\right)$, with the variance-covariance matrix of the limiting distribution given by:

```
Inverse[ i0 ]
```

$$\begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\,\sigma_0^4 \end{pmatrix}$$

From this result we can find, for example, the asymptotic distribution of the MLE

$$\hat{\theta} \overset{a}{\sim} N\left(\begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix},\, \begin{pmatrix} \sigma_0^2/n & 0 \\ 0 & 2\sigma_0^4/n \end{pmatrix}\right).$$

This can be contrasted against the small sample distributions: $\hat{\mu} \sim N(\mu_0, \sigma_0^2/n)$ independent of $n\hat{\sigma}^2/\sigma_0^2 \sim$ Chi-squared$(n-1)$. ■

11.5 B Non-identically Distributed Samples

Suppose that the statistical setup described in §11.5 A is further extended such that ML estimation is based on a random sample which does *not* consist of identically distributed random variables. Despite the loss of identicality, mutual independence between the variables $(X_1, \ldots, X_n)$ in the sample ensures that the log-likelihood remains a sum:

$$\log L(\theta) = \sum_{i=1}^{n} \log f_i(x_i;\theta)$$

where $f_i(x_i;\theta)$ is the pdf of X_i. Accordingly, for the MLE to have the usual trio of asymptotic properties (see Table 1), the regularity conditions will need to be weakened even further in order that certain forms of the Central Limit Theorem and Law of Large Numbers relevant to sums of non-identically distributed random variables remain valid. The conditions requiring weakening are 4a, 5a(i) and 5a(ii):

4b. In an open neighbourhood of θ_0, the information identity for *asymptotic* Fisher Information (a $(k \times k)$ symmetric matrix) holds. That is:

$$i_0^{(\infty)} = \lim_{n\to\infty} E\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\partial}{\partial\theta}\log f(X_i;\theta_0)\Big)\Big(\frac{\partial}{\partial\theta}\log f(X_i;\theta_0)\Big)^{\mathrm{T}}\Big]$$

$$= \lim_{n\to\infty} E\Big[-\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^{\mathrm{T}}}\log L(\theta_0)\Big].$$

Moreover, every element of $i_0^{(\infty)}$ is finite, and $i_0^{(\infty)}$ is positive definite.

5b. In an open neighbourhood of θ_0:

(i) $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log L(\theta_0) \xrightarrow{d} N\big(\vec{0}, i_0^{(\infty)}\big)$

(ii) $-\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^{\mathrm{T}}}\log L(\theta_0) \xrightarrow{p} i_0^{(\infty)}$.

Should these conditions hold, as well as 1a, 2, 3 and 5a(iii), then the MLE $\hat{\theta}$ is a consistent estimator of θ, $\sqrt{n}\,\big(\hat{\theta}-\theta_0\big) \xrightarrow{d} N\big(\vec{0}, (i_0^{(\infty)})^{-1}\big)$, and $\hat{\theta}$ is asymptotically efficient.

⊕ ***Example 15:*** The Asymptotic Distribution of $\hat{\theta}$: Exponential Regression

Suppose that a positive-valued random variable Y depends on another random variable X, both of which are observed in pairs $((Y_1, X_1), (Y_2, X_2), \ldots)$. For example, Y may represent sales of a firm, and X may represent the firms advertising expenditure. We may represent this dependence by specifying a *conditional* statistical model for Y; that is, by specifying a pdf for Y, given that a value $x \in \mathbb{R}$ is assigned to X. One such model is the *Exponential Regression*, where $Y \,\big|\, (X = x) \sim$ Exponential$(\exp(\alpha_0 + \beta_0 x))$, with pdf $f(y \,|\, X = x; \theta_0)$:

```
f = 1/Exp[α0 + β0 x] Exp[-y/Exp[α0 + β0 x]];
domain[f] = {y, 0, ∞} && {α0 ∈ Reals, β0 ∈ Reals, x ∈ Reals};
```

The parameter $\theta = (\alpha, \beta) \in \mathbb{R}^2$, and its true value $\theta_0 = (\alpha_0, \beta_0)$ is unknown. The regression function is given by the conditional mean $E[Y \,|\, (X = x)]$, and this is equal to:

```
Expect[y, f]
```

$$e^{\alpha_0 + x\,\beta_0}$$

Despite the fact that the functional form of the MLE $\hat{\theta}$ cannot be derived in this case,[8] we can still obtain the asymptotic properties of the MLE by determining if the regularity conditions 1a, 2, 3, 4b, 5b(i), 5b(ii) and 5a(iii) are satisfied. In this example, we shall focus on obtaining the asymptotic Fisher Information matrix $i_0^{(\infty)}$ given in 4b. We begin by deriving the Fisher Information:

```
FisherInformation[{α0, β0}, f]
```

$$\begin{pmatrix} 1 & x \\ x & x^2 \end{pmatrix}$$

This output reflects the non-identicality of the distribution of $Y \mid (X = x)$, for Fisher Information quite clearly depends on the value assigned to X. Let $((Y_1, X_1), \ldots, (Y_n, X_n))$ denote a random sample of size n on the pair (Y, X). Because the distribution of $Y_i \mid (X_i = x_i)$ need not be identical to the distribution of $Y_j \mid (X_j = x_j)$ (for x_i need not equal x_j), then the Sample Information matrix is no longer given by Fisher Information multiplied by sample size; rather, Sample Information is given by the sample sum:

$$I_0 = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}.$$

Under independence, the log-likelihood is made up of a sum of contributions,

$$\log L(\theta) = \sum_{i=1}^n \log f\left(y_i \mid (X_i = x_i); \theta\right)$$

implying that $\frac{1}{n} I_0$ is exactly the expectation given in regularity condition 4b, when computed either way because

```
FisherInformation[{α0, β0}, f, Method → 1] ==
 FisherInformation[{α0, β0}, f, Method → 2]
```

```
True
```

