Chapter 7 ESTIMATION

7.1 Statistical Inference

Recall our various clinical trial examples. What would we say is the probability that a future patient will respond successfully to treatment after we observe the results from a collection of other patients? This is the kind of question that statistical inference is designed to address. In general, statistical inference consists of making probabilistic statements about unknown quantities. For example, we can compute means, variances, quantiles, probabilities, and some other quantities yet to be introduced concerning unobserved random variables and unknown parameters of distributions. Our goal will be to say what we have learned about the unknown quantities after observing some data that we believe contain relevant information. Here are some other examples of questions that statistical inference can try to answer. What can we say about whether a machine is functioning properly after we observe some of its output? In a civil lawsuit, what can we say about whether there was discrimination after observing how different ethnic groups were treated? The methods of statistical inference, which we shall develop to address these questions, are built upon the theory of probability covered in the earlier chapters of this text.

Probability and Statistical Models

In the earlier chapters of this book, we discussed the theory and methods of probability. As new concepts in probability were introduced, we also introduced examples of the use of these concepts in problems that we shall now recognize as *statistical inference*. Before discussing statistical inference formally, it is useful to remind ourselves of those probability concepts that will underlie inference.

Example 7.1.1 Lifetimes of Electronic Components. A company sells electronic components and they are interested in knowing as much as they can about how long each component is likely to last. They can collect data on components that have been used under typical conditions. They choose to use the family of exponential distributions to model the length of time (in years) from when a component is put into service until it fails. They would like to model the components as all having the same failure rate θ, but there is uncertainty about the specific numerical value of θ. To be more precise,

let $X_1, X_2, \ldots$ stand for a sequence of component lifetimes in years. The company believes that if they knew the failure rate θ, then $X_1, X_2, \ldots$ would be i.i.d. random variables having the exponential distribution with parameter θ. (See Sec. 5.7 for the definition of exponential distributions. We are using the symbol θ for the parameter of our exponential distributions rather than β to match the rest of the notation in this chapter.) Suppose that the data that the company will observe consist of the values of $X_1, \ldots, X_m$ but that they are still interested in $X_{m+1}, X_{m+2}, \ldots$. They are also interested in θ because it is related to the average lifetime. As we saw in Eq. (5.7.17), the mean of an exponential random variable with parameter θ is $1/\theta$, which is why the company thinks of θ as the failure rate.

We imagine an experiment whose outcomes are sequences of lifetimes as described above. As mentioned already, if we knew the value θ, then $X_1, X_2, \ldots$ would be i.i.d. random variables. In this case, the law of large numbers (Theorem 6.2.4) says that the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to the mean $1/\theta$. And Theorem 6.2.5 says that $n/\sum_{i=1}^{n} X_i$ converges in probability to θ. Because θ is a function of the sequence of lifetimes that constitute each experimental outcome, it can be treated as a random variable. Suppose that, before observing the data, the company believes that the failure rate is probably around 0.5/year but there is quite a bit of uncertainty about it. They model θ as a random variable having the gamma distribution with parameters 1 and 2. To rephrase what was stated earlier, they also model $X_1, X_2, \ldots$ as conditionally i.i.d. exponential random variables with parameter θ given θ. They hope to learn more about θ from examining the sample data $X_1, \ldots, X_m$. They can never learn θ precisely, because that would require observing the entire infinite sequence $X_1, X_2, \ldots$. For this reason, θ is only hypothetically observable. ◀

Example 7.1.1 illustrates several features that will be common to most statistical inference problems and which constitute what we call a statistical model.

Definition 7.1.1 Statistical Model. A *statistical model* consists of an identification of random variables of interest (both observable and only hypothetically observable), a specification of a joint distribution or a family of possible joint distributions for the observable random variables, the identification of any parameters of those distributions that are assumed unknown and possibly hypothetically observable, and (if desired) a specification for a (joint) distribution for the unknown parameter(s). When we treat the unknown parameter(s) θ as random, then the joint distribution of the observable random variables indexed by θ is understood as the conditional distribution of the observable random variables given θ.

In Example 7.1.1, the observable random variables of interest form the sequence $X_1, X_2, \ldots$, while the failure rate θ is hypothetically observable. The family of possible joint distributions of $X_1, X_2, \ldots$ is indexed by the parameter θ. The joint distribution of the observables corresponding to the value θ is that $X_1, X_2, \ldots$ are i.i.d. random variables each having the exponential distribution with parameter θ. This is also the conditional distribution of $X_1, X_2, \ldots$ given θ because we are treating θ as a radom variable. The distribution of θ is the gamma distribution with parameters 1 and 2.

Note: Redefining Old Ideas. The reader will notice that a statistical model is nothing more than a formal identification of many features that we have been using in various examples throughout the earlier chapters of this book. Some examples need only a few of the features that make up a complete specification of a statistical model, while other examples use the complete specification. In Sections 7.1–7.4, we shall

introduce a considerable amount of terminology, most of which is mere formalization of concepts that have been introduced and used in several places earlier in the book. The purpose of all of this formalism is to help us to keep the concepts organized so that we can tell when we are applying the same ideas in new ways and when we are introducing new ideas.

We are now ready formally to introduce statistical inference.

Definition 7.1.2 Statistical Inference. A *statistical inference* is a procedure that produces a probabilistic statement about some or all parts of a statistical model.

By a "probabilistic statement" we mean a statement that makes use of any of the concepts of probability theory that were discussed earlier in the text or are yet to be discussed later in the text. Some examples include a mean, a conditional mean, a quantile, a variance, a conditional distribution for a random variable given another, the probability of an event, a conditional probability of an event given something, and so on. In Example 7.1.1, here are some examples of statistical inferences that one might wish to make:

- Produce a random variable Y (a function of $X_1, \ldots, X_m$) such that $\Pr(Y \geq \theta|\theta) = 0.9$.
- Produce a random variable Y that we expect to be close to θ.
- Compute how likely it is that the average of the next 10 lifetimes, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$, is at least 2.
- Say something about how confident we are that $\theta \leq 0.4$ after observing $X_1, \ldots, X_m$.

All of these types of inference and others will be discussed in more detail later in this book.

In Definition 7.1.1, we distinguished between observable and hypothetically observable random variables. We reserved the name *observable* for a random variable that we are essentially certain that we could observe if we devoted the necessary effort to observe it. The name *hypothetically observable* was used for a random variable that would require infinite resources to observe, such as the limit (as $n \to \infty$) of the sample averages of the first n observables. In this text, such hypothetically observable random variables will correspond to the parameters of the joint distribution of the observables as in Example 7.1.1. Because these parameters figure so prominently in many of the types of inference problems that we will see, it pays to formalize the concept of parameter.

Definition 7.1.3 Parameter/Parameter space. In a problem of statistical inference, a characteristic or combination of characteristics that determine the joint distribution for the random variables of interest is called a *parameter* of the distribution. The set Ω of all possible values of a parameter θ or of a vector of parameters $(\theta_1, \ldots, \theta_k)$ is called the *parameter space*.

All of the families of distributions introduced earlier (and to be introduced later) in this book have parameters that are included in the names of the individual members of the family. For example, the family of binomial distributions has parameters that we called n and p, the family of normal distributions is parameterized by the mean μ and variance σ^2 of each distribution, the family of uniform distributions on intervals is parameterized by the endpoints of the intervals, the family of exponential distributions is parameterized by the rate parameter θ, and so on.

In Example 7.1.1, the parameter θ (the failure rate) must be positive. Therefore, unless certain positive values of θ can be explicitly ruled out as possible values of θ, the parameter space Ω will be the set of all positive numbers. As another example, suppose that the distribution of the heights of the individuals in a certain population is assumed to be the normal distribution with mean μ and variance σ^2, but that the exact values of μ and σ^2 are unknown. The mean μ and the variance σ^2 determine the particular normal distribution for the heights of individuals. So (μ, σ^2) can be considered a pair of parameters. In this example of heights, both μ and σ^2 must be positive. Therefore, the parameter space Ω can be taken as the set of all pairs (μ, σ^2) such that $\mu > 0$ and $\sigma^2 > 0$. If the normal distribution in this example represents the distribution of the heights in inches of the individuals in some particular population, we might be certain that $30 < \mu < 100$ and $\sigma^2 < 50$. In this case, the parameter space Ω could be taken as the smaller set of all pairs (μ, σ^2) such that $30 < \mu < 100$ and $0 < \sigma^2 < 50$.

The important feature of the parameter space Ω is that it must contain all possible values of the parameters in a given problem, in order that we can be certain that the actual value of the vector of parameters is a point in Ω.

Example 7.1.2

A Clinical Trial. Suppose that 40 patients are going to be given a treatment for a condition and that we will observe for each patient whether or not they recover from the condition. We are most likely also intersted in a large collection of additional patients besides the 40 to be observed. To be specific, for each patient $i = 1, 2, \ldots$, let $X_i = 1$ if patient i recovers, and let $X_i = 0$ if not. As a collection of possible distributions for $X_1, X_2, \ldots$, we could choose to say that the X_i are i.i.d. having the Bernoulli distribution with parameter p for $0 \le p \le 1$. In this case, the parameter p is known to lie in the closed interval $[0, 1]$, and this interval could be taken as the parameter space. Notice also that the law of large numbers (Theorem 6.2.4) says that p is the limit as n goes to infinity of the proportion of the first n patients who recover. ◀

In most problems, there is a natural interpretation for the parameter as a feature of the possible distributions of our data. In Example 7.1.2, the parameter p has a natural interpretation as the proportion out of a large population of patients given the treatment who recover from the condition. In Example 7.1.1, the parameter θ has a natural interpretation as a failure rate, that is, one over the average lifetime of a large population of lifetimes. In such cases, inference about parameters can be interpreted as inference about the feature that the parameter represents. In this text, all parameters will have such natural interpretations. In examples that one encounters outside of an introductory course, interpretations may not be as straightforward.

Examples of Statistical Inference

Here are some of the examples of statistical models and inferences that were introduced earlier in the text.

Example 7.1.3

A Clinical Trial. The clinical trial introduced in Example 2.1.4 was concerned with how likely patients are to avoid relapse while under various treatments. For each i, let $X_i = 1$ if patient i in the imipramine group avoids relapse and $X_i = 0$ otherwise. Let P stand for the proportion of patients who avoid relapse out of a large group receiving imipramine treatment. If P is unknown, we can model $X_1, X_2, \ldots$ as i.i.d.

Bernoulli random variables with parameter p conditional on $P = p$. The patients in the imipramine column of Table 2.1 should provide us with some information that changes our uncertainty about P. A statistical inference would consist of making a probability statement about the data and/or P, and what the data and P tell us about each other. For instance, in Example 4.7.8, we assumed that P had the uniform distribution on the interval $[0, 1]$, and we found the conditional distribution of P given the observed results of the study. We also computed the conditional mean of P given the study results as well as the M.S.E. for trying to predict P both before and after observing the results of the study. ◀

Example 7.1.4 Radioactive Particles. In Example 5.7.8, radioactive particles reach a target according to a Poisson process with unknown rate β. In Exercise 22 of Sec. 5.7, you were asked to find the conditional distribution of β after observing the Poisson process for a certain amount of time. ◀

Example 7.1.5 Anthropometry of Flea Beetles. In Example 5.10.2, we plotted two physical measurements from a sample of 31 flea beetles together with contours of a bivariate normal distribution. The family of bivariate normal distributions is parameterized by five quantities: the two means, the two variances, and the correlation. The choice of which set of five parameters to use for the fitted distribution is a form of statistical inference known as *estimation*. ◀

Example 7.1.6 Interval for Mean. Suppose that the heights of men in a certain population follow the normal distribution with mean μ and variance 9, as in Example 5.6.7. This time, assume that we do not know the value of the mean μ, but rather we wish to learn about it by sampling from the population. Suppose that we decide to sample $n = 36$ men and let $\overline{X}_n$ stand for the average of their heights. Then the interval $\left(\overline{X}_n - 0.98, \overline{X}_n + 0.98\right)$ computed in Example 5.6.8 has the property that it will contain the value of μ with probability 0.95. ◀

Example 7.1.7 Discrimination in Jury Selection. In Example 5.8.4, we were interested in whether there was evidence of discrimination against Mexican Americans in juror selection. Figure 5.8 shows how people who came into the case with different opinions about the extent of discrimination (if any) could alter their opinions in the light of learning the numerical evidence presented in the case. ◀

Example 7.1.8 Service Times in a Queue. Suppose that customers in a queue must wait for service, and that we get to observe the service times of several customers. Suppose that we are interested in the rate at which customers are served. In Example 5.7.3, we let Z stand for the service rate, and in Example 5.7.4, we showed how to find the conditional distribution of Z given several observed service times. ◀

General Classes of Inference Problems

Prediction One form of inference is to try to predict random variables that have not yet been observed. In Example 7.1.1, we might be interested in the average of the next 10 lifetimes, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$. In the clinical trial example (Example 7.1.3), we might be interested in predicting how many patients from the next set of patients in the imipramine group will have successful outcome. In virtually every statistical inference problem, in which we have not observed all of the relevant data, prediction

is possible. When the unobserved quantity to be predicted is a parameter, prediction is usually called *estimation,* as in Example 7.1.5.

Statistical Decision Problems In many statistical inference problems, after the experimental data have been analyzed, we must choose a decision from some available class of decisions with the property that the consequences of each available decision depend on the unknown value of some parameter. For example, we might have to estimate the unknown failure rate θ of our electronic components when the consequences depend on how close our estimate is to the correct value θ. As another example, we might have to decide whether the unknown proportion P of patients in the imipramine group (Example 7.1.3) is larger or smaller than some specified constant when the consequences depend on where P lies relative to the constant. This last type of inference is closely related to *hypothesis testing*, the subject of Chapter 9.

Experimental Design In some statistical inference problems, we have some control over the type or the amount of experimental data that will be collected. For example, consider an experiment to determine the mean tensile strength of a certain type of alloy as a function of the pressure and temperature at which the alloy is produced. Within the limits of certain budgetary and time constraints, it may be possible for the experimenter to choose the levels of pressure and temperature at which experimental specimens of the alloy are to be produced, and also to specify the number of specimens to be produced at each of these levels.

Such a problem, in which the experimenter can choose (at least to some extent) the particular experiment that is to be carried out, is called a problem of *experimental design*. Of course, the design of an experiment and the statistical analysis of the experimental data are closely related. One cannot design an effective experiment without considering the subsequent statistical analysis that is to be carried out on the data that will be obtained. And one cannot carry out a meaningful statistical analysis of experimental data without considering the particular type of experiment from which the data were derived.

Other Inferences The general classes of problems described above, as well as the more specific examples that appeared earlier, are intended as illustrations of types of statistical inferences that we will be able to perform with the theory and methods introduced in this text. The range of possible models, inferences, and methods that can arise when data are observed in real research problems far exceeds what we can introduce here. It is hoped that gaining an understanding of the problems that we can cover here will give the reader an appreciation for what needs to be done when a more challenging statistical problem arises.

Definition of a Statistic

Example 7.1.9

Failure Times of Ball Bearings. In Example 5.6.9, we had a sample of the numbers of millions of revolutions before failure for 23 ball bearings. We modeled the lifetimes as a random sample from a lognormal distribution. We might suppose that the parameters μ and σ^2 of that lognormal distribution are unknown and that we might wish to make some inference about them. We would want to make use of the 23 observed values in making any such inference. But do we need to keep track of all 23 values or are there some summaries of the data on which our inference will be based? ◀

Each statistical inference that we will learn how to perform in this book will be based on one or a few summaries of the available data. Such data summaries arise so often and are so fundamental to inference that they receive a special name.

Definition 7.1.4

Statistic. Suppose that the observable random variables of interest are $X_1, \ldots, X_n$. Let r be an arbitrary real-valued function of n real variables. Then the random variable $T = r(X_1, \ldots, X_n)$ is called a *statistic*.

Three examples of statistics are the sample mean $\overline{X}_n$, the maximum Y_n of the values of $X_1, \ldots, X_n$, and the function $r(X_1, \ldots, X_n)$, which has the constant value 3 for all values of $X_1, \ldots, X_n$.

Example 7.1.10

Failure Times of Ball Bearings. In Example 7.1.9, suppose that we were interested in making a statement about how far μ is from 40. Then we might want to use the statistic

$$T = \left| \frac{1}{36} \sum_{i=1}^{36} \log(X_i) - 4 \right|$$

in our inference procedure. In this case, T is a naïve measure of how far the data suggest that μ is from 40. ◀

Example 7.1.11

Interval for Mean. In Example 7.1.6, we constructed an interval that has probability 0.95 of containing μ. The endpoints of that interval, namely, $\overline{X}_n - 0.98$ and $\overline{X}_n + 0.98$, are statistics. ◀

Many inferences can proceed without explicitly constructing statistics as a preliminary step. However, most inferences will involve the use of statistics that could be identified in advance. And knowing which statistics are useful in which inferences can greatly simplify the implementation of the inference. Expressing an inference in terms of statistics can also help us to decide how well the inference meets out needs. For instance, in Example 7.1.10, if we estimate $|\mu - 40|$ by T, we can use the distribution of T to help determine how likely it is that T differs from $|\mu - 40|$ by a large amount. As we construct specific inferences later in this book, we will draw attention to those statistics that play important roles in the inference.

Parameters as Random Variables

There is some controversy over whether parameters should be treated as random variables or merely as numbers that index a distribution. For instance, in Example 7.1.3, we let P stand for the proportion of the patients who avoid relapse from a large group receiving imipramine. We then say that $X_1, X_2, \ldots$ are i.i.d. Bernoulli random variables with parameter p conditional on $P = p$. Here, we are explicitly thinking of P as a random variable, and we give it a distribution. An alternative would be to say that $X_1, X_2, \ldots$ are i.i.d. Bernoulli random variables with parameter p where p is unknown and leave it at that.

If we really want to compute something like the conditional probability that the proportion P is greater than 0.5 given the observations of the first 40 patients, then we need the conditional distribution of P given the first 40 patients, and we must treat P as a random variable. On the other hand, if we are only interested in making probability statements that are indexed by the value of p, then we do not need to think about a random variable called P. For example, we might wish to find two random variables Y_1 and Y_2 (functions of $X_1, \ldots, X_{40}$) such that, no matter what p

equals, the probability that $Y_1 \le p \le Y_2$ is at least 0.9. Some of the inferences that we shall discuss later in this book are of the former type that require treating P as a random variable, and some are of the latter type in which p is merely an index for a distribution.

Some statisticians believe that it is possible and useful to treat parameters as random variables in every statistical inference problem. They believe that the distribution of the parameter is a subjective probability distribution in the sense that it represents an individual experimenter's information and subjective beliefs about where the true value of the parameter is likely to lie. Once they assign a distribution for a parameter, that distribution is no different from any other probability distribution used in the field of statistics, and all of the rules of probability theory apply to every distribution. Indeed, in all of the cases described in this book, the parameters can actually be identified as limits of functions of large collections of potential observations. Here is a typical example.

Example 7.1.12

Parameter as a Limit of Random Variables. In Example 7.1.3, the parameter P can be understood as follows: Imagine an infinite sequence of potential patients receiving imipramine treatment. Assume that for every integer n, the outcomes of every ordered subset of n patients from that infinite sequence has the same joint distribution as the outcomes of every other ordered subset of n patients. In other words, assume that the order in which the patients appear in the sequence is irrelevant to the joint distribution of the patient outcomes. Let P_n be the proportion of patients who don't relapse out of the first n. It can be shown that the probability is 1 that P_n converges to something as $n \to \infty$. That something can be thought of as P, which we have been calling the proportion of successes in a very large population. In this sense, P is a random variable because it is a function of other random variables. A similar argument can be made in all of the statistical models in this book involving parameters, but the mathematics needed to make these arguments precise is too advanced to present here. (Chapter 1 of Schervish (1995) contains the necessary details.) Statisticians who argue as in this example are said to adhere to the Bayesian philosophy of statistics and are called *Bayesians*. ◀

There is another line of reasoning that leads naturally to treating P as a random variable in Example 7.1.12 without relying on an infinite sequence of potential patients. Suppose that the number of potential patients is enough larger than any sample that we will see to make the approximation in Theorem 5.3.4 applicable. Then P is just the proportion of successes among the large population of potential patients. Conditional on $P = p$, the number of successes in a sample of n patients will be approximately a binomial random variable with paramters n and p according to Theorem 5.3.4. If the outcomes of the patients in the sample are random variables, then it makes sense that the proportion of successes among those patients is also random.

There is another group of statisticians who believe that in many problems it is not appropriate to assign a distribution to a parameter but claim instead that the true value of the parameter is a certain fixed number whose value happens to be unknown to the experimenter. These statisticians would assign a distribution to a parameter only when there is extensive previous information about the relative frequencies with which similar parameters have taken each of their possible values in past experiments. If two different scientists could agree on which past experiments were similar to the present experiment, then they might agree on a distribution to be assigned to the parameter. For example, suppose that the proportion θ of defective items in a certain large manufactured lot is unknown. Suppose also that

the same manufacturer has produced many such lots of items in the past and that detailed records have been kept about the proportions of defective items in past lots. The relative frequencies for past lots could then be used to construct a distribution for θ. Statisticians who would argue this way are said to adhere to the frequentist philosophy of statistics and are called *frequentists*.

The frequentists rely on the assumption that there exist infinite sequences of random variables in order to make sense of most of their probability statements. Once one assumes the existence of such an infinite sequence, one finds that the parameters of the distributions being used are limits of functions of the infinite sequences, just as do the Bayesians described above. In this way, the parameters are random variables because they are functions of random variables. The point of disagreement between the two groups is whether it is useful or even possible to assign a distribution to such parameters.

Both Bayesians and frequentists agree on the usefulness of families of distributions for observations indexed by parameters. Bayesians refer to the distribution indexed by parameter value θ as the conditional distribution of the observations given that the parameter equals θ. Frequentists refer to the distribution indexed by θ as the distribution of the observations when θ is the true value of the parameter. The two groups agree that whenever a distribution can be assigned to a parameter, the theory and methods to be described in this chapter are applicable and useful. In Sections 7.2–7.4, we shall explicitly assume that each parameter is a random random variable and we shall assign it a distribution that represents the probabilities that the parameter lies in various subsets of the parameter space. Beginning in Sec. 7.5, we shall consider techniques of estimation that are not based on assigning distributions to parameters.

References

In the remainder of this book, we shall consider many different problems of statistical inference, statistical decision, and experimental design. Some books that discuss statistical theory and methods at about the same level as they will be discussed in this book were mentioned at the end of Sec. 1.1. Some statistics books written at a more advanced level are Bickel and Doksum (2000), Casella and Berger (2002), Cramér (1946), DeGroot (1970), Ferguson (1967), Lehmann (1997), Lehmann and Casella (1998), Rao (1973), Rohatgi (1976), and Schervish (1995).

Exercises

1. Identify the components of the statistical model (as defined in Definition 7.1.1) in Example 7.1.3.

2. Identify two statistical inferences mentioned in Example 7.1.3.

3. In Examples 7.1.4 and 5.7.8 (page 323), identify the components of the statistical model as defined in Definition 7.1.1.

4. In Example 7.1.6, identify the components of the statistical model as defined in Definition 7.1.1.

5. In Example 7.1.6, identify any statistical inference mentioned.

6. In Example 5.8.3 (page 328), identify the components of the statistical model as defined in Definition 7.1.1.

7. In Example 5.4.7 (page 293), identify the components of the statistical model as defined in Definition 7.1.1.

7.2 Prior and Posterior Distributions

The distribution of a parameter before observing any data is called the prior distribution of the parameter. The conditional distribution of the parameter given the observed data is called the posterior distribution. If we plug the observed values of the data into the conditional p.f. or p.d.f. of the data given the parameter, the result is a function of the parameter alone, which is called the likelihood function.

The Prior Distribution

Example 7.2.1 Lifetimes of Electronic Components. In Example 7.1.1, lifetimes $X_1, X_2, \ldots$ of electronic components were modeled as i.i.d. exponential random variables with parameter θ conditional on θ, and θ was interpreted as the failure rate of the components. Indeed, we noted that $n/\sum_{i=1}^{n} X_i$ should converge in probability to θ as n goes to ∞. We then said that θ had the gamma distribution with parameters 1 and 2. ◀

The distribution of θ mentioned at the end of Example 7.2.1 was assigned before observing any of the component lifetimes. For this reason, we call it a *prior distribution*.

Definition 7.2.1 Prior Distribution/p.f./p.d.f. Suppose that one has a statistical model with parameter θ. If one treats θ as random, then the distribution that one assigns to θ before observing the other random variables of interest is called its *prior distribution*. If the parameter space is at most countable, then the prior distribution is discrete and its p.f. is called the *prior p.f.* of θ. If the prior distribution is a continuous distribution, then its p.d.f. is called the *prior p.d.f.* of θ. We shall commonly use the symbol $\xi(\theta)$ to denote the prior p.f. or p.d.f. as a function of θ.

When one treats the parameter as a random variable, the name "prior distribution" is merely another name for the marginal distribution of the parameter.

Example 7.2.2 Fair or Two-Headed Coin. Let θ denote the probability of obtaining a head when a certain coin is tossed, and suppose that it is known that the coin either is fair or has a head on each side. Therefore, the only possible values of θ are $\theta = 1/2$ and $\theta = 1$. If the prior probability that the coin is fair is 0.8, then the prior p.f. of θ is $\xi(1/2) = 0.8$ and $\xi(1) = 0.2$. ◀

Example 7.2.3 Proportion of Defective Items. Suppose that the proportion θ of defective items in a large manufactured lot is unknown and that the prior distribution assigned to θ is the uniform distribution on the interval [0, 1]. Then the prior p.d.f. of θ is

$$\xi(\theta) = \begin{cases} 1 & \text{for } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7.2.1}$$

◀

The prior distribution of a parameter θ must be a probability distribution over the parameter space Ω. We assume that the experimenter or statistician will be able to summarize his previous information and knowledge about where in Ω the value of θ is likely to lie by constructing a probability distribution on the set Ω. In other words, before the experimental data have been collected or observed, the experimenter's past experience and knowledge will lead him to believe that θ is more likely to lie in certain regions of Ω than in others. We shall assume that the relative likelihoods

of the different regions can be expressed in terms of a probability distribution on Ω, namely, the prior distribution of θ.

Example 7.2.4 Lifetimes of Fluorescent Lamps. Suppose that the lifetimes (in hours) of fluorescent lamps of a certain type are to be observed and that the the lifetime of any particular lamp has the exponential distribution with parameter θ. Suppose also that the exact value of θ is unknown, and on the basis of previous experience the prior distribution of θ is taken as the gamma distribution for which the mean is 0.0002 and the standard deviation is 0.0001. We shall determine the prior p.d.f. of θ.

Suppose that the prior distribution of θ is the gamma distribution with parameters α_0 and β_0. It was shown in Theorem 5.7.5 that the mean of this distribution is α_0/β_0 and the variance is α_0/β_0^2. Therefore, $\alpha_0/\beta_0 = 0.0002$ and $\alpha_0^{1/2}/\beta_0 = 0.0001$. Solving these two equations gives $\alpha_0 = 4$ and $\beta_0 = 20{,}000$. It follows from Eq. (5.7.13) that the prior p.d.f. of θ for $\theta > 0$ is as follows:

$$\xi(\theta) = \frac{(20{,}000)^4}{3!}\theta^3 e^{-20{,}000\theta}. \tag{7.2.2}$$

Also, $\xi(\theta) = 0$ for $\theta \le 0$. ◀

In the remainder of this section and Sections 7.3 and 7.4, we shall focus on statistical inference problems in which the parameter θ is a random variable of interest and hence will need to be assigned a distribution. In such problems, we shall refer to the distribution indexed by θ for the other random variables of interest as the conditional distribution for those random variables given θ. For example, this is precisely the language used in Example 7.2.1 where the parameter is θ, the failure rate. In referring to the conditional p.f. or p.d.f. of random variables, such as $X_1, X_2, \ldots$ in Example 7.2.1, we shall use the notation of conditional p.f.'s and p.d.f.'s. For example, if we let $\boldsymbol{X} = (X_1, \ldots, X_m)$ in Example 7.2.1, the conditional p.d.f. of $\boldsymbol{X}$ given θ is

$$f_m(\boldsymbol{x}|\theta) = \begin{cases} \theta^m \exp(-\theta[x_1 + \cdots + x_m]) & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7.2.3}$$

In many problems, such as Example 7.2.1, the observable data $X_1, X_2, \ldots$ are modeled as a random sample from a univariate distribution indexed by θ. In these cases, let $f(x|\theta)$ denote the p.f. or p.d.f. of a single random variable under the distribution indexed by θ. In such a case, using the above notation,

$$f_m(\boldsymbol{x}|\theta) = f(x_1|\theta) \cdots f(x_m|\theta).$$

When we treat θ as a random variable, $f(x|\theta)$ is the conditional p.f. or p.d.f. of each observation X_i given θ, and the observations are conditionally i.i.d. given θ. In summary, the following two expressions are to be understood as equivalent:

- $X_1, \ldots, X_n$ form a random sample with p.f. or p.d.f. $f(x|\theta)$.
- $X_1, \ldots, X_n$ are conditionally i.i.d. given θ with conditional p.f. or p.d.f. $f(x|\theta)$.

Although we shall generally use the wording in the first bullet above for simplicity, it is often useful to remember that the two wordings are equivalent when we treat θ as a random variable.

Sensitivity Analysis and Improper Priors In Example 2.3.8 on page 84, we saw a situation in which two very different sets of prior probabilities were used for a collection of events. After we observed data, however, the posterior probabilities were

quite similar. In Example 5.8.4 on page 330, we used a large collection of prior distributions for a parameter in order to see how much impact the prior distribution had on the posterior probability of a single important event. It is a common practice to compare the posterior distributions that arise from several different prior distributions in order to see how much effect the prior distribution has on the answers to important questions. Such comparisons are called *sensitivity analysis*.

It is very often the case that different prior distributions do not make much difference after the data have been observed. This is especially true if there are a lot of data or if the prior distributions being compared are very spread out. This observation has two important implications. First, the fact that different experimenters might not agree on a prior distribution becomes less important if there are a lot of data. Second, experimenters might be less inclined to spend time specifying a prior distribution if it is not going to matter much which one is specified. Unfortunately, if one does not specify some prior distribution, there is no way to calculate a conditional distribution of the parameter given the data.

As an expedient, there are some calculations available that attempt to capture the idea that the data contain much more information than is available a priori. Usually, these calculations involve using a function $\xi(\theta)$ as if it were a prior p.d.f. for the parameter θ but such that $\int \xi(\theta)\, d\theta = \infty$, which clearly violates the definition of p.d.f. Such priors are called *improper*. We shall discuss improper priors in more detail in Sec. 7.3.

The Posterior Distribution

Example 7.2.5 Lifetimes of Fluorescent Lamps. In Example 7.2.4, we constructed a prior distribution for the parameter θ that specifies the exponential distribution for a collection of lifetimes of fluorescent lamps. Suppose that we observe a collection of n such lifetimes. How would we change the distribution of θ to take account of the observed data? ◀

Definition 7.2.2 Posterior Distribution/p.f./p.d.f. Consider a statistical inference problem with parameter θ and random variables $X_1, \ldots, X_n$ to be observed. The conditional distribution of θ given $X_1, \ldots, X_n$ is called the *posterior distribution* of θ. The conditional p.f. or p.d.f. of θ given $X_1 = x_1, \ldots, X_n = x_n$ is called the *posterior p.f.* or *posterior p.d.f.* of θ and is typically denoted $\xi(\theta|x_1, \ldots, x_n)$.

When one treats the parameter as a random variable, the name "posterior distribution" is merely another name for the conditional distribution of the parameter given the data. Bayes' theorem for random variables (3.6.13) and for random vectors (3.7.15) tells us how to compute the posterior p.d.f. or p.f. of θ after observing data. We shall review the derivation of Bayes' theorem here using the specific notation of prior distributions and parameters.

Theorem 7.2.1 Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$. Suppose also that the value of the parameter θ is unknown and the prior p.d.f. or p.f. of θ is $\xi(\theta)$. Then the posterior p.d.f. or p.f. of θ is

$$\xi(\theta|\boldsymbol{x}) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\boldsymbol{x})} \quad \text{for } \theta \in \Omega,$$

where g_n is the marginal joint p.d.f. or p.f. of $X_1, \ldots, X_n$.

Proof For simplicity, we shall assume that the parameter space Ω is either an interval of the real line or the entire real line and that $\xi(\theta)$ is a prior p.d.f. on Ω, rather than a prior p.f. However, the proof that will be given here can be adapted easily to a problem in which $\xi(\theta)$ is a p.f.

Since the random variables $X_1, \ldots, X_n$ form a random sample from the distribution for which the p.d.f. is $f(x|\theta)$, it follows from Sec. 3.7 that their conditional joint p.d.f. or p.f. $f_n(x_1, \ldots, x_n|\theta)$ given θ is

$$f_n(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta). \tag{7.2.4}$$

If we use the vector notation $\boldsymbol{x} = (x_1, \ldots, x_n)$, then the joint p.d.f. in Eq. (7.2.4) can be written more compactly as $f_n(\boldsymbol{x}|\theta)$. Eq. (7.2.4) merely expresses the fact that $X_1, \ldots, X_n$ are conditionally independent and identically distributed given θ, each having p.d.f. or p.f. $f(x|\theta)$.

If we multiply the conditional joint p.d.f. or p.f. by the p.d.f. $\xi(\theta)$, we obtain the $(n+1)$-dimensional joint p.d.f. (or p.f./p.d.f.) of $X_1, \ldots, X_n$ and θ in the form

$$f(\boldsymbol{x}, \theta) = f_n(\boldsymbol{x}|\theta)\xi(\theta). \tag{7.2.5}$$

The marginal joint p.d.f. or p.f. of $X_1, \ldots, X_n$ can now be obtained by integrating the right-hand side of Eq. (7.2.5) over all values of θ. Therefore, the n-dimensional marginal joint p.d.f. or p.f. $g_n(\boldsymbol{x})$ of $X_1, \ldots, X_n$ can be written in the form

$$g_n(\boldsymbol{x}) = \int_\Omega f_n(\boldsymbol{x}|\theta)\xi(\theta)\, d\theta. \tag{7.2.6}$$

Eq. (7.2.6) is just an instance of the law of total probability for random vectors (3.7.14).

Furthermore, the conditional p.d.f. of θ given that $X_1 = x_1, \ldots, X_n = x_n$, namely, $\xi(\theta|\boldsymbol{x})$, must be equal to $f(\boldsymbol{x}, \theta)$ divided by $g_n(\boldsymbol{x})$. Thus, we have

$$\xi(\theta|\boldsymbol{x}) = \frac{f_n(\boldsymbol{x}|\theta)\xi(\theta)}{g_n(\boldsymbol{x})} \quad \text{for } \theta \in \Omega, \tag{7.2.7}$$

which is Bayes' theorem restated for parameters and random samples. If $\xi(\theta)$ is a p.f., so that the prior distribution is discrete, just replace the integral in (7.2.6) by the sum over all of the possible values of θ. ■

Example 7.2.6 Lifetimes of Fluorescent Lamps. Suppose again, as in Examples 7.2.4 and 7.2.5, that the distribution of the lifetimes of fluorescent lamps of a certain type is the exponential distribution with parameter θ, and the prior distribution of θ is a particular gamma distribution for which the p.d.f. $\xi(\theta)$ is given by Eq. (7.2.2). Suppose also that the lifetimes $X_1, \ldots, X_n$ of a random sample of n lamps of this type are observed. We shall determine the posterior p.d.f. of θ given that $X_1 = x_1, \ldots, X_n = x_n$.

By Eq. (5.7.16), the p.d.f. of each observation X_i is

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint p.d.f. of $X_1, \ldots, X_n$ can be written in the following form, for $x_i > 0$ $(i = 1, \ldots, n)$:

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta y},$$

where $y = \sum_{i=1}^{n} x_i$. As $f_n(\boldsymbol{x}|\theta)$ will be used in constructing the posterior distribution of θ, it is now apparent that the statistic $Y = \sum_{i=1}^{n} X_i$ will be used in any inference that makes use of the posterior distribution.

Since the prior p.d.f. $\xi(\theta)$ is given by Eq. (7.2.2), it follows that for $\theta > 0$,

$$f_n(\boldsymbol{x}|\theta)\xi(\theta) = \theta^{n+3}e^{-(y+20{,}000)\theta}. \tag{7.2.8}$$

We need to compute $g_n(\boldsymbol{x})$, which is the integral of (7.2.8) over all θ:

$$g_n(\boldsymbol{x}) = \int_0^\infty \theta^{n+3}e^{-(y+20{,}000)\theta}d\theta = \frac{\Gamma(n+4)}{(y+20{,}000)^{n+4}},$$

where the last equality follows from Theorem 5.7.3. Hence,

$$\begin{aligned}\xi(\theta|\boldsymbol{x}) &= \frac{\theta^{n+3}e^{-(y+20{,}000)\theta}}{\dfrac{\Gamma(n+4)}{(y+20{,}000)^{n+4}}} \\ &= \frac{(y+20{,}000)^{n+4}}{\Gamma(n+4)}e^{-(y+20{,}000)\theta},\end{aligned} \tag{7.2.9}$$

for $\theta > 0$. When we compare this expression with Eq. (5.7.13), we can see that it is the p.d.f. of the gamma distribution with parameters $n + 4$ and $y + 20{,}000$. Hence, this gamma distribution is the posterior distribution of θ.

As a specific example, suppose that we observe the following $n = 5$ lifetimes in hours: 2911, 3403, 3237, 3509, and 3118. Then $y = 16{,}178$, and the posterior distribution of θ is the gamma distribution with parameters 9 and 36,178. The top panel of Fig. 7.1 displays both the prior and posterior p.d.f.'s in this example. It is clear that the data have caused the distribution of θ to change somewhat from the prior to the posterior.

At this point, it might be appropriate to perform a sensitivity analysis. For example, how would the posterior distribution change if we had chosen a different prior distribution? To be specific, consider the gamma prior with parameters 1 and 1000. This prior has the same standard deviation as the original prior, but the mean is five times as big. The posterior distribution would then be the gamma distribution with parameters 6 and 17,178. The p.d.f.'s of this pair of prior and posterior are plotted in the lower panel of Fig. 7.1. One can see that both the prior and the posterior in the bottom panel are more spread out than their counterparts in the upper panel. It

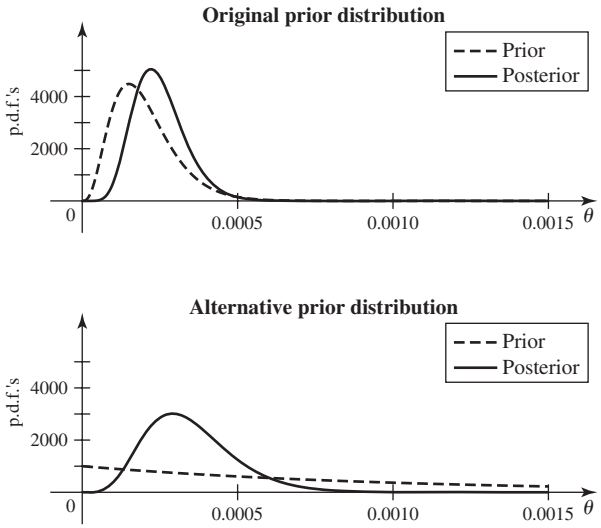

Figure 7.1 Prior and posterior p.d.f.'s in Example 7.2.6. The top panel is based on the original prior. The bottom panel is based on the alternative prior that was part of the sensitivity analysis.

is clear that the choice of prior distribution is going to make a difference with this small data set. ◀

The names "prior" and "posterior" derive from the Latin words for "former" and "coming after." The prior distribution is the distribution of θ that comes before observing the data, and posterior distribution comes after observing the data.

The Likelihood Function

The denominator on the right side of Eq. (7.2.7) is simply the integral of the numerator over all possible values of θ. Although the value of this integral depends on the observed values $x_1, \ldots, x_n$, it does not depend on θ and it may be treated as a constant when the right-hand side of Eq. (7.2.7) is regarded as a p.d.f. of θ. We may therefore replace Eq. (7.2.7) with the following relation:

$$\xi(\theta|\boldsymbol{x}) \propto f_n(\boldsymbol{x}|\theta)\xi(\theta). \tag{7.2.10}$$

The proportionality symbol $\propto$ is used here to indicate that the left side is equal to the right side except possibly for a constant factor, the value of which may depend on the observed values $x_1, \ldots, x_n$ but does not depend on θ. The appropriate constant factor that will establish the equality of the two sides in the relation (7.2.10) can be determined at any time by using the fact that $\int_\Omega \xi(\theta|\boldsymbol{x})\, d\theta = 1$, because $\xi(\theta|\boldsymbol{x})$ is a p.d.f. of θ.

One of the two functions on the right-hand side of Eq. (7.2.10) is the prior p.d.f. of θ. The other function has a special name also.

Definition 7.2.3 Likelihood Function. When the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \ldots, x_n$, it is called the *likelihood function*.

The relation (7.2.10) states that the posterior p.d.f. of θ is proportional to the product of the likelihood function and the prior p.d.f. of θ.

By using the proportionality relation (7.2.10), it is often possible to determine the posterior p.d.f. of θ without explicitly performing the integration in Eq. (7.2.6). If we can recognize the right side of the relation (7.2.10) as being equal to one of the standard p.d.f.'s introduced in Chapter 5 or elsewhere in this book, except possibly for a constant factor, then we can easily determine the appropriate factor that will convert the right side of (7.2.10) into a proper p.d.f. of θ. We shall illustrate these ideas by considering again Example 7.2.3.

Example 7.2.7 Proportion of Defective Items. Suppose again, as in Example 7.2.3, that the proportion θ of defective items in a large manufactured lot is unknown and that the prior distribution of θ is a uniform distribution on the interval [0, 1]. Suppose also that a random sample of n items is taken from the lot, and for $i = 1, \ldots, n$, let $X_i = 1$ if the ith item is defective, and let $X_i = 0$ otherwise. Then $X_1, \ldots, X_n$ form n Bernoulli trials with parameter θ. We shall determine the posterior p.d.f. of θ.

It follows from Eq. (5.2.2) that the p.f. of each observation X_i is

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{for } x = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, if we let $y = \sum_{i=1}^n x_i$, then the joint p.f. of $X_1, \ldots, X_n$ can be written in the following form for $x_i = 0$ or 1 $(i = 1, \ldots, n)$:

$$f_n(\boldsymbol{x}|\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.11}$$

Since the prior p.d.f. $\xi(\theta)$ is given by Eq. (7.2.1), it follows that for $0 < \theta < 1$,

$$f_n(\boldsymbol{x}|\theta)\xi(\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.12}$$

When we compare this expression with Eq. (5.8.3), we can see that, except for a constant factor, it is the p.d.f. of the beta distribution with parameters $\alpha = y + 1$ and $\beta = n - y + 1$. Since the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is proportional to the right side of Eq. (7.2.12), it follows that $\xi(\theta|\boldsymbol{x})$ must be the p.d.f. of the beta distribution with parameters $\alpha = y + 1$ and $\beta = n - y + 1$. Therefore, for $0 < \theta < 1$,

$$\xi(\theta|\boldsymbol{x}) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}. \tag{7.2.13}$$

In this example, the statistic $Y = \sum_{i=1}^{n} X_i$ is being used to construct the posterior distribution, and hence will be used in any inference that is based on the posterior distribution. ◀

Note: Normalizing Constant for Posterior p.d.f. The steps that got us from (7.2.12) to (7.2.13) are an example of a very common technique for determining a posterior p.d.f. We can drop any inconvenient constant factor from the prior p.d.f. and from the likelihood function before we multiply them together as in (7.2.10). Then we look at the resulting product, call it $g(\theta)$, to see if we recognize it as looking like part of a p.d.f. that we have seen elsewhere. If indeed we find a named distribution with p.d.f. equal to $cg(\theta)$, then our posterior p.d.f. is also $cg(\theta)$, and our posterior distribution has the corresponding name, just as in Example 7.2.7.

Sequential Observations and Prediction

In many experiments, the observations $X_1, \ldots, X_n$, which form the random sample, must be obtained sequentially, that is, one at a time. In such an experiment, the value of X_1 is observed first, the value of X_2 is observed next, the value of X_3 is then observed, and so on. Suppose that the prior p.d.f. of the parameter θ is $\xi(\theta)$. After the value x_1 of X_1 has been observed, the posterior p.d.f. $\xi(\theta|x_1)$ can be calculated in the usual way from the relation

$$\xi(\theta|x_1) \propto f(x_1|\theta)\xi(\theta). \tag{7.2.14}$$

Since X_1 and X_2 are conditionally independent given θ, the conditional p.f. or p.d.f. of X_2 given θ and $X_1 = x_1$ is the same as that given θ alone, namely, $f(x_2|\theta)$. Hence, the posterior p.d.f. of θ in Eq. (7.2.14) serves as the prior p.d.f. of θ when the value of X_2 is to be observed. Thus, after the value x_2 of X_2 has been observed, the posterior p.d.f. $\xi(\theta|x_1, x_2)$ can be calculated from the relation

$$\xi(\theta|x_1, x_2) \propto f(x_2|\theta)\xi(\theta|x_1). \tag{7.2.15}$$

We can continue in this way, calculating an updated posterior p.d.f. of θ after each observation and using that p.d.f. as the prior p.d.f. of θ for the next observation. The posterior p.d.f. $\xi(\theta|x_1, \ldots, x_{n-1})$ after the values $x_1, \ldots, x_{n-1}$ have been observed will ultimately be the prior p.d.f. of θ for the final observed value of X_n. The posterior p.d.f. after all n values $x_1, \ldots, x_n$ have been observed will therefore be specified by the relation

$$\xi(\theta|\boldsymbol{x}) \propto f(x_n|\theta)\xi(\theta|x_1, \ldots, x_{n-1}). \tag{7.2.16}$$

Alternatively, after all n values $x_1, \ldots, x_n$ have been observed, we could calculate the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ in the usual way by combining the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ with the original prior p.d.f. $\xi(\theta)$, as indicated in Eq. (7.2.7). It can be shown (see

Exercise 8) that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ will be the same regardless of whether it is calculated directly by using Eq. (7.2.7) or sequentially by using Eqs. (7.2.14), (7.2.15), and (7.2.16). This property was illustrated in Sec. 2.3 (see page 80) for a coin that is known either to be fair or to have a head on each side. After each toss of the coin, the posterior probability that the coin is fair is updated.

The proportionality constants in Eqs. (7.2.14)–(7.2.16) have a useful interpretation. For example, in (7.2.16) the proportionality constant is 1 over the integral of the right side with respect to θ. But this integral is the conditional p.d.f. or p.f. of X_n given $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$, according to the conditional version of the law of total probability (3.7.16). For example, if θ has a continuous distribution,

$$f(x_n|x_1, \ldots, x_{n-1}) = \int f(x_n|\theta)\xi(\theta|x_1, \ldots, x_{n-1})d\theta. \tag{7.2.17}$$

The proportionality constant in (7.2.16) is 1 over (7.2.17). So, if we are interested in predicting the nth observation in a sequence after observing the first $n-1$, we can use (7.2.17), which is also 1 over the proportionality constant in Eq. (7.2.16), as the conditional p.f. or p.d.f. of X_n given the first $n-1$ observations.

Example 7.2.8

Lifetimes of Fluorescent Lamps. In Example 7.2.6, conditional on θ, the lifetimes of fluorescent lamps are independent exponential random variables with parameter θ. We also observed the lifetimes of five lamps, and the posterior distribution of θ was found to be the gamma distribution with parameters 9 and 36,178. Suppose that we want to predict the lifetime X_6 of the next lamp.

The conditional p.d.f. of X_6, the lifetime of the next lamp, given the first five lifetimes equals the integral of $\xi(\theta|\boldsymbol{x})f(x_6|\theta)$ with respect to θ. The posterior p.d.f. of θ is $\xi(\theta|\boldsymbol{x}) = 2.633 \times 10^{36}\theta^8 e^{-36,178\theta}$ for $\theta > 0$. So, for $x_6 > 0$

$$\begin{aligned} f(x_6|\boldsymbol{x}) &= \int_0^\infty 2.633 \times 10^{36}\theta^8 e^{-36,178\theta}\theta e^{-x_6\theta}\, d\theta \\ &= 2.633 \times 10^{36} \int_0^\infty \theta^9 e^{-(x_6+36,178)\theta}\, d\theta \\ &= 2.633 \times 10^{36}\frac{\Gamma(10)}{(x_6+36,178)^{10}} = \frac{9.555 \times 10^{41}}{(x_6+36,178)^{10}}. \end{aligned} \tag{7.2.18}$$

We can use this p.d.f. to perform any calculation we wish concerning the distribution of X_6 given the observed lifetimes. For example, the probability that the sixth lamp lasts more than 3000 hours equals

$$\Pr(X_6 > 3000|\boldsymbol{x}) = \int_{3000}^\infty \frac{9.555 \times 10^{41}}{(x_6+36,178)^{10}}\, dx_6 = \frac{9.555 \times 10^{41}}{9 \times 39,178^9} = 0.4882.$$

Finally, we can continue the sensitivity analysis that was started in Example 7.2.6. If it is important to know the probability that the next lifetime is at least 3000, we can see how much influence the choice of prior distribution has made on this calculation. Using the second prior distribution (gamma with parameters 1 and 1000), we found that the posterior distribution of θ was the gamma distribution with parameters 6 and 17,178. We could compute the conditional p.d.f. of X_6 given the observed data in the same way as we did with the original posterior, and it would be

$$f(x_6|\boldsymbol{x}) = \frac{1.542 \times 10^{26}}{(x_6+17,178)^7}, \quad \text{for } x_6 > 0. \tag{7.2.19}$$

With this p.d.f., the probability that $X_6 > 3000$ is

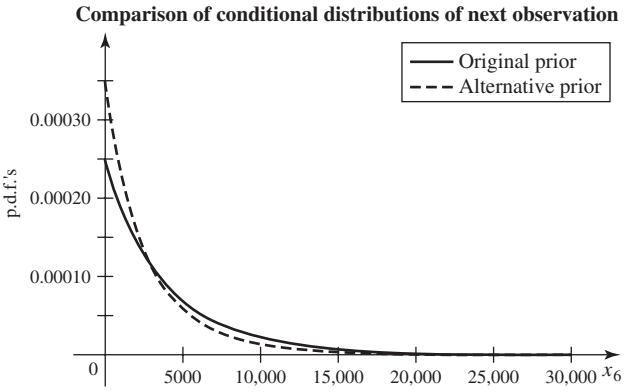

Figure 7.2 Two possible conditional p.d.f.'s, Eqs. (7.2.18) and (7.2.19) for X_6 given the observed data in Example 7.2.8. The two p.d.f.'s were computed using the two different posterior distributions that were derived from the two different prior distributions in Example 7.2.6.

$$\Pr(X_6 > 3000|\boldsymbol{x}) = \int_{3000}^{\infty} \frac{1.542 \times 10^{26}}{(x_6 + 17{,}178)^7}\, dx_6 = \frac{1.542 \times 10^{26}}{6 \times 20{,}178^6} = 0.3807.$$

As we noted at the end of Example 7.2.6, the different priors make a considerable difference in the inferences that we can make. If it is important to have a precise value of $\Pr(X_6 > 3000|\boldsymbol{x})$, we need a larger sample. The two different p.d.f.'s of X_6 given $\boldsymbol{x}$ can be compared in Fig. 7.2. The p.d.f. from Eq. (7.2.18) is higher for intermediate values of x_6, while the one from Eq. (7.2.19) is higher for the extreme values of x_6. ◀

Summary

The prior distribution of a parameter describes our uncertainty about the parameter before observing any data. The likelihood function is the conditional p.d.f. or p.f. of the data given the parameter when regarded as a function of the parameter with the observed data plugged in. The likelihood tells us how much the data will alter our uncertainty. Large values of the likelihood correspond to parameter values where the posterior p.d.f. or p.f. will be higher than the prior. Low values of the likelihood occur at parameter values where the posterior will be lower than the prior. The posterior distribution of the parameter is the conditional distribution of the parameter given the data. It is obtained using Bayes' theorem for random variables, which we first saw on page 148. We can predict future observations that are conditionally independent of the observed data given θ by using the conditional version of the law of total probability that we saw on page 163.

Exercises

1. Consider again the situation described in Example 7.2.8. This time, suppose that the experimenter believes that the prior distribution of θ is the gamma distribution with parameters 1 and 5000. What would this experimenter compute as the value of $\Pr(X_6 > 3000|\boldsymbol{x})$?

2. Suppose that the proportion θ of defective items in a large manufactured lot is known to be either 0.1 or 0.2, and the prior p.f. of θ is as follows:

$$\xi(0.1) = 0.7 \quad \text{and} \quad \xi(0.2) = 0.3.$$

Suppose also that when eight items are selected at random from the lot, it is found that exactly two of them are defective. Determine the posterior p.f. of θ.

3. Suppose that the number of defects on a roll of magnetic recording tape has a Poisson distribution for which the mean λ is either 1.0 or 1.5, and the prior p.f. of λ is as

follows:

$$\xi(1.0) = 0.4 \quad \text{and} \quad \xi(1.5) = 0.6.$$

If a roll of tape selected at random is found to have three defects, what is the posterior p.f. of λ?

4. Suppose that the prior distribution of some parameter θ is a gamma distribution for which the mean is 10 and the variance is 5. Determine the prior p.d.f. of θ.

5. Suppose that the prior distribution of some parameter θ is a beta distribution for which the mean is 1/3 and the variance is 1/45. Determine the prior p.d.f. of θ.

6. Suppose that the proportion θ of defective items in a large manufactured lot is unknown, and the prior distribution of θ is the uniform distribution on the interval [0, 1]. When eight items are selected at random from the lot, it is found that exactly three of them are defective. Determine the posterior distribution of θ.

7. Consider again the problem described in Exercise 6, but suppose now that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} 2(1-\theta) & \text{for } 0 < \theta < 1, \\ 0 & \text{otherwise.} \end{cases}$$

As in Exercise 6, suppose that in a random sample of eight items exactly three are found to be defective. Determine the posterior distribution of θ.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, the value of θ is unknown, and the prior p.d.f. of θ is $\xi(\theta)$. Show that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is the same regardless of whether it is calculated directly by using Eq. (7.2.7) or sequentially by using Eqs. (7.2.14), (7.2.15), and (7.2.16).

9. Consider again the problem described in Exercise 6, and assume the same prior distribution of θ. Suppose now, however, that instead of selecting a random sample of eight items from the lot, we perform the following experiment: Items from the lot are selected at random one by one until exactly three defectives have been found. If we find that we must select a total of eight items in this experiment, what is the posterior distribution of θ at the end of the experiment?

10. Suppose that a single observation X is to be taken from the uniform distribution on the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, the value of θ is unknown, and the prior distribution of θ is the uniform distribution on the interval [10, 20]. If the observed value of X is 12, what is the posterior distribution of θ?

11. Consider again the conditions of Exercise 10, and assume the same prior distribution of θ. Suppose now, however, that six observations are selected at random from the uniform distribution on the interval $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, and their values are 11.0, 11.5, 11.7, 11.1, 11.4, and 10.9. Determine the posterior distribution of θ.

7.3 Conjugate Prior Distributions

For each of the most popular statistical models, there exists a family of distributions for the parameter with a very special property. If the prior distribution is chosen to be a member of that family, then the posterior distribution will also be a member of that family. Such a family of distributions is called a conjugate family. Choosing a prior distribution from a conjugate family will typically make it particularly simple to calculate the posterior distribution.

Sampling from a Bernoulli Distribution

Example 7.3.1 A Clinical Trial. In Example 5.8.5 (page 330), we were observing patients in a clinical trial. The proportion P of successful outcomes among all possible patients was a random variable for which we chose a distribution from the family of beta distributions. This choice made the calculation of the conditional distribution of P given the observed data very simple at the end of that example. Indeed, the conditional distribution of P given the data was another member of the beta family. ◀

That the result in Example 7.3.1 occurs in general is the subject of the next theorem.

Theorem 7.3.1 Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown ($0 < \theta < 1$). Suppose also that the prior distribution

of θ is the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the beta distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n - \sum_{i=1}^{n} x_i$.

Theorem 7.3.1 is just a restatement of Theorem 5.8.2 (page 329), and its proof is essentially the calculation in Example 5.8.3.

Updating the Posterior Distribution One implication of Theorem 7.3.1 is the following: Suppose that the proportion θ of defective items in a large shipment is unknown, the prior distribution of θ is the beta distribution with parameters α and β, and n items are selected one at a time at random from the shipment and inspected. Assume that the items are conditionally independent given θ. If the first item inspected is defective, the posterior distribution of θ will be the beta distribution with parameters $\alpha + 1$ and β. If the first item is nondefective, the posterior distribution will be the beta distribution with parameters α and $\beta + 1$. The process can be continued in the following way: Each time an item is inspected, the current posterior beta distribution of θ is changed to a new beta distribution in which the value of either the parameter α or the parameter β is increased by one unit. The value of α is increased by one unit each time a defective item is found, and the value of β is increased by one unit each time a nondefective item is found.

Definition 7.3.1

Conjugate Family/Hyperparameters. Let $X_1, X_2, \ldots$ be conditionally i.i.d. given θ with common p.f. or p.d.f. $f(x|\theta)$. Let Ψ be a family of possible distributions over the parameter space Ω. Suppose that, no matter which prior distribution ξ we choose from Ψ, no matter how many observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ we observe, and no matter what are their observed values $\boldsymbol{x} = (x_1, \ldots, x_n)$, the posterior distribution $\xi(\theta|\boldsymbol{x})$ is a member of Ψ. Then Ψ is called a *conjugate family of prior distributions* for samples from the distributions $f(x|\theta)$. It is also said that the family Ψ is *closed under sampling* from the distributions $f(x|\theta)$. Finally, if the distributions in Ψ are parametrized by further parameters, then the associated parameters for the prior distribution are called the *prior hyperparameters* and the associated parameters of the posterior distribution are called the *posterior hyperparameters*.

Theorem 7.3.1 says that the family of beta distributions is a conjugate family of prior distributions for samples from a Bernoulli distribution. If the prior distribution of θ is a beta distribution, then the posterior distribution at each stage of sampling will also be a beta distribution, regardless of the observed values in the sample. Also, the family of beta distributions is closed under sampling from Bernoulli distributions. The parameters α and β in Theorem 7.3.1 are the prior hyperparameters. The corresponding parameters of the posterior distributions ($\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n - \sum_{i=1}^{n} x_i$) are the posterior hyperparameters. The statistic $\sum_{i=1}^{n} X_i$ is needed to compute the posterior distribution, hence it will be needed to perform any inference based on the posterior distribution. Exercises 23 and 24 introduce a general collection of p.d.f.'s $f(x|\theta)$ for which conjugate families of priors exist. Most of the familiar named distributions are covered by these exercises. The various uniform distributions are notable exceptions.

Example 7.3.2

The Variance of the Posterior Beta Distribution. Suppose that the proportion θ of defective items in a large shipment is unknown, the prior distribution of θ is the uniform distribution on the interval [0, 1], and items are to be selected at random from the shipment and inspected until the variance of the posterior distribution of θ

has been reduced to the value 0.01 or less. We shall determine the total number of defective and nondefective items that must be obtained before the sampling process is stopped.

As stated in Sec. 5.8, the uniform distribution on the interval [0, 1] is the beta distribution with parameters 1 and 1. Therefore, after y defective items and z nondefective items have been obtained, the posterior distribution of θ will be the beta distribution with $\alpha = y + 1$ and $\beta = z + 1$. It was shown in Theorem 5.8.3 that the variance of the beta distribution with parameters α and β is $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Therefore, the variance V of the posterior distribution of θ will be

$$V = \frac{(y + 1)(z + 1)}{(y + z + 2)^2(y + z + 3)}.$$

Sampling is to stop as soon as the number of defectives y and the number of nondefectives z that have been obtained are such that $V \leq 0.01$. It can be shown (see Exercise 2) that it will not be necessary to select more than 22 items, but it is necessary to select at least seven items. ◀

Example 7.3.3

Glove Use by Nurses. Friedland et al. (1992) studied 23 nurses in an inