Chapter One

The Bernoulli model

In this chapter and in Chapter 2, we will consider a data set recording the number of newborn girls and boys in the UK in 2004 and investigate whether the distribution of the sexes is even among newborn children. This question could be of interest to an economist thinking about the wider issue of incentives facing parents who are expecting a baby. Sometimes the incentives are so strong that parents take actions that actually change basic statistics like the sex ratio.

When analyzing such a question using econometrics, an important and basic distinction is between sample and population distributions. In short, the sample distribution describes the variation in a particular data set, whereas we imagine that the data are sampled from some population about which we would like to learn. This first chapter describes that distinction in more detail. Building on that basis, we formulate a model using a class of possible population distributions. The population distribution within this class, which is the one most likely to have generated the data, can then be found. In Chapter 2, we can then proceed to question whether the distribution of the sexes is indeed even.

1.1 SAMPLE AND POPULATION DISTRIBUTIONS

We start by looking at a simple demographic data set showing the number of newborn girls and boys in the UK in 2004. This allows us to consider the question whether the chance that a newborn child is a girl is 50%. By examining the frequency of the two different outcomes, we obtain a *sample distribution*. Subsequently, we will turn to the general population of newborn children from which the data set has been sampled, and establish the notion of a *population distribution*. The econometric tools will be developed with a view toward learning about this population distribution from a sample distribution.

1.1.1 Sample distributions

In 2004, the number of newborn children in the UK was 715996, see Office for National Statistics (2006). Of these, 367586 were boys and 348410 were girls. These data have come about by observing $n = 715996$ newborn children. This

i	sex	Y_i
1	boy	0
2	girl	1
$\vdots$	$\vdots$	$\vdots$
715996	boy	0

Table 1.1 Cross-sectional data set of the sex of children born in 2004

gives us a *cross-sectional* data set as illustrated in Table 1.1. The name cross-section data refers to its origins in surveys that sought to interview a cross section of society. In a convenient notation, we let $i = 1, \ldots, n$ be the child index, and for each child we introduce a *random variable* Y_i, which can take the numerical value 0 or 1 representing "boy"or "girl,"respectively. While the data set shows a particular set of outcomes, or observations, of the random variables $Y_1, \ldots, Y_n$, the econometric analysis will be based on a *model* for the possible variation in the random variables $Y_1, \ldots, Y_n$. As in this example, random variables always take numerical values.

To obtain an overview of a data set like that reported Table 1.1, the number of cases in each category would be counted, giving a summary as in Table 1.2. This reduction of the data, of course, corresponds to the actual data obtained from the Office of National Statistics.

Y_i	0	1
count	367586	348410

Table 1.2 Sex of newborn children in the UK in 2004

The magnitudes of the numbers in the cells in Table 1.2 depend on the numbers born in 2004. We can standardize by dividing each entry by the total number of newborn children, with the result shown in Table 1.3.

y	0	1
$\widehat{\mathsf{f}}(y)$	0.513	0.487

Table 1.3 Sample frequency of newborn boys and girls for 2004

Each cell of Table 1.3 then shows:

$$\widehat{\mathsf{f}}\,(y) = \text{“frequency of sex } y \text{ among } n = 715996 \text{ newborn children.”}$$

We say that Table 1.3 gives the frequency distribution of the random variables $Y_1, \ldots, Y_n$. There are two aspects of the notation $\widehat{\mathsf{f}}(y)$ that need explanation. First, the argument y of the function $\widehat{\mathsf{f}}$ represents the potential outcomes of child births, as opposed to the realization of a particular birth. Second, the function $\widehat{\mathsf{f}}$, said as "f-hat", is an *observed*, or *sample*, quantity, in that it is computed from the observations $Y_1, \ldots, Y_n$. The notation $\widehat{\mathsf{f}}$, rather than f, is used to emphasize the *sample* aspect, in contrast to the *population* quantities we will discuss later on.