To obtain the asymptotic Fisher Information matrix, we must examine the limiting behaviour of the elements of $\frac{1}{n} I_0$. This will require further assumptions about the marginal distribution of X. If the random variable X has finite mean μ, finite variance σ^2, with neither moment depending on n, then by Khinchine's Weak Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mu \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} \sigma^2 + \mu^2.$$

Under these further assumptions, we obtain the asymptotic Fisher Information matrix as

$$i_0^{(\infty)} = \begin{pmatrix} 1 & \mu \\ \mu & \sigma^2 + \mu^2 \end{pmatrix}$$

which is positive definite. Establishing conditions 5b(i), 5b(ii) and 5a(iii) involves similar manipulations, and in this case can be shown to hold under the assumptions concerning the behaviour of X. In conclusion, the asymptotic distribution of the MLE $\hat{\theta}$ of $\theta_0 = (\alpha_0, \beta_0)$ is, under the assumptions placed on X, given by

$$\hat{\theta} \overset{a}{\sim} N\left(\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}, \frac{1}{n\sigma^2}\begin{pmatrix} \sigma^2 + \mu^2 & -\mu \\ -\mu & 1 \end{pmatrix}\right). \qquad \blacksquare$$

11.6 Exercises

1. Let $X \sim \text{Poisson}(\lambda)$, where parameter $\lambda \in \mathbb{R}_+$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X. (i) Derive $\hat{\lambda}$, the ML estimator of λ. (ii) Obtain the exact distribution of $\hat{\lambda}$. (iii) Obtain the asymptotic distribution of $\hat{\lambda}$ (check regularity conditions).

2. Let $X \sim \text{Geometric}(p)$, where parameter p is such that $0 < p < 1$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X. Derive $\hat{p}$, the ML estimator of p, and obtain its asymptotic distribution.

3. Let $X \sim N(\mu, 1)$, where parameter $\mu \in \mathbb{R}$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X. (i) Derive $\hat{\mu}$, the ML estimator of μ. (ii) Obtain the exact distribution of $\hat{\mu}$. (iii) Obtain the asymptotic distribution of $\hat{\mu}$ (check regularity conditions).

4. Let $X \sim \text{ExtremeValue}(\theta)$, with pdf $f(x; \theta) = \exp(-(x - \theta) - e^{-(x-\theta)})$, where $\theta \in \mathbb{R}$ is an unknown parameter. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X. (i) Obtain $\hat{\theta}$, the ML estimator of θ. (ii) Obtain the asymptotic distribution of $\hat{\theta}$ (check regularity conditions).

5. For the pdf of the $N(0, \sigma^2)$ distribution, specify a *replacement rule* that serves to replace σ and its powers in the pdf. In particular, the rule you construct should act to convert the pdf from an input of

$$\frac{1}{\sigma\sqrt{2\pi}}\,\text{Exp}\left[-\frac{x^2}{2\sigma^2}\right] \quad \text{to an output of} \quad \frac{1}{\sqrt{\theta}\sqrt{2\pi}}\,\text{Exp}\left[-\frac{x^2}{2\theta}\right]$$

6. Let $X \sim N(0, \sigma^2)$, where parameter $\sigma^2 \in \mathbb{R}_+$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X.

(i) Derive $\hat{\sigma}^2$, the ML estimator of σ^2.

(ii) Obtain the exact distribution of $\hat{\sigma}^2$.

(iii) Obtain the asymptotic distribution of $\hat{\sigma}^2$ (check regularity conditions).

Hint: use your solution to Exercise 5.

7. Let $X \sim \text{Rayleigh}(\sigma^2)$, where parameter $\sigma^2 \in \mathbb{R}_+$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X.

(i) Derive $\hat{\sigma}^2$, the ML estimator of σ^2.

(ii) Obtain the exact distribution of $\hat{\sigma}^2$.

(iii) Obtain the asymptotic distribution of $\hat{\sigma}^2$ (check regularity conditions).

8. Let $X \sim \text{Uniform}(0, \theta)$, where parameter $\theta \in \mathbb{R}_+$ is unknown, and, of course, $X < \theta$. Let $(X_1, X_2, \ldots, X_n)$ denote a size n random sample drawn on X. Show that the largest order statistic $\hat{\theta} = X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ is the ML estimator of θ. Using **mathStatica**'s `OrderStat` function, obtain the exact distribution of $\hat{\theta}$. Transform $\hat{\theta} \to Y$ such that $Y = n(\theta - \hat{\theta})$. Then derive the limiting distribution of $n(\theta - \hat{\theta})$. Propose an asymptotic approximation to the exact distribution of $\hat{\theta}$.