21

Maximum likelihood

In previous chapters we could easily construct estimators for various parameters of interest because these parameters had a natural sample analogue: expectation versus sample mean, probabilities versus relative frequencies, etc. However, in some situations such an analogue does not exist. In this chapter, a general principle to construct estimators is introduced, the so-called *maximum likelihood principle*. *Maximum likelihood estimators* have certain attractive properties that are discussed in the last section.

21.1 Why a general principle?

In Section 4.4 we modeled the number of cycles up to pregnancy by a random variable X with a geometric distribution with (unknown) parameter p. Weinberg and Gladen studied the effect of smoking on the number of cycles and obtained the data in Table 21.1 for 100 smokers and 486 nonsmokers.

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

Is the parameter p, which equals the probability of becoming pregnant after one cycle, different for smokers and nonsmokers? Let us try to find out by estimating p in the two cases.

What would be reasonable ways to estimate p? Since $p = \mathrm{P}(X = 1)$, the law of large numbers (see Section 13.3) motivates use of

$$S = \frac{\text{number of } X_i \text{ equal to } 1}{n}$$

as an estimator for p. This yields estimates $p = 29/100 = 0.29$ for smokers and $p = 198/486 = 0.41$ for nonsmokers. We know from Section 19.4 that S is an unbiased estimator for p. However, one cannot escape the feeling that S is a "bad" estimator: S does not use all the information in the table, i.e., the way the women are distributed over the numbers $2, 3, \ldots$ of observed numbers of cycles is not used. One would like to have an estimator that incorporates all the available information. Due to the way the data are given, this seems to be difficult. For instance, estimators based on the average cannot be evaluated, because 7 smokers and 12 nonsmokers had an unknown number of cycles up to pregnancy (larger than 12). If one simply ignores the last column in Table 21.1 as we did in Exercise 17.5, the average can be computed and yields $1/\bar{x}_{93} = 0.2809$ as an estimate of p for smokers and $1/\bar{x}_{474} = 0.3688$ for nonsmokers. However, because we discard seven values larger than 12 in case of the smokers and twelve values larger than 12 in case of the nonsmokers, we overestimate p in both cases.

In the next section we introduce a general principle to find an estimate for a parameter of interest, the *maximum likelihood principle*. This principle yields good estimators and will solve problems such as those stated earlier.

21.2 The maximum likelihood principle

Suppose a dealer of computer chips is offered on the black market two batches of 10 000 chips each. According to the seller, in one batch about 50% of the chips are defective, while this percentage is about 10% in the other batch. Our dealer is only interested in this last batch. Unfortunately the seller cannot tell the two batches apart. To help him to make up his mind, the seller offers our dealer one batch, from which he is allowed to select and test 10 chips. After selecting 10 chips arbitrarily, it turns out that only the second one is defective. Our dealer at once decides to buy this batch. Is this a wise decision?

With the batch where 50% of the chips are defective it is *more likely* that defective chips will appear, whereas with the other batch one would expect hardly any defective chip. Clearly, our dealer chooses the batch for which it is *most likely* that only one chip is defective. This is also the guiding idea behind the maximum likelihood principle.

> The maximum likelihood principle. Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

Set $R_i = 1$ in case the ith tested chip was defective and $R_i = 0$ in case it was operational, where $i = 1, \ldots, 10$. Then $R_1, \ldots, R_{10}$ are ten independent $Ber(p)$ distributed random variables, where p is the probability that a randomly selected chip is defective. The probability that the observed data occur is equal to

$$\mathrm{P}(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = p(1-p)^9.$$

For the batch where about 10% of the chips are defective we find that

$$\mathrm{P}(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = \frac{1}{10}\left(\frac{9}{10}\right)^9 = 0.039,$$

whereas for the other batch

$$\mathrm{P}(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = \frac{1}{2}\left(\frac{1}{2}\right)^9 = 0.00098.$$

So the probability for the batch with only 10% defective chips is about 40 times larger than the probability for the other batch. Given the data, our dealer made a sound decision.

Quick exercise 21.1 Which batch should the dealer choose if only the first three chips are defective?