The variables $Y_1, \ldots, Y_n$ (denoted Y_i in shorthand) take the values 0 or 1. That is, Y_i takes $J = 2$ distinct values for $j = 1, \ldots, J$. Thus, the sum of the cell values in Table 1.3 is unity:

$$\sum_{j=1}^{J} \widehat{\mathsf{f}}\,(y_j) = 1.$$

1.1.2 Population distributions

We will think of a sample distribution as a random realization from a population distribution. In the above example, the sample is all newborn children in the UK in 2004, whereas the population distribution is thought of as representing the biological causal mechanism that determines the sex of children. Thus, although the sample here is actually the population of all newborn children in the UK in 2004, the population from which that sample is drawn is a hypothetical one.

The notion of a population distribution can be made a little more concrete with a coin-flipping example. The outcome of a coin toss is determined by the coin and the way it is tossed. As a model of this, we imagine a symmetric coin is tossed fairly such that there is an equal chance of the outcome being heads or tails, so the probability of each is $1/2$. It is convenient to think in terms of a random variable X describing the outcome of this coin-flipping experiment, so X takes values 0 and 1 if the outcome is tails and heads, respectively. The distribution of the outcomes can be described in terms of an underlying probability measure, P say, giving rise to the imagined population frequencies:

$$\mathsf{f}\,(0) = \mathsf{P}\,(X = 0) = 1/2 \qquad \text{and} \qquad \mathsf{f}\,(1) = \mathsf{P}\,(X = 1) = 1/2.$$

Here f appears without a "hat" as it is a population quantity, and $\mathsf{P}(X = 0)$ is read as "the probability of the event $X = 0$". We think of the frequency f as related to the random variable X, although that aspect is suppressed in the notation. In contrast, the probability measure P is more generic. We could introduce a new random variable $Y = 1 - X$, which takes the value 0 and 1 for heads and tails, rather than tails and heads, and write $\mathsf{P}(Y = 1) = \mathsf{P}(X = 0) = 1/2$.

We can sample from this population distribution as many times as we want. If, for instance, we toss the coin $n = 27$ times, we may observe 12 heads, so the sample frequency of heads is $\widehat{\mathsf{f}}(1) = 12/27$. In fact, when sampling an odd number of times, we can never observe that $\widehat{\mathsf{f}}(1) = \mathsf{f}(1) = 1/2$. One important difference between $\mathsf{f}(x)$ and $\widehat{\mathsf{f}}(x)$ is that f is a deterministic function describing the distribution of possible outcomes for a random variable X, whereas $\widehat{\mathsf{f}}$ is a random function describing the observed frequency of the outcomes in a sample of random variables $X_1, \ldots, X_n$; another sample of tosses would lead to different values of $\widehat{\mathsf{f}}$, but not f.

1.2 DISTRIBUTION FUNCTIONS AND DENSITIES

We need a structured way of thinking about distributions in order to build appropriate models. From probability theory, we can use the concepts of distribution functions and densities.

1.2.1 Distribution functions and random variables

Distribution theory is centered around *cumulative distribution functions* or just *distribution functions*. This is the function $\mathsf{F}(x) = \mathsf{P}(X \leq x)$, which is well defined regardless of the type of random variable. For the coin example, the distribution function is plotted in panel (a) of Figure 1.1. It has the defining property of any distribution function: it starts at zero on the far left and increases toward unity, reaching one at the far right. The probability is zero of observing an outcome less than zero, jumps to $1/2$ at 0 (tails), then to unity at 1 (heads), and stays there.

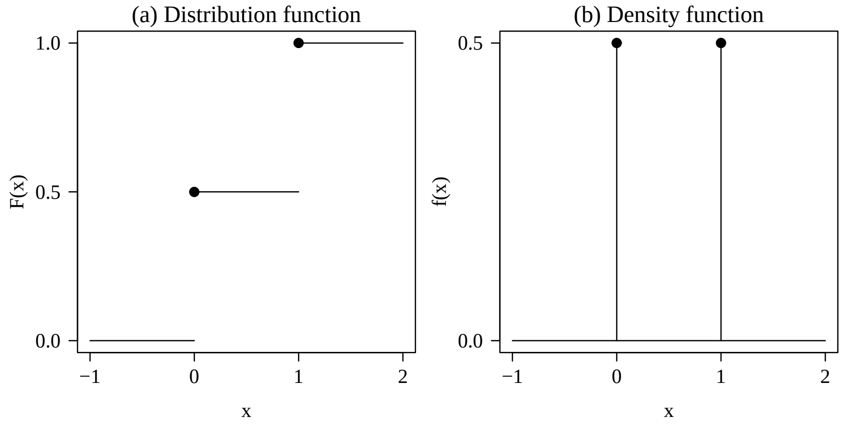

Figure 1.1 Distribution and density function for coin toss example

If we consider the inverse of the distribution function, we get the *quantiles* of a distribution. The 50% quantile, also called the median, is the smallest value of x such that $\mathsf{P}(X \leq x) = 0.5$. For the coin-flipping example considered in Figure 1.1, the median is 0.

When dealing with two random variables X and Y that could, for instance, describe the outcomes of two coin tosses, we have the *joint distribution function*:

$$\mathsf{F}(x, y) = \mathsf{P}(X \leq x \text{ and } Y \leq y).$$

We get the *marginal distribution function* of Y by allowing X to take any value:

$$\mathsf{F}(y) = \mathsf{P}(Y \leq y) = \mathsf{P}(X < \infty \text{ and } Y \leq y).$$

For example, when X and Y refer respectively to whether the mother is young/old and the child is boy/girl, then the marginal distribution of the sex of the child is so called because it refers to the distribution in the margin of the 2×2 table of possible outcomes irrespective of the mother's age: see Table 4.2 below.

If we flip a coin twice and let X and Y describe the two outcomes, we do not expect any influence between the two outcomes, and hence obtain:

$$\mathsf{P}(X \leq x \text{ and } Y \leq y) = \mathsf{P}(X \leq x)\mathsf{P}(Y \leq y). \tag{1.2.1}$$

If so, we say that the variables X and Y are *independent*. More generally, the variables $X_1, \ldots, X_n$ are said to be independent if their joint distribution function equals the product of the marginal distribution functions:

$$\mathsf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \leq x_i).$$

For example, the probability of 3 heads in a row is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$.

1.2.2 Density functions

In an econometric analysis, it is often convenient to consider the rate of increase of the distribution function rather than the distribution function itself.

For the birth and the coin-flipping experiments, the jumps of the distribution function determine the distribution uniquely. This is generally the case for any distribution function that is piecewise constant and therefore associated with a *discrete* distribution. The jumps are the *probability mass function* or the *density*:

$$\mathsf{f}(x) = \mathsf{P}(X = x).$$

Since the distribution function is piecewise constant, we can find the size of the jump at a point x as the value of the distribution function at x minus the value immediately before x, which we could write as:

$$\mathsf{f}(x) = \mathsf{P}(X = x) = \mathsf{P}(X \leq x) - \lim_{h \downarrow 0} \mathsf{P}(X \leq x - h).$$

As an example, compare the distribution function and the corresponding density shown in Figure 1.1. Here it is seen that the density is 0.5 when $x = 0$ or $x = 1$ and otherwise 0.

We can recover the distribution function from the density by summation:

$$\mathsf{F}(x) = \mathsf{P}(X \leq x) = \sum_{\substack{z:\text{ possible value} \\ \text{for } X \text{ so } z \leq x}} \mathsf{f}(z).$$

In particular, the sum over all possible outcomes is always unity.

We can also work with joint densities. It follows from (1.2.1) that two discrete random variables X, Y are independent if and only if:

$$\mathsf{f}(x, y) = \mathsf{P}(X = x \text{ and } Y = y) = \mathsf{P}(X = x)\mathsf{P}(Y = y) = \mathsf{f}(x)\,\mathsf{f}(y), \quad (1.2.2)$$

whereas the *marginal density* of X is:

$$\mathsf{f}(x) = \mathsf{P}(X = x) = \mathsf{P}(X = x \text{ and } Y \leq \infty) = \sum_{y\text{: possible values for } Y} \mathsf{f}(x, y), \quad (1.2.3)$$

where the sum is taken over all possible outcomes for Y.

1.2.3 A discrete distribution: the Bernoulli distribution

If a variable X takes the values 0 and 1, as with the variable for sex, it is said to be binary or dichotomous. It then has what is called a *Bernoulli* . The probability of a unit outcome:

$$\theta = \mathsf{P}(X = 1)$$

is the *success probability*. The parameter θ takes a value in the range $[0, 1]$. It follows that the probability of a *failure* is $\mathsf{P}(X = 0) = 1 - \theta$. In short, we write $X \overset{\mathsf{D}}{=} \mathsf{Bernoulli}[\theta]$ to indicate that X is Bernoulli-distributed with parameter θ. Figure 1.1 shows the distribution function and the density for a Bernoulli distribution with success parameter $\theta = 0.5$.

The density for the Bernoulli distribution can be written in a compact form:

$$\mathsf{f}(x) = \theta^{x}(1 - \theta)^{1-x} \qquad \text{for } x = 0, 1. \quad (1.2.4)$$

In (1.2.4), it holds that $\mathsf{P}(X = 0) = \mathsf{f}(0) = (1 - \theta)$ and $\mathsf{P}(X = 1) = \mathsf{f}(1) = \theta$.

1.3 THE BERNOULLI MODEL

We are now ready to develop our first statistical model. Using the above distribution theory, a statistical model and its associated likelihood function can be defined for the birth data. The likelihood function can then be used to find the specific member of the statistical model that is most likely to have generated the observed data.

1.3.1 A statistical model

Reconsider the birth data summarized in Table 1.2, where we are interested in learning about the population frequency of girls among newborn children. To do this, we will build a simple *statistical model* for the sex of newborn children. The objective is to make a good description of the sample distribution, which will eventually allow us to make inferences about, in this case, the frequency of girl births in the population of possible births. Here we will concentrate on describing the distribution with a view toward checking any assumptions we make.

Let Y_i denote the sex for child i, and consider n random variables $Y_1, \ldots, Y_n$ representing the data. We will make four assumptions:

(i) *independence*: $Y_1, \ldots, Y_n$ are mutually independent;
(ii) *identical distribution*: all children are drawn from the same population;
(iii) *Bernoulli distribution*: $Y_i \overset{\mathsf{D}}{=} \mathsf{Bernoulli}[\theta]$;
(iv) *parameter space*: $0 < \theta < 1$, which we write as $\theta \in \Theta = (0, 1)$.

We need to think about whether these assumptions are reasonable. Could they be so wrong that all inferences we draw from the model are misleading? In that case, we say the model is mis-specified. For example, the assumptions of independence or an identical distribution could well be wrong. In cases of identical twins, the independence assumption (i) is indeed not correct. Perhaps young and old mothers could have different chances of giving birth to girls, which could be seen as a violation of (ii). Is that something to worry about? In this situation, no more data are available, so we either have to stick to speculative arguments, turn to an expert in the field, or find more detailed data. We will proceed on the basis that any violations are not so large as to seriously distort our conclusions. Assumption (iii), however, is not in question in this model as the Bernoulli distribution is the only available distribution for binary data. Assumption (iv) is also not problematic here, even though the parameter space is actually restrictive in that it is chosen as $0 < \theta < 1$ as opposed to $0 \leq \theta \leq 1$. The resolution is that since we have observed both girls and boys, it is not possible that $\theta = 0$ or $\theta = 1$. These two points can therefore be excluded.

1.3.2 The likelihood function

Based on the statistical model, we can analyze how *probable* different outcomes $y_1, \ldots, y_n$ of $Y_1, \ldots, Y_n$ are for any given choice of the parameter θ. This is done by writing down the joint density of $Y_1, \ldots, Y_n$. Using the notation $\mathsf{f}_\theta(y_1, \ldots, y_n)$ for the joint density and the rules from §1.2.2, we get:

$$\begin{aligned} \mathsf{f}_\theta(y_1, \ldots, y_n) &= \prod_{i=1}^{n} \mathsf{f}_\theta(y_i) && [\ (i): \text{independence, see (1.2.2)}\ ] \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} && [\ (ii, iii): \text{Bernoulli, see (1.2.4)}\ ] \end{aligned}$$

This expression can be reduced further using the fact that:

$$\theta^a \theta^b = \theta^{a+b}, \tag{1.3.1}$$

which is the functional equation for power functions. Then (1.3.1) implies that:

$$\prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{\sum_{i=1}^{n}(1-y_i)}.$$

Introducing the notation $\overline{y}$ for the average $n^{-1}\sum_{i=1}^{n} y_i$, the joint density becomes:

$$\mathsf{f}_\theta(y_1,\ldots,y_n) = \theta^{n\overline{y}} (1-\theta)^{n(1-\overline{y})} = \left\{\theta^{\overline{y}} (1-\theta)^{(1-\overline{y})}\right\}^n. \qquad (1.3.2)$$

For known θ, we can calculate the density for any value of $\overline{y}$.

In practice, however, the premises are turned around: we have observed a data set that is a realization of the random variables $Y_1,\ldots,Y_n$, while θ is unknown. The aim is now to find the most *likely* value of θ for this particular outcome. To that end, we define the *likelihood function*:

$$\mathsf{L}_{Y_1,\ldots,Y_n}(\theta) = \mathsf{f}_\theta(Y_1,\ldots,Y_n), \qquad (1.3.3)$$

where the argument becomes the parameter θ varying in the parameter space Θ, rather than the possible data outcomes. Inserting (1.3.2) for the joint density, but expressed in terms of Y, we get:

$$\mathsf{L}_{Y_1,\ldots,Y_n}(\theta) = \left\{\theta^{\overline{Y}} (1-\theta)^{(1-\overline{Y})}\right\}^n. \qquad (1.3.4)$$

Two steps have been taken:

(1) y_i is replaced by Y_i to indicate that the likelihood function is based on the random variables representing the data;

(2) the expression (1.3.4) is viewed as a function of θ rather than Y_i.

Figure 1.2 illustrates the link between the joint density and the likelihood function. In panel (a), rather than showing the joint density as a function of its n-dimensional argument, it is shown as a function of $\overline{y}$. This is done for three different choices of θ. Panel (b) shows the corresponding likelihood function as a function of θ for $\overline{Y} = 0.487$. The three marked points indicate how the three different densities link up with the likelihood function. What would happen if the likelihood function were not raised to the power $1/n$ as in the figure?

Notice that the likelihood function depends only on the observations through $\overline{Y}$, so $\overline{Y}$ is said to be a *sufficient statistic* for θ. The summary statistics of Table 1.2 are therefore sufficient for our analysis, once it has been established that the model is not mis-specified.

1.3.3 Estimation

We will now seek to find the most likely parameter value by maximizing the likelihood function (1.3.4). The likelihood function has a product structure. Here, the

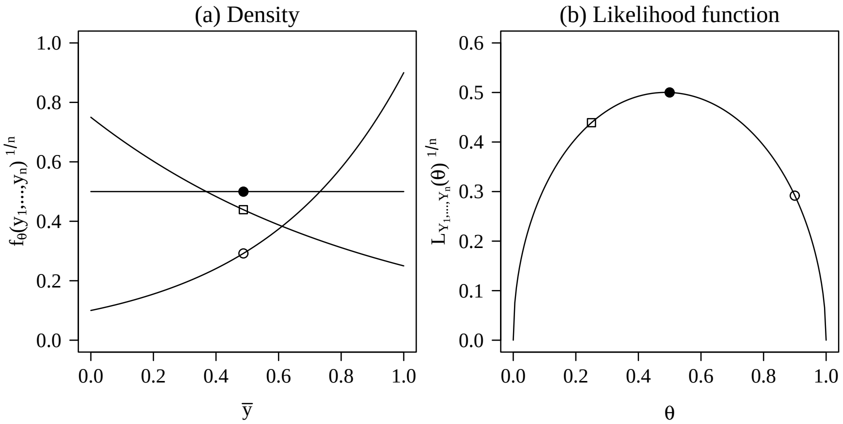

Figure 1.2 (a) Bernoulli densities as function of $\overline{y}$ given $\theta = 0.25,\ 0.5,\ 0.9$ for $\Box, \bullet, \circ$, respectively. (b) Bernoulli likelihood given $\overline{Y} = 0.487$, where the points $\Box, \bullet, \circ$ mark $\theta = 0.25,\ 0.5,\ 0.9$

product structure arises from the independence assumption, but we will later see, in Chapter 12, that such a structure can also arise in other ways. Since sums are much easier to deal with than products, it is convenient to linearize the likelihood function by applying the (natural) logarithm. The *log-likelihood function* is:

$$\ell_{Y_1,\ldots,Y_n}(\theta) \overset{\text{def}}{=} \log \mathsf{L}_{Y_1,\ldots,Y_n}(\theta) = \log\left[\left\{\theta^{\overline{Y}}(1-\theta)^{(1-\overline{Y})}\right\}^n\right].$$

Due to the monotonicity of the log transformation, the maxima of L and ℓ occur at the same value of the parameter. The above expression can be simplified by noting that the (natural) logarithm satisfies the functional equation:

$$\log(ab) = \log(a) + \log(b). \tag{1.3.5}$$

The log-likelihood function therefore reduces to:

$$\ell_{Y_1,\ldots,Y_n}(\theta) = \log \mathsf{L}_{Y_1,\ldots,Y_n}(\theta) = n\left\{\overline{Y}\log(\theta) + \left(1-\overline{Y}\right)\log(1-\theta)\right\}. \tag{1.3.6}$$

Figure 1.2(b) indicates that the likelihood function, and hence the log-likelihood function, has a unique maximum. To find this maximum, we differentiate with respect to θ:

$$\frac{\partial}{\partial\theta}\ell_{Y_1,\ldots,Y_n}(\theta) = n\left(\frac{\overline{Y}}{\theta} - \frac{1-\overline{Y}}{1-\theta}\right).$$

We set this expression equal to zero to find the value $\widehat{\theta}$ for which the likelihood takes its maximum:

$$n\left(\frac{\overline{Y}}{\widehat{\theta}} - \frac{1-\overline{Y}}{1-\widehat{\theta}}\right) = 0.$$

This first-order equation is called the *likelihood equation* for θ. Rearranging the likelihood equation:

$$\frac{\overline{Y}}{\widehat{\theta}} = \frac{1-\overline{Y}}{1-\widehat{\theta}} \qquad \Leftrightarrow \qquad \frac{\overline{Y}(1-\widehat{\theta}\,)}{\widehat{\theta}\,(1-\widehat{\theta}\,)} = \frac{(1-\overline{Y})\widehat{\theta}}{\widehat{\theta}\,(1-\widehat{\theta}\,)} \qquad \Leftrightarrow \qquad \widehat{\theta} = \overline{Y}.$$

Thus, $\widehat{\theta}$ is the value, among all possible parameter values θ, that maximizes the likelihood function. Thus, the *maximum likelihood estimator* for θ is:

$$\widehat{\theta} = \overline{Y}.$$

Once again, the hat over θ is used to indicate that $\widehat{\theta}$ is a function of the observed random variables, and hence the sample version of the parameter θ that is a population quantity. The maximum for the log-likelihood function is then:

$$\max_{0<\theta<1} \ell_{Y_1,\ldots,Y_n}(\theta) = \ell_{Y_1,\ldots,Y_n}\left(\widehat{\theta}\right) = n\left\{\overline{Y}\log\left(\widehat{\theta}\right) + (1-\overline{Y})\log\left(1-\widehat{\theta}\right)\right\}.$$

In the above analysis, $\widehat{\theta}$ was found to be a unique maximum by appealing to Figure 1.2(b). This can alternatively be proved by checking the second derivative of the log-likelihood function:

$$\frac{\partial^2}{\partial\theta^2}\ell_{Y_1,\ldots,Y_n}(\theta) = -n\left\{\frac{\overline{Y}}{\theta^2} + \frac{1-\overline{Y}}{(1-\theta)^2}\right\}.$$

Since $1 > \overline{Y} > 0$, this expression is negative for any value of $\theta \in \Theta$, so the log-likelihood function is concave with a unique maximum at $\widehat{\theta} = \overline{Y}$.

