Lecture Notes on Evolutionary Genetics

Alan R. Rogers[1]

February 29, 2008

[1] Department of Anthropology, University of Utah, Salt Lake City, UT 84112

Lecture 1

The Method of Maximum Likelihood

1.1 Introduction

Statisticians have developed several methods for estimating parameters from data. Of these, the method of maximum likelihood is both the most powerful and the most flexible. Yet it is seldom covered in introductory statistics courses because it is hard to present either as a canned computer program or within a cookbook-style textbook. It is much too flexibile for that.

The method of maximum likelihood is easiest to describe by example, so I begin with the simplest experiment that I can imagine.

1.2 Tossing a coin once

You are given a (possibly unfair) coin, and you toss it once. Since the probability of tossing heads is unknown, let us call it p. Let x represent the number of heads observed (either zero or one). If the coin comes up heads, then $x = 1$ and you have observed an event with probability p. On the other hand, if the coin comes up tails then $x = 0$ and you have observed an event with probability $1 - p$. In symbols,

$$\left.\begin{array}{rcl} \Pr[x=1] & = & p \\ \Pr[x=0] & = & 1-p \end{array}\right\} \tag{1.1}$$

1.2.1 The likelihood function

If we have performed a coin tossing experiment, then we will know the value of x but not that of p. Consequently, we can think of formula 1.1 as a function of this unknown parameter value. This is the reverse of the usual situation in probability theory, where the parameters are taken as given and the outcomes are allowed to vary. To distinguish these two situations, equation 1.1 is called a *probability distribution* if taken as a function of the outcome variable x, but is called a *likelihood* if taken as a function of its parameter, p. For example, if we toss one coin and observe heads, then we have observed an event of probability p. The likelihood function in this case is

$$L(p) = p$$

On the other hand, if the coin had come up tails, the likelihood function would be

$$L(p) = 1 - p$$

The likelihood function is useful because it summarizes the information that the data provide about the parameters. To estimate a parameter, we make use of the principle of maximum likelihood:

Principle 1 *To estimate a parameter, the method of maximum likelhood chooses the parameter value that makes L as large as possible.*

● EXAMPLE **1–1**
If the coin came up heads, then $L(p) = p$. This function reaches its maximum value when $p = 1$. The maximum likelihood estimate of p is therefore $\hat{p} = 1$.

(Here, $\hat{p}$ is pronounced "pee-hat." It is the conventional symbol for a maximum likelihood estimate of p.)

● EXAMPLE **1–2**
If the coin came up tails, then $L(p) = 1 - p$ and $\hat{p} = 0$.

1.3 Tossing the coin twice

Had we tossed the coin twice, there would have been four possible ordered outcomes:

Table 1.1: Ordered outcomes from two tosses of a coin

Outcome		
toss 1	toss 2	Probability
heads	heads	p^2
heads	tails	$p(1-p)$
tails	heads	$p(1-p)$
tails	tails	$(1-p)^2$

It is often inconvenient to keep track of the outcome of each toss, so the table above is usually abbreviated as:

Table 1.2: Unordered outcomes from two tosses of a coin

x	Probability
2	p^2
1	$2p(1-p)$
0	$(1-p)^2$

● EXAMPLE **1–3**
If you observe one head in two tosses, then $L(p) = 2p(1-p)^2$ and $\hat{p} = 1/2$.

● EXAMPLE **1–4**
If you observe two heads in two tosses, then $L(p) = p^2$ and $\hat{p} = 1$.

1.4 Tossing the coin several times

In the general case, the coin is tossed N times yielding x heads and $N - x$ tails. The probability of observing x heads in N tosses is given by the binomial distribution function:

$$\Pr[x; N, p] = \binom{N}{x} p^x (1-p)^{N-x} \tag{1.2}$$

In this expression, the notation $\binom{N}{x}$ is pronounced "N choose x" and represents the number of ways of choosing x heads out of N tosses.

● EXAMPLE **1–5**
In the case of two tosses, table 1.1 shows two ways of getting a single head. Consequently, $\binom{2}{1} = 2$, and equation 1.2 gives $\Pr[1; 2, p] = 2p(1-p)$. Note that this result agrees with that of the preceding paragraph.

● EXAMPLE **1–6**
Write down all the outcomes that can result from three tosses of a coin.

○ ANSWER

toss 1	toss 2	toss 3
heads	heads	heads
heads	heads	tails
heads	tails	heads
heads	tails	tails
tails	heads	heads
tails	heads	tails
tails	tails	heads
tails	tails	tails

● EXAMPLE **1–7**
In how many outcomes is $x = 0$? In other words, what is $\binom{3}{0}$?
○ ANSWER
$\binom{3}{0} = 1$

⋆ EXERCISE **1–1** What are $\binom{3}{1}$, $\binom{3}{2}$, and $\binom{3}{3}$?

The form of the binomial distribution function is not hard to understand. To see why, consider the probability of the following outcome:

$$\overbrace{\overbrace{H\;H\;\ldots H}^{x \text{ tosses}}\quad\overbrace{T\;T\;\ldots T}^{N-x \text{ tosses}}}^{N \text{ tosses}}$$

Here, the coin has been tossed N times, producing heads on the first x tosses and tails on the remaining $N - x$. Since each heads is an event of probability p and each tails is an event of probability $1 - p$, the probability of the outcome observed on this sequence of tosses is $p^x(1 - p)^{N-x}$. But this is not the only outcome that would yield x heads in N tosses. No matter what order the heads and tails appear in, if there are x heads in N tosses we have observed an event of probability $p^x(1 - p)^{N-x}$. If we don't know the order in which the heads and tails appear, we have to sum across all the ways in which x heads and $N - x$ tails can be re-ordered. This sum accounts for the term $\binom{N}{x}$ in equation 1.2.

Equation 1.2 becomes a likelihood function if we think of it as a function of p rather than x. For example, if we toss three coins and observe one head, the likelihood function is

$$L(p) = 3p(1 - p)^2$$

To estimate p, the method of maximum likelihood chooses the value of p that maximizes the likelihood. The value of p that maximizes L will also maximize $\ln L$, so we can work with either function. It is often more convenient to work with $\ln L$ rather than with L itself. If $x = 1$ and $N = 3$, the log likelihood function is

$$\ln L(p) = \ln 3 + \ln p + 2\ln(1 - p) \tag{1.3}$$

Figure 1.1 graphs $\ln L$ for several coin tossing experiments in each of which 1/3 of the tosses come up heads. Each curve reaches a maximum at roughly $p = 1/3$. Thus, the maximum likelihood estimator must be close to 1/3 in each case. We can see this much by studying the graph.

The curves in figure 1.1 also provide information about the precision of the estimates: The likelihood function is flat when the sample size is small, but is narrow and peaked when the sample size is large. This is a crucial point. It means that when N is large, our estimate of p is unlikely to be far from 1/3, the true value. The larger the data set, the stronger this claim becomes.

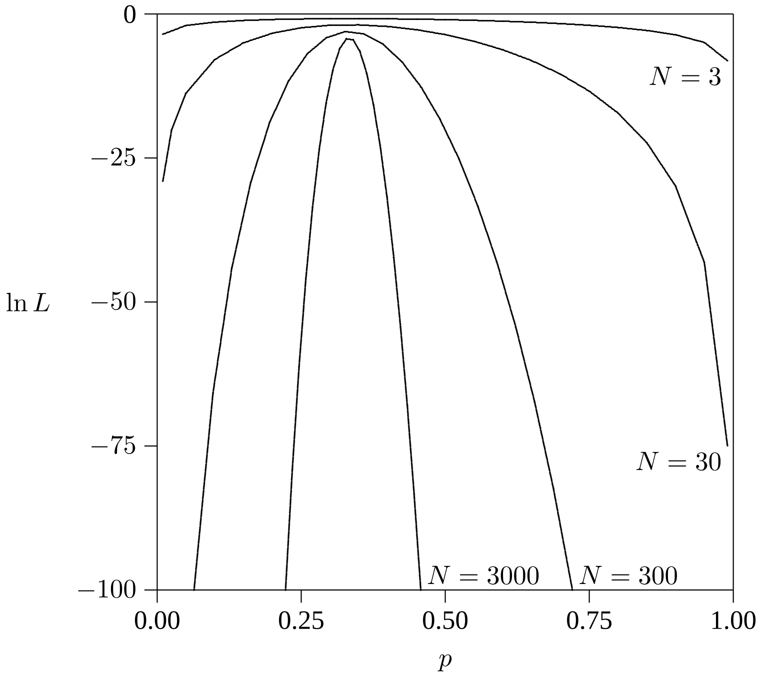

Figure 1.1: Log likelihood functions for binomial experiments in which $x/n = 1/3$

1.5 A maximum likelihood estimator for p

To obtain an estimator, it is convenient to work with the logarithm of L rather than with L itself. Taking the log of equation 1.2 gives

$$\ln L(p) = \ln \binom{N}{x} + x \ln p + (N - x) \ln(1 - p)$$

The maximum likelihood estimator of p is the value of p that makes $\ln L(p)$ as large as possible. This estimator turns out to be

$$\hat{p} = x/N \tag{1.4}$$

in agreement with the examples above where $x/N = 1/3$.

● EXAMPLE **1–8**

Verify equation 1.4 using Maple.

○ ANSWER

The first step is to define $\ln L$. Here, `Const` is $\ln\binom{N}{x}$, a constant that will have no effect on the answer. I have included it here only for the sake of clarity.

```
> lnL := Const + x*log(p) + (N-x)*log(1-p);
    lnL := Const + x ln(p)
    + (N - x) ln(1 - p)
```

The next step is to find $\hat{p}$ by solving the equation $d \ln L/dp = 0$:

```
> phat := solve(diff(lnL, p) = 0, p);
    phat := x/N
```

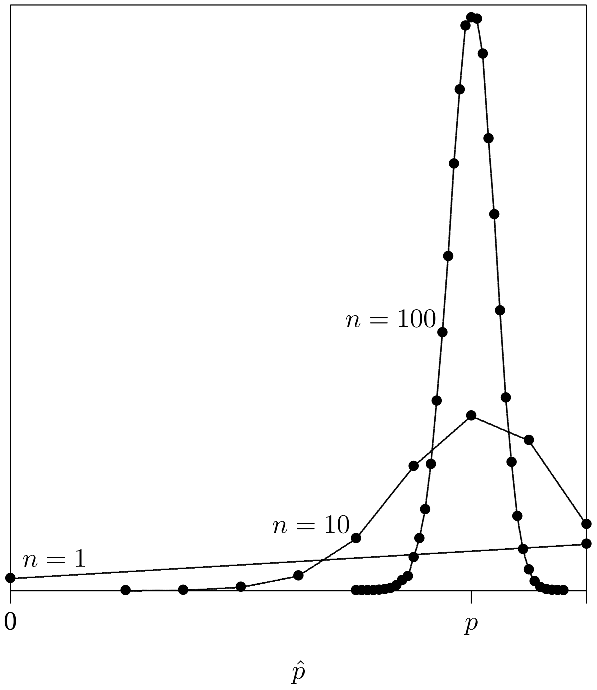

Figure 1.2: Frequency distributions of $\hat{p}$ when $p = 0.8$.

Finally, examine the second derivative to verify that this is indeed a local maximum:

```
> simplify(subs(p=phat, diff(lnL, p, p)));
              3
            N
        ----------
        (-N + x) x
```

If $x < N$, the entire expression is negative and $\hat{p}$ is at a local maximum. Thus, in this case $\hat{p}$ is a maximum likelihood estimator of p. When $x = N$, however, the second derivative is undefined, so the second-order condition provides no information.

How well does this formula work? To find out, I analyzed data from computer simulations in which $p = 0.8$. The results are shown in figure 1.2 and cover three values of N. For each value of N, I generated thousands of simulated data sets and estimated p from each data sets. The distributions of these estimates are shown in figure 1.2. First look at the distribution for $N = 100$. In that case, the distribution is centered narrowly around the true parameter value, $p = 0.8$. Few of the simulated estimates are far from the true value, so we could have high confidence in an estimate from 100 real coin tosses. Now look at the distribution for $N = 10$. In that case, the distribution is spread much more widely—it would be easy to find estimates that differ from the true value by 0.2 or so. With a sample of 10 we get only a crude idea of the value of p. Finally, look at the distribution for $N = 1$. Here the distribution goes from wall to wall. A single toss of the coin would tell little about p.

1.6 Sampling variance, standard error, and confidence intervals

1.6.1 Sampling variance

The variance of a maximum likelihood estimator is the reciprocal of the expectation of the 2nd derivative,

$$v = -1 \Bigg/ E\left\{\frac{d^2 \ln L}{dp^2}\right\} \tag{1.5}$$

For example, with equation 1.2 the first and second derivatives are

$$\begin{aligned} \frac{\partial \ln L}{\partial p} &= \frac{x}{p} - \frac{N-x}{1-p} \\ \frac{\partial^2 \ln L}{\partial p^2} &= -\frac{x}{p^2} - \frac{N-x}{(1-p)^2} \end{aligned}$$

The expected value of x is Np and that of $N-x$ is $N(1-p)$. Thus,

$$\begin{aligned} E\left\{\frac{\partial^2 \ln L}{\partial p^2}\right\} &= -\frac{Np}{p^2} - \frac{N(1-p)}{(1-p)^2} \\ &= -\frac{N}{p(1-p)} \end{aligned}$$

Plugging this into equation 1.5 gives the sampling variance for our estimate of p:

$$v = \frac{p(1-p)}{N}$$

This expresses the sampling variance of $\hat{p}$ in terms of the unknown parameter p. To use this answer with data, we would have to use $\hat{p}$ as an approximation for p. Thus, in practice the standard error is estimated as

$$v = \frac{\hat{p}(1-\hat{p})}{N} \tag{1.6}$$

It often turns out to be difficult to take the expectation of the second derivative. In such cases, the usual practice is to approximate equation 1.5 by

$$v \approx -1 \Bigg/ \left.\frac{d^2 \ln L}{dp^2}\right|_{p=\hat{p}} \tag{1.7}$$

Instead of taking the expectation of the second derivative, one simply evaluates it at the point where the parameter is equal to its estimate. With the example above,

$$\left.\frac{d^2 \ln L}{dp^2}\right|_{p=\hat{p}} = -\frac{N}{\hat{p}(1-\hat{p})}$$

so equations 1.7 and 1.5 both give the answer shown in equation 1.6.

● EXAMPLE **1–9**

When $\hat{p} = 0.8$ and $N = 100$, the estimated sampling variance is $v = 0.8 \times 0.2/100 = 0.0016$.

● EXAMPLE **1–10**

Use equations 1.5 and 1.7 to obtain a formula for the sampling variance of $\hat{p}$, which estimates the probabity of heads in a binomial experiment. Compare your answer to equation 1.6.

○ ANSWER

Begin by defining `lnL` and `phat` as in example 1–8 above. Then calculate the second derivative of `lnL`.

```
> # d2 is the 2nd derivative of lnL with
> # respect to p
> d2 := diff(lnL, p, p);
                x      N - x
    d2 := - ---- - --------
                 2          2
                p     (1 - p)
```

To take the expectation of this expression, replace x with its expected value Nx.

```
> # Ed2 is the expectation of d2
> Ed2 := simplify(subs(x = N*p, d2));
                  N
    Ed2 := ----------
              p (-1 + p)
```

Students who have studied probability theory will recognize that this substitution is justified because d2 is a linear function of the random variable x. The rest of you must take it on faith.

Now, equation 1.5 gives

```
> # v1 is the sampling variance of phat
> v1 := -1/Ed2;
              p (-1 + p)
    v1 := - ----------
                  N
```

To use this formula with data, we would have to substitute $\hat{p}$ for p, and this would give equation 1.6.

The next step is to use equation 1.7, which gives

```
> # v2 is also the sampling variance of phat
> v2 := simplify(-1 / subs(p=phat, d2));
              (-N + x) x
    v2 := - ----------
                    3
                  N
```

But $x = N\hat{p}$, so this can be re-expressed as:

```
> phat := 'phat';
    phat := phat

> v2 := simplify(subs(x = N*phat, v2));
              (-1 + phat) phat
    v2 := - ----------------
                     N
```

This is also equivalent to equation 1.6.

1.6.2 Standard error

The standard error, like the sampling variance, is an estimate of the error in an estimate. The larger the standard error of an estimate, the less acurrate that estimate is likely to be. The two estimates of error are closely related: the standard error is the square root of the sampling variance.

1.6.3 Confidence interval

What is a 95% confidence interval? Since a confidence interval is calculated from data, and the data themselves are random, the confidence interval is a random quantity too. If we repeat some experiment again and again and calculate a confidence interval from each fresh set of data, we will likely get a different confidence interval each time.

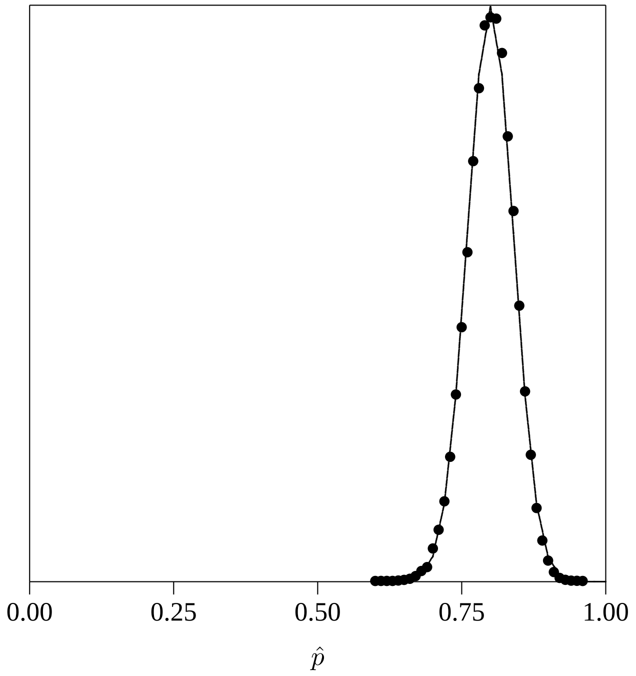

Figure 1.3: Normal approximation to sampling distribution

The solid line shows the normal approximation to the sampling distribution of $\hat{p}$ in the case where $p = 0.8$ and $N = 100$. The bullets are copied from the the corresponding curve in figure 1.2, and show the true sampling distribution of $\hat{p}$ as estimated by computer simulation.

The procedure for constructing confidence intervals is devised so that, on average, 95% of the intervals we construct will contain the true parameter value.

If a data set consists of a large number of independent observations, the maximum likelihood estimates that we obtain from it will have sampling distributions that are approximately normal. Consequently, normal distribution theory is usually used to place approximate confidence intervals around maximum likelihood estimates. In a normal distribution, 95% of the probability mass is within 1.96 standard deviations of the mean. Consequently, a 95% confidence interval for some parameter θ is the interval between

$$\hat{\theta} - 1.96S.E. \quad \text{and} \quad \hat{\theta} + 1.96S.E.$$

In words, the lower bound of the interval is 1.96 standard errors below the estimate and the upper bound is 1.96 standard errors above.

● EXAMPLE **1–11**

Use the data from example 1–9 to calculate the standard error and the 95% confidence interval for $\hat{p}$.

○ ANSWER

In that example $\hat{p} = 0.8$ and the sampling variance was 0.0016. The standard error is thus $\sqrt{0.0016} = 0.04$, and the 95% confidence interval is $[0.7216, 0.8784]$.

In this example, the sampling distribution of p should be normal with mean 0.8 and standard deviation 0.04. This distribution is shown as a solid line in figure 1.3. The bullets in the figure are remarkably close to the solid line, but they were not drawn using the normal approximation to the sampling distribution. They are simply copied from figure 1.2. They show that the normal distribution does a remarkable job of approximating the sampling distribution of $\hat{p}$.

1.6.4 How well does the normal theory work?

The sampling distribution of $\hat{p}$ is only approximately normal, but with moderately large samples the approximation is very good. To see just how good it is, let us reconsider the sampling distribution shown in figure 1.2 for the case in which $n = 100$. In those data $\hat{p} = 0.8$ and the sampling variance was $0.8 \times 0.2/100 = 0.0016$. The standard error is thus $\sqrt{0.0016} = 0.04$. The normal distribution with these parameters is shown as a solid line in figure 1.3. The bullets in the figure are remarkably close to the solid line, but they were not drawn using the normal approximation to the sampling distribution. They were simply copied from figure 1.2. They show that the normal distribution does a remarkable job of approximating the sampling distribution of $\hat{p}$, at least in this case. If n had been smaller, the agreement would not have been as good. The method of maximum likelihood works best with large samples.

1.7 Maximum likelihood exercises with genetics problems

$\star$ EXERCISE **1–2** Suppose that we have data from a genetic system with two alleles, and that we observe N_1 individuals of genotype A_1A_1, N_2 of genotype A_1A_2, and N_3 of genotype A_2A_2. If the (unknown) genotype frequencies are P_1, P_2, and P_3, then the likelihood function is

$$L \propto P_1^{N_1} P_2^{N_2} (1 - P_1 - P_2)^{N_3}$$

I have used the symbol "$\propto$" (which stands for "is proportional to") rather than the equals sign because this expression ignores a proportional constant that will not affect the answer. The log likelihood is

$$\ln L = \text{const.} + N_1 \ln P_1 + N_2 \ln P_2 + N_3 \ln(1 - P_1 - P_2)$$

Find the values of P_1, P_2, and P_3 that maximize the likelihood.

$\star$ EXERCISE **1–3** If we assume that the population is in Hardy-Weinberg equilibrium then the likelihood and log likelihood functions are

$$\begin{aligned} L &\propto [p^2]^{N_1} [2p(1-p)]^{N_2} [(1-p)^2]^{N_3} \\ \ln L &= \text{const.} + (2N_1 + N_2) \ln p + (N_2 + 2N_3) \ln(1-p) \end{aligned}$$

Find the value of p that maximizes the likelihood.

1.8 Maximum likelihood exercises with the Smith-Fretwell model of clutch size

The Smith-Fretwell model assumes that a child survives with a probability that increases with the resources that are available to it. To use the model with data, we would need to estimate the shape of this function. I have used a computer program to simulate a set of data that will be used in the exercises below. The data are available on the web at

```
http://www.anthro.utah.edu/˜rogers/ant4471
```

The data are in Maple format, so you will not have to retype them. You can import them into Maple using the "read" statement. Once you do so, two variables will be defined. The variable x is a vector containing 100 values that represent the resources available to each of 100 children. The variable y also has 100 values, and indicates whether each child survived or died. $y[i] = 1$ if the ith child survived and equals 0 if that child died.[1]

[1] For a real data set, see [1, table 4 and figure 6].

To estimate parameters, we need a parametric description of this relationship. To find one, we have to guess. Having guessed the form of the relationship, we can then use Maximum Likelihood to fit parameters and determine whether the guess was a good one. Let us begin by guessing that $s(x)$ has the following form:

$$s(x, a) = (1 - \exp(-ax))^4 \tag{1.8}$$

With real data, one is never sure that the guess is a good one. But these data were simulated, so I can assure you that this is the right formula to use.

To use the method of Maximum Likelihood, we need a likelihood function. If $y_i = 1$, then the likelihood of the observation on the ith chick is $L_i = s(x_i, a)$. On the other hand, if $y_i = 0$ then $L_i = 1 - s(x_i, a)$. In general,

$$L_i = y_i s(x_i, a) + (1 - y_i)(1 - s(x_i, a))$$

For the data set as a whole, the likelihood is

$$L = \prod_{i=1}^{100} L_i$$

The log likelihood is

$$\ln L = \sum_{i=1}^{100} \ln L_i$$

⋆ EXERCISE **1–4** Plot $s(x, a)$ with various values of a. Describe the function's shape and how the parameter affects its shape.

⋆ EXERCISE **1–5** Find the value of a that maximizes the likelihood (or the log likelihood), using any method you please. (If one method fails, try another. Make sure that you have found a local maximum rather than a local minimum.)

⋆ EXERCISE **1–6** Use the second derivative of $\ln L$ to estimate the sampling variance and standard error of $\hat{a}$. Then use the standard error to compute a 95 percent confidence interval.

⋆ EXERCISE **1–7** Use the Smith-Fretwell result, together with your estimate of a, to predict the optimal allocation of resources to each offspring.

Bibliography

[1] C.M. Perrins. Population fluctuations and clutch-size in the Great Tit. *Journal of Animal Ecology*, 34:601–647, 1965.

Appendix B

Answers to Exercises

⋆ EXERCISE **1–1** $\binom{3}{1} = 3$, $\binom{3}{2} = 3$, and $\binom{3}{3} = 1$.
⋆ EXERCISE **1–2** The parameter values that maximize this expression are

$$\hat{P}_1 = \frac{N_1}{N_1 + N_2 + N_3} \tag{B.1}$$

$$\hat{P}_2 = \frac{N_2}{N_1 + N_2 + N_3} \tag{B.2}$$

⋆ EXERCISE **1–4** $s(x, a)$ increases with x. When x is small, s increases at an increasing rate—its derivative increases. When x is larger, s increases at a decreasing rate, eventually reaching an asymptote where $s = 1$. The large the value of a, the more rapidly does s reach this asymptote.
⋆ EXERCISE **1–5** The derivative of $\ln L$ with respect to a is zero when $\hat{a} = 1.9141$. We can be sure that this is a local maximum because the second derivative is a negative number:

$$\left. \frac{d^2 \ln L}{da^2} \right|_{a=\hat{a}} = -45.38.$$

In simulating these data, I set $a = 2$. Thus, the maximum likelihood value is not far from the true value.
⋆ EXERCISE **1–6** The sampling variance is

$$-1 \bigg/ \left. \frac{d^2 \ln L}{da^2} \right|_{a=\hat{a}} = 1/45.38 = 0.022$$

The standard error is the square root of this,

$$\texttt{se} = \sqrt{0.022} = 0.1484$$

The 95 percent confidence interval is

$$[\hat{a} - 1.96\texttt{se}, \hat{a} - 1.96\texttt{se}]$$

⋆ EXERCISE **1–7** The optimal allocation is $\hat{x} = 1.221$ units of resource per child.