Returning to the example of the number of cycles up to pregnancy, denoting X_i as the number of cycles up to pregnancy of the ith smoker, recall that

$$\mathrm{P}(X_i = k) = (1-p)^{k-1}p$$

and

$$\mathrm{P}(X_i > 12) = \mathrm{P}(\text{no success in cycle 1 to 12}) = (1-p)^{12};$$

cf. Quick exercise 4.6. From Table 21.1 we see that there are 29 smokers for which $X_i = 1$, that there are 16 for which $X_i = 2$, etc. Since we model the data as a random sample from a geometric distribution, the probability of the data—as a function of p—is given by

$$\begin{aligned} L(p) &= C \cdot \mathrm{P}(X_i = 1)^{29} \cdot \mathrm{P}(X_i = 2)^{16} \cdots \mathrm{P}(X_i = 12)^3 \cdot \mathrm{P}(X_i > 12)^7 \\ &= C \cdot p^{29} \cdot ((1-p)p)^{16} \cdots \left((1-p)^{11}p\right)^3 \cdot \left((1-p)^{12}\right)^7 \\ &= C \cdot p^{93} \cdot (1-p)^{322}. \end{aligned}$$

Here C is the number of ways we can assign 29 ones, 16 twos, …, 3 twelves, and 7 numbers larger than 12 to 100 smokers.[1] According to the *maximum likelihood principle* we now choose p, with $0 \leq p \leq 1$, in such a way, that $L(p)$

[1] $C = 311657028822819441451842682167854800096263625208359116504431153487280760832000000000$.

is maximal. Since C does not depend on p, we do not need to know the value of C explicitly to find for which p the function $L(p)$ is maximal.

Differentiating $L(p)$ with respect to p yields that

$$\begin{aligned} L'(p) &= C\left[93p^{92}(1-p)^{322} - 322p^{93}(1-p)^{321}\right] \\ &= Cp^{92}(1-p)^{321}\left[93(1-p) - 322p\right] \\ &= Cp^{92}(1-p)^{321}(93 - 415p). \end{aligned}$$

Now $L'(p) = 0$ if $p = 0$, $p = 1$, or $p = 93/415 = 0.224$, and $L(p)$ attains its unique maximum in this last point (check this!). We say that $93/415 = 0.224$ is the *maximum likelihood estimate* of p for the smokers. Note that this estimate is quite a lot smaller than the estimate 0.29 for the smokers we found in the previous section, and the estimate 0.2809 you obtained in Exercise 17.5.

Quick exercise 21.2 Check that for the nonsmokers the probability of the data is given by

$$L(p) = \text{constant} \cdot p^{474}(1-p)^{955}.$$

Compute the maximum likelihood estimate for p.

Remark 21.1 (Some history). The method of maximum likelihood estimation was propounded by Ronald Aylmer Fisher in a highly influential paper. In fact, this paper does not contain the original statement of the method, which was published by Fisher in 1912 [9], nor does it contain the original definition of *likelihood*, which appeared in 1921 (see [10]). The roots of the maximum likelihood method date back as far as 1713, when Jacob Bernoulli's *Ars Conjectandi* ([1]) was posthumously published. In the eighteenth century other important contributions were by Daniel Bernoulli, Lambert, and Lagrange (see also [2], [16], and [17]). It is interesting to remark that another giant of statistics, Karl Pearson, had not understood Fisher's method. Fisher was hurt by Pearson's lack of understanding, which eventually led to a violent confrontation.

21.3 Likelihood and loglikelihood

Suppose we have a dataset $x_1, x_2, \ldots, x_n$, modeled as a realization of a random sample from a distribution characterized by a parameter θ. To stress the dependence of the distribution on θ, we write

$$p_\theta(x)$$

for the probability mass function in case we have a sample from a discrete distribution and

$$f_\theta(x)$$

for the probability density function when we have a sample from a continuous distribution.

For a dataset $x_1, x_2, \ldots, x_n$ modeled as the realization of a random sample $X_1, \ldots, X_n$ from a *discrete* distribution, the maximum likelihood principle now tells us to estimate θ by that value, for which the function $L(\theta)$, given by

$$L(\theta) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = p_\theta(x_1) \cdots p_\theta(x_n)$$

is maximal. This value is called the maximum likelihood estimate of θ. The function $L(\theta)$ is called the *likelihood function*. This is a function of θ, determined by the numbers $x_1, x_2, \ldots, x_n$.

In case the sample is from a *continuous* distribution we clearly need to define the likelihood function $L(\theta)$ in a way different from the discrete case (if we would define $L(\theta)$ as in the discrete case, one always would have that $L(\theta) = 0$). For a reasonable definition of the likelihood function we have the following motivation. Let f_θ be the probability density function of X, and let $\varepsilon > 0$ be some fixed, small number. It is sensible to choose θ in such a way, that the probability $\mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon, \ldots, x_n - \varepsilon \le X_n \le x_n + \varepsilon)$ is maximal. Since the X_i are independent, we find that

$$\begin{aligned} &\mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon, \ldots, x_n - \varepsilon \le X_n \le x_n + \varepsilon) \\ &\quad = \mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon) \cdots \mathrm{P}(x_n - \varepsilon \le X_n \le x_n + \varepsilon) \\ &\quad \approx f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) (2\varepsilon)^n, \end{aligned} \tag{21.1}$$

where in the last step we used that (see also Equation (5.1))

$$\mathrm{P}(x_i - \varepsilon \le X_i \le x_i + \varepsilon) = \int_{x_i - \varepsilon}^{x_i + \varepsilon} f_\theta(x)\,\mathrm{d}x \approx 2\varepsilon f_\theta(x_i).$$

Note that the right-hand side of (21.1) is maximal whenever the function $f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$ is maximal, irrespective of the value of ε. In view of this, given a dataset $x_1, x_2, \ldots, x_n$, the likelihood function $L(\theta)$ is defined by

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$$

in the continuous case.

MAXIMUM LIKELIHOOD ESTIMATES. The *maximum likelihood estimate* of θ is the value $t = h(x_1, x_2, \ldots, x_n)$ that maximizes the likelihood function $L(\theta)$. The corresponding random variable

$$T = h(X_1, X_2, \ldots, X_n)$$

is called the *maximum likelihood estimator* for θ.

As an example, suppose we have a dataset $x_1, x_2, \ldots, x_n$ modeled as a realization of a random sample from an $Exp(\lambda)$ distribution, with probability density function given by $f_\lambda(x) = 0$ if $x < 0$ and

$$f_\lambda(x) = \lambda \mathrm{e}^{-\lambda x} \qquad \text{for} \quad x \geq 0.$$

Then the likelihood is given by

$$\begin{aligned} L(\lambda) &= f_\lambda(x_1) f_\lambda(x_2) \cdots f_\lambda(x_n) \\ &= \lambda \mathrm{e}^{-\lambda x_1} \cdot \lambda \mathrm{e}^{-\lambda x_2} \cdots \lambda \mathrm{e}^{-\lambda x_n} \\ &= \lambda^n \cdot \mathrm{e}^{-\lambda(x_1 + x_2 + \cdots + x_n)}. \end{aligned}$$

To obtain the maximum likelihood estimate of λ it is enough to find the maximum of $L(\lambda)$. To do so, we determine the derivative of $L(\lambda)$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\lambda} L(\lambda) &= n\lambda^{n-1} \mathrm{e}^{-\lambda \sum_{i=1}^n x_i} - \lambda^n \left(\sum_{i=1}^n x_i \right) \mathrm{e}^{-\lambda \sum_{i=1}^n x_i} \\ &= n \left(\lambda^{n-1} \mathrm{e}^{-\lambda \sum_{i=1}^n x_i} \left(1 - \frac{\lambda}{n} \sum_{i=1}^n x_i \right) \right). \end{aligned}$$

We see that $\mathrm{d}\left(L(\lambda)\right)/\mathrm{d}\lambda = 0$ if and only if

$$1 - \lambda \bar{x}_n = 0,$$

i.e., if $\lambda = 1/\bar{x}_n$. Check that for this value of λ the likelihood function $L(\lambda)$ attains a maximum! So the maximum likelihood estimator for λ is $1/\bar{X}_n$.

In the example of the number of cycles up to pregnancy of smoking women, we have seen that $L(p) = C \cdot p^{93} \cdot (1-p)^{322}$. The maximum likelihood estimate of p was found by differentiating $L(p)$. Differentiating is not always possible, as the following example shows.

Estimating the upper endpoint of a uniform distribution

Suppose the dataset $x_1 = 0.98$, $x_2 = 1.57$, and $x_3 = 0.31$ is the realization of a random sample from a $U(0, \theta)$ distribution with $\theta > 0$ unknown. The probability density function of each X_i is now given by $f_\theta(x) = 0$ if x is not in $[0, \theta]$ and

$$f_\theta(x) = \frac{1}{\theta} \qquad \text{for} \quad 0 \leq x \leq \theta.$$

The likelihood $L(\theta)$ is zero if θ is smaller than at least one of the x_i, and equals $1/\theta^3$ if θ is greater than or equal to each of the three x_i, i.e.,

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) f_\theta(x_3) = \begin{cases} \frac{1}{\theta^3} & \text{if} \quad \theta \geq \max(x_1, x_2, x_3) = 1.57 \\ 0 & \text{if} \quad \theta < \max(x_1, x_2, x_3) = 1.57. \end{cases}$$

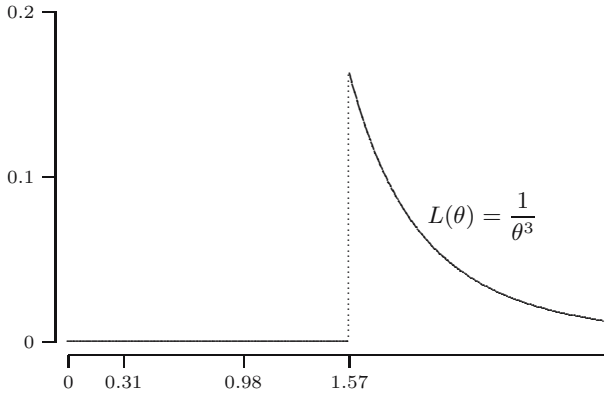

Fig. 21.1. Likelihood function $L(\theta)$ of a sample from a $U(0,\theta)$ distribution.

Figure 21.1 depicts this likelihood function. One glance at this figure is enough to realize that $L(\theta)$ attains its maximum at $\max(x_1, x_2, x_3) = 1.57$.

In general, given a dataset $x_1, x_2, \ldots, x_n$ originating from a $U(0,\theta)$ distribution, we see that $L(\theta) = 0$ if θ is smaller than at least one of the x_i and that $L(\theta) = 1/\theta^n$ if θ is greater than or equal to the largest of the x_i. We conclude that the maximum likelihood estimator of θ is given by $\max\{X_1, X_2, \ldots, X_n\}$.

Loglikelihood

In the preceding example it was easy to find the value of the parameter for which the likelihood is maximal. Usually one can find the maximum by differentiating the likelihood function $L(\theta)$. The calculation of the derivative of $L(\theta)$ may be tedious, because $L(\theta)$ is a product of terms, all involving θ (see also Quick exercise 21.3). To differentiate $L(\theta)$ we have to apply the product rule from calculus. Considering the logarithm of $L(\theta)$ changes the product of the terms involving θ into a *sum* of logarithms of these terms, which makes the process of differentiating easier. Moreover, because the logarithm is an increasing function, the likelihood function $L(\theta)$ and the *loglikelihood function* $\ell(\theta)$, defined by

$$\ell(\theta) = \ln(L(\theta)),$$

attain their extreme values for the same values of θ. In particular, $L(\theta)$ is maximal if and only if $\ell(\theta)$ is maximal. This is illustrated in Figure 21.2 by the likelihood function $L(p) = Cp^{93}(1-p)^{322}$ and the loglikelihood function $\ell(p) = \ln(C) + 93\ln(p) + 322\ln(1-p)$ for the smokers.

In the situation that we have a dataset $x_1, x_2, \ldots, x_n$ modeled as a realization of a random sample from an $Exp(\lambda)$ distribution, we found as likelihood function $L(\lambda) = \lambda^n \cdot \mathrm{e}^{-\lambda(x_1+x_2+\cdots+x_n)}$. Therefore, the loglikelihood function is given by

$$\ell(\lambda) = n\ln(\lambda) - \lambda\,(x_1 + x_2 + \cdots + x_n)\,.$$

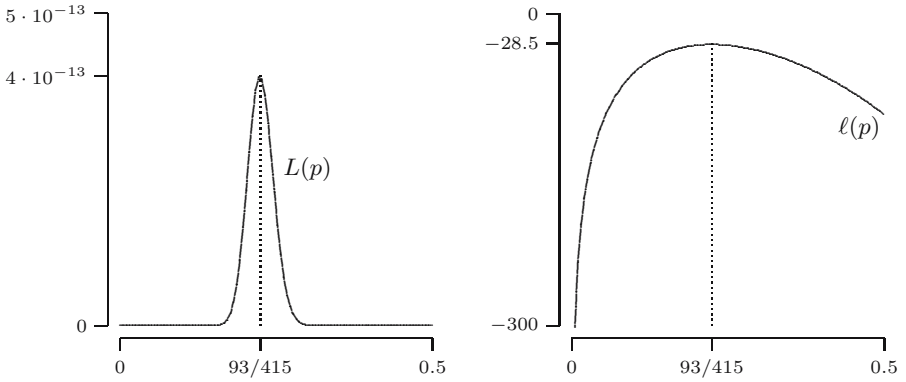

Fig. 21.2. The graphs of the likelihood function $L(p)$ and the loglikelihood function $\ell(p)$ for the smokers.

QUICK EXERCISE 21.3 In this example, use the loglikelihood function $\ell(\lambda)$ to show that the maximum likelihood estimate of λ equals $1/\bar{x}_n$.

Estimating the parameters of the normal distribution

Suppose that the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample from an $N(\mu, \sigma^2)$ distribution, with μ and σ unknown. What are the maximum likelihood estimates for μ and σ?
In this case θ is the vector (μ, σ), and therefore the likelihood function is a function of two variables:

$$L(\mu, \sigma) = f_{\mu,\sigma}(x_1) f_{\mu,\sigma}(x_2) \cdots f_{\mu,\sigma}(x_n),$$

where each $f_{\mu,\sigma}(x)$ is the $N(\mu, \sigma^2)$ probability density function:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Since

$$\ln\left(f_{\mu,\sigma}(x)\right) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2,$$

one finds that

$$\begin{aligned}\ell(\mu, \sigma) &= \ln\left(f_{\mu,\sigma}(x_1)\right) + \cdots + \ln\left(f_{\mu,\sigma}(x_n)\right) \\ &= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left((x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right).\end{aligned}$$

The partial derivatives of ℓ are

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\left((x_1-\mu)+(x_2-\mu)+\cdots+(x_n-\mu)\right) = \frac{n}{\sigma^2}\left(\bar{x}_n-\mu\right)$$
$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\left((x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2\right)$$
$$= -\frac{n}{\sigma^3}\left(\sigma^2-\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Solving $\dfrac{\partial \ell}{\partial \mu} = 0$ and $\dfrac{\partial \ell}{\partial \sigma} = 0$ yields

$$\mu = \bar{x}_n \quad \text{and} \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}.$$

It is not hard to show that for these values of μ and σ the likelihood function $L(\mu,\sigma)$ attains a maximum. We find that $\bar{x}_n$ is the maximum likelihood estimate for μ and that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}$$

is the maximum likelihood estimate for σ.

21.4 Properties of maximum likelihood estimators

Apart from the fact that the maximum likelihood principle provides a general principle to construct estimators, one can also show that maximum likelihood estimators have several desirable properties.

Invariance principle

In the previous example, we saw that

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

is the maximum likelihood estimator for the parameter σ of an $N(\mu,\sigma^2)$ distribution. Does this imply that D_n^2 is the maximum likelihood estimator for σ^2? This is indeed the case! In general one can show that if T is the maximum likelihood estimator of a parameter θ and $g(\theta)$ is an invertible function of θ, then $g(T)$ is the maximum likelihood estimator for $g(\theta)$.

Asymptotic unbiasedness

The maximum likelihood estimator T may be biased. For example, because $D_n^2 = \frac{n-1}{n} S_n^2$, for the previously mentioned maximum likelihood estimator D_n^2 of the parameter σ^2 of an $N(\mu, \sigma^2)$ distribution, it follows from Section 19.4 that

$$\mathrm{E}\left[D_n^2\right] = \mathrm{E}\left[\frac{n-1}{n} S_n^2\right] = \frac{n-1}{n} \mathrm{E}\left[S_n^2\right] = \frac{n-1}{n}\,\sigma^2.$$