While an estimator is a random variable, a realization for a particular dataset is called an *estimate*. Thus, using the birth data, we estimate the chance of a newborn child being female by:

$$\widehat{\theta} = 0.4874 = 48.74\%, \tag{1.3.7}$$

while the log-likelihood function has its maximum value of:

$$\ell_{Y_1,\ldots,Y_n}(\widehat{\theta}\,) = -496033.8.$$

It is worth noting that these numbers are numerical approximations and subject to rounding errors, so that the estimate 0.4874 is only an approximation to the fraction of 348410 divided by 715996. In Chapter 2, we will see that this rounding error is small compared to the more important sampling error, so we choose to apply the equality symbol even when numbers are subject to rounding error.

We could reparametrize the model in terms of the proportion of boys, η say, satisfying $\eta = 1-\theta$. Going through everything above would deliver a new log-likelihood function $\widetilde{\ell}$ leading to the maximum likelihood estimator $\widehat{\eta} = 1-\overline{Y}$, taking the value 51.26% in this case. We immediately see that the maximum likelihood estimators from these two parametrizations satisfy $\widehat{\eta} = 1 - \widehat{\theta}$, and the two likelihood functions have the same maximum value:

$$\max_{0<\theta<1} \ell_{Y_1,\dots,Y_n}(\theta) = \max_{0<\eta<1} \widetilde{\ell}_{Y_1,\dots,Y_n}(\eta).$$

This is no coincidence, but a fundamental equivariance property of likelihood theory, namely that a likelihood function has the same maximum value for all, but very abstract, one-one parametrizations of the parameter space.

1.3.4 Restricting the statistical model

The motivation for this particular data analysis is to consider whether the chance that a newborn child is a girl could possibly be 50%. To do this, we can compare the values of the likelihood function at the unrestricted estimate, $\widehat{\theta}$, found above, and at the hypothesized point $\theta = 50\%$. We formalize this analysis as follows.

The analysis above represents an unrestricted model, where the likelihood function is maximized over an unrestricted parameter space, which we will now denote $\Theta_U = (0, 1)$, with maximum likelihood estimate $\widehat{\theta}_U = \overline{Y} = 48.74\%$. Our hypothesis is that $\theta = 50\%$, which restricts the parameter space to a single point $\Theta_R = \{0.5\}$. It is easy to maximize the likelihood function in the case of such a simple hypothesis, and we find:

$$\max_{\theta\in\Theta_R} \ell_{Y_1,\dots,Y_n}(\theta) = \ell_{Y_1,\dots,Y_n}(0.5) = -496290.6,$$

where, of course, the restricted maximum likelihood estimate is $\widehat{\theta}_R = 50\%$.

We now evaluate that restriction by comparing the relative likelihoods of the unrestricted maximum likelihood estimator $\widehat{\theta}_U$ and the restricted maximum likelihood estimator $\widehat{\theta}_R$, in terms of the ratio or quotient:

$$\mathsf{Q} = \frac{\max_{\theta\in\Theta_R} \mathsf{L}_{Y_1,\dots,Y_n}(\theta)}{\max_{\theta\in\Theta_U} \mathsf{L}_{Y_1,\dots,Y_n}(\theta)},$$

satisfying $0 \le \mathsf{Q} \le 1$. The closer Q is to unity, the more likely it is that θ could satisfy the restriction. In practice, we usually look at a transformation of Q called the *log-likelihood ratio test statistic* or simply the likelihood ratio test statistic:

$$\mathsf{LR} = -2\log\mathsf{Q} = 2\left\{\max_{\theta\in\Theta_U} \ell_{Y_1,\dots,Y_n}(\theta) - \max_{\theta\in\Theta_R} \ell_{Y_1,\dots,Y_n}(\theta)\right\}.$$

This takes non-negative values, so $\mathsf{LR} \ge 0$, and the closer LR is to zero, the more likely it is that θ could satisfy the restriction.

In our example we have:

$$\mathsf{LR} = 2(-496033.8 + 496290.6) = 513.6.$$

The crucial question is whether this is a large or a small value. In the following chapter, we will learn that there is an easy criterion for judging this. We will find that it is actually very large indeed, which in turn will lead us to reject the hypothesis that there is an equal chance that a newborn child is a boy or a girl.

1.4 SUMMARY AND EXERCISES

Sample quantities		Population quantities	
$\widehat{\mathsf{f}}$	frequency	f	density
$\widehat{\theta}$	estimator	θ	parameter

Table 1.4 Sample and population quantities

Summary: Sample and population quantities were introduced. It is fundamental to distinguish between those. The former are computed from the data. The latter are related to a postulated population. The notation distinguishes the two concepts by using a hat for sample quantities (Table 1.4).

We think of observations as outcomes of random variables. The probability theory is set up in terms of distribution functions. The density is the rate of increase of a distribution function. For likelihood theory, the multiplicative decomposition of joint densities in the independence case is crucial.

A statistical model specifies a parametrized class of distributions, one of which could have generated the data. This is in contrast with most other subjects where the notion of a model is reserved for a single distribution or other generating mechanism. For each value of the parameters, the joint density describes how probable outcomes are. Interpreted as a likelihood function, the joint density describes how likely parameters are. The maximum likelihood estimator is the most likely parameter value.

Bibliography: Many texts give detailed introductions to probability theory. A favorite choice is Hoel, Port and Stone (1971). Maximum likelihood was first suggested by Thiele (1886), see Lauritzen (2002) for an English translation, but the proposal did not take off at the time. The idea is usually attributed to R. A. Fisher, who, apparently unaware of Thiele's work, rediscovered the idea in Fisher (1922). Fisher understood the general applicability of likelihood and this, along with his many other contributions, revolutionized the way statistics was done.

In our notation, we have not formally distinguished between a random variable and a realization of a random variable computed from the data. This would require a little more probability theory than we need, and such a notation would actually become a hindrance later on in the book. For the same reason, our definition of the likelihood function in (1.3.3) differs from that in many classical texts on statistical theory, such as that of Cox and Hinkley (1974), and our notation is the same for estimators and estimates.

Key questions:

- What is the difference between sample and population distributions?
- Describe the notion of independence.
- How are joint densities and likelihood functions related?
- What is a statistical model?
- Discuss the validity of the assumptions of the statistical model for the newborn children data.

Exercise 1.1. *Let Y_1 and Y_2 be independent* $\mathsf{Bernoulli}[0.5]$*-distributed random variables. Find the possible outcomes, the density, and the distribution function for* $\overline{Y} = (Y_1 + Y_2)/2$.

Exercise 1.2. *Table 1.5 shows the number of newborn boys and girls in the UK in 2003 and 2004.*
(a) *Set up a Bernoulli model for the 2003 data and estimate the success parameter.*
(b) *Consider a joint model for the data for 2003 and 2004, where the success parameters can be different for the two years, and where all observations are independent. Argue that the joint likelihood is found by multiplying the two marginal likelihoods for 2003 and for 2004. How would you estimate the success parameters in this model?*

	boys	girls
2003	356578	338971
2004	367586	348410

Table 1.5 Sex of newborn children in the UK in 2003 and 2004. Sources: Office for National Statistics (2005, 2006)

Exercise 1.3. *Table 1.6 shows the number of newborn boys and girls in the US in 2002, measured in thousands. Set up a Bernoulli model for the data and estimate the success parameter.*

	boys	girls
2002	2058	1964

Table 1.6 Sex of newborn children in the US in 2002, measured in thousands. Source: Census Bureau (2005)

Exercise 1.4. * *Consider a pair of random variables* $X,\ Y$ *taking values* $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$ *with equal probability.*
(a) *Find the marginal distribution of* X.
(b) *Are* X *and* Y *independent?*