We see that D_n^2 is a biased estimator for σ^2, but also that as n goes to infinity, the expected value of D_n^2 converges to σ^2. This holds more generally. Under mild conditions on the distribution of the random variables X_i under consideration (see, e.g., [36]), one can show that asymptotically (that is, as the size n of the dataset goes to infinity) maximum likelihood estimators are unbiased. By this we mean that if $T_n = h(X_1, X_2, \ldots, X_n)$ is the maximum likelihood estimator for a parameter θ, then

$$\lim_{n \to \infty} \mathrm{E}[T_n] = \theta.$$

Asymptotic minimum variance

The variance of an unbiased estimator for a parameter θ is always larger than or equal to a certain positive number, known as the Cramér-Rao lower bound (see Remark 20.2). Again under mild conditions one can show that maximum likelihood estimators have asymptotically the smallest variance among unbiased estimators. That is, asymptotically the variance of the maximum likelihood estimator for a parameter θ attains the Cramér-Rao lower bound.

21.5 Solutions to the quick exercises

21.1 In the case that only the first three chips are defective, the probability that the observed data occur is equal to

$$\mathrm{P}(R_1 = 1, R_2 = 1, R_3 = 1, R_4 = 0, \ldots, R_{10} = 0) = p^3(1-p)^7.$$

For the batch where about 10% of the chips are defective we find that

$$\mathrm{P}(R_1 = 1, R_2 = 1, R_3 = 1, R_4 = 0, \ldots, R_{10} = 0) = \left(\frac{1}{10}\right)^3 \left(\frac{9}{10}\right)^7 = 0.00048,$$

whereas for the other batch this probability is equal to $\left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^7 = 0.00098$. So the probability for the batch with about 50% defective chips is about 2 times larger than the probability for the other batch. In view of this, it would be reasonable to choose the other batch, not the tested one.

21.2 From Table 21.1 we derive

$$\begin{aligned} L(p) &= \text{constant} \cdot \mathrm{P}(X_i = 1)^{198}\, \mathrm{P}(X_i = 2)^{107} \cdots \mathrm{P}(X_i = 12)^{6}\, \mathrm{P}(X_i > 12)^{12} \\ &= \text{constant} \cdot p^{198} \cdot \left[(1-p)p\right]^{107} \cdots \left[(1-p)^{11}p\right]^{6} \cdot \left[(1-p)^{12}\right]^{12} \\ &= \text{constant} \cdot p^{474} \cdot (1-p)^{955}. \end{aligned}$$

Here the constant is the number of ways we can assign 198 ones, 107 twos, ..., 6 twelves, and 12 numbers larger than 12 to 486 nonsmokers. Differentiating $L(p)$ with respect to p yields that

$$\begin{aligned} L'(p) &= \text{constant} \cdot \left[474p^{473}(1-p)^{955} - 955p^{474}(1-p)^{954}\right] \\ &= \text{constant} \cdot p^{473}(1-p)^{954}\left[474(1-p) - 955p\right] \\ &= \text{constant} \cdot p^{473}(1-p)^{954}(474 - 1429p). \end{aligned}$$

Now $L'(p) = 0$ if $p = 0$, $p = 1$, or $p = 474/1429 = 0.33$, and $L(p)$ attains its unique maximum in this last point.

21.3 The loglikelihood function $L(\lambda)$ has derivative

$$\ell'(\lambda) = \frac{n}{\lambda} - (x_1 + x_2 + \cdots + x_n) = n\left(\frac{1}{\lambda} - \bar{x}_n\right).$$

One finds that $\ell'(\lambda) = 0$ if and only if $\lambda = 1/\bar{x}_n$ and that this is a maximum. The maximum likelihood estimate for λ is therefore $1/\bar{x}_n$.

21.6 Exercises

21.1 ⊞ Consider the following situation. Suppose we have two fair dice, D_1 with 5 red sides and 1 white side and D_2 with 1 red side and 5 white sides. We pick one of the dice randomly, and throw it repeatedly until *red* comes up for the first time. With the same die this experiment is repeated two more times. Suppose the following happens:

First experiment:	first red appears in	3rd throw
Second experiment:	first red appears in	5th throw
Third experiment:	first red appears in	4th throw.

Show that for die D_1 this happens with probability $5.7424 \cdot 10^{-8}$, and for die D_2 the probability with which this happens is $8.9725 \cdot 10^{-4}$. Given these probabilities, which die do you think we picked?

21.2 ⊡ We throw an unfair coin repeatedly until heads comes up for the first time. We repeat this experiment three times (with the same coin) and obtain the following data:

First experiment:	heads first comes up in 3rd throw
Second experiment:	heads first comes up in 5th throw
Third experiment:	heads first comes up in 4th throw.

Let p be the probability that heads comes up in a throw with this coin. Determine the maximum likelihood estimate $\hat{p}$ of p.

21.3 In Exercise 17.4 we modeled the hits of London by flying bombs by a Poisson distribution with parameter μ.

a. Use the data from Exercise 17.4 to find the maximum likelihood estimate of μ.

b. Suppose the summarized data from Exercise 17.4 got corrupted in the following way:

Number of hits	0 or 1	2	3	4	5	6	7
Number of squares	440	93	35	7	0	0	1

Using this new data, what is the maximum likelihood estimate of μ?

21.4 ⊞ In Section 19.1, we considered the arrivals of packages at a network server, where we modeled the number of arrivals per minute by a $Pois(\mu)$ distribution. Let $x_1, x_2, \ldots, x_n$ be a realization of a random sample from a $Pois(\mu)$ distribution. We saw on page 286 that a natural estimate of the probability of zeros in the dataset is given by

$$\frac{\text{number of } x_i \text{ equal to zero}}{n}.$$

a. Show that the likelihood $L(\mu)$ is given by

$$L(\mu) = \frac{\mathrm{e}^{-n\mu}}{x_1! \cdots x_n!} \mu^{x_1+x_2+\cdots+x_n}.$$

b. Determine the loglikelihood $\ell(\mu)$ and the formula of the maximum likelihood estimate for μ.

c. What is the maximum likelihood estimate for the probability $\mathrm{e}^{-\mu}$ of zero arrivals?

21.5 ⊡ Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a normal distribution.

a. Let the probability density of this normal distribution be given by

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^2} \quad \text{for } -\infty < x < \infty.$$

Determine the maximum likelihood estimate for μ.

b. Now suppose that the density of this normal distribution is given by

$$f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2/\sigma^2} \quad \text{for } -\infty < x < \infty.$$

Determine the maximum likelihood estimate for σ.

21.6 Let $x_1, x_2, \ldots, x_n$ be a dataset that is a realization of a random sample from a distribution with probability density $f_\delta(x)$ given by

$$f_\delta(x) = \begin{cases} \mathrm{e}^{-(x-\delta)} & \text{for } x \geq \delta \\ 0 & \text{for } x < \delta. \end{cases}$$

a. Draw the likelihood $L(\delta)$.

b. Determine the maximum likelihood estimate for δ.

21.7 ⊡ Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a Rayleigh distribution, which is a continuous distribution with probability density function given by

$$f_\theta(x) = \frac{x}{\theta^2} \mathrm{e}^{-\frac{1}{2}x^2/\theta^2} \quad \text{for } x \geq 0.$$

In this case what is the maximum likelihood estimate for θ?

21.8 ⊞ (Exercises 19.7 and 20.7 continued) A certain type of plant can be divided into four types: starchy-green, starchy-white, sugary-green, and sugary-white. The following table lists the counts of the various types among 3839 leaves.

Type	Count
Starchy-green	1997
Sugary-white	32
Starchy-white	906
Sugary-green	904

Setting

$$X = \begin{cases} 1 & \text{if the observed leave is of type starchy-green} \\ 2 & \text{if the observed leave is of type sugary-white} \\ 3 & \text{if the observed leave is of type starchy-white} \\ 4 & \text{if the observed leave is of type sugary-green,} \end{cases}$$

the probability mass function p of X is given by

a	1	2	3	4
$p(a)$	$\frac{1}{4}(2+\theta)$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$

and $p(a) = 0$ for all other a. Here $0 < \theta < 1$ is an unknown parameter, which was estimated in Exercise 19.7. We want to find a maximum likelihood estimate of θ.

a. Use the data to find the likelihood $L(\theta)$ and the loglikelihood $\ell(\theta)$.

b. What is the maximum likelihood estimate of θ using the data from the preceding table?

c. Suppose that we have the counts of n different leaves: n_1 of type starchy-green, n_2 of type sugary-white, n_3 of type starchy-white, and n_4 of type sugary-green (so $n = n_1 + n_2 + n_3 + n_4$). Determine the general formula for the maximum likelihood estimate of θ.

21.9 ⊡ Let $x_1, x_2, \ldots, x_n$ be a dataset that is a realization of a random sample from a $U(\alpha, \beta)$ distribution (with α and β unknown, $\alpha < \beta$). Determine the maximum likelihood estimates for α and β.

21.10 Let $x_1, x_2, \ldots, x_n$ be a dataset, which is a realization of a random sample from a $Par(\alpha)$ distribution. What is the maximum likelihood estimate for α?

21.11 ⊞ In Exercise 4.13 we considered the situation where we have a box containing an unknown number—say N—of identical bolts. In order to get an idea of the size of N we introduced three random variables X, Y, and Z. Here we will use X and Y, and in the next exercise Z, to find maximum likelihood estimates of N.

a. Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a $Geo(1/N)$ distribution. Determine the maximum likelihood estimate for N.

b. Suppose that $y_1, y_2, \ldots, y_n$ is a dataset, which is a realization of a random sample from a discrete uniform distribution on $1, 2, \ldots, N$. Determine the maximum likelihood estimate for N.

21.12 (Exercise 21.11 continued.) Suppose that m bolts in the box were marked and then r bolts were selected from the box; Z is the number of marked bolts in the sample. (Recall that it was shown in Exercise 4.13 **c** that Z has a hypergeometric distribution, with parameters m, N, and r.) Suppose that k bolts in the sample were marked. Show that the likelihood $L(N)$ is given by

$$L(N) = \frac{\binom{m}{k}\binom{N-m}{r-k}}{\binom{N}{r}}.$$

Next show that $L(N)$ increases for $N < mr/k$ and decreases for $N > mr/k$, and conclude that mr/k is the maximum likelihood estimate for N.

21.13 Often one can model the times that customers arrive at a shop rather well by a Poisson process with (unknown) rate λ (customers/hour). On a certain day, one of the attendants noticed that between noon and 12.45 p.m.

two customers arrived, and another attendant noticed that on the same day one customer arrived between 12.15 and 1 p.m. Use the observations of the attendants to determine the maximum likelihood estimate of λ.

21.14 A very inexperienced archer shoots n times an arrow at a disc of (unknown) radius θ. The disc is hit every time, but at completely random places. Let $r_1, r_2, \ldots, r_n$ be the distances of the various hits to the center of the disc. Determine the maximum likelihood estimate for θ.

21.15 On January 28, 1986, the main fuel tank of the space shuttle *Challenger* exploded shortly after takeoff. Essential in this accident was the leakage of some of the six O-rings of the *Challenger*. In Section 1.4 the probability of failure of an O-ring is given by

$$p(t) = \frac{\mathrm{e}^{a+b\cdot t}}{1+\mathrm{e}^{a+b\cdot t}},$$

where t is the temperature at launch in degrees Fahrenheit. In Table 21.2 the temperature t (in °F, rounded to the nearest integer) and the number of failures N for 23 missions are given, ordered according to increasing temperatures. (See also Figure 1.3, where these data are graphically depicted.) Give the likelihood $L(a, b)$ and the loglikelihood $\ell(a, b)$.

Table 21.2. Space shuttle failure data of pre-*Challenger* missions.

t	53	57	58	63	66	67	67	67
N	2	1	1	1	0	0	0	0
t	68	69	70	70	70	70	72	73
N	0	0	0	0	1	1	0	0
t	75	75	76	76	78	79	81	
N	0	2	0	0	0	0	0	

21.16 In the 18th century Georges-Louis Leclerc, Comte de Buffon (1707–1788) found an amusing way to approximate the number π using probability theory and statistics. Buffon had the following idea: take a needle and a large sheet of paper, and draw horizontal lines that are a needle-length apart. Throw the needle a number of times (say n times) on the sheet, and count how often it hits one of the horizontal lines. Say this number is s_n, then s_n is the realization of a $Bin(n, p)$ distributed random variable S_n. Here p is the probability that the needle hits one of the horizontal lines. In Exercise 9.20 you found that $p = 2/\pi$. Show that

$$T = \frac{2n}{S_n}$$

is the maximum likelihood estimator for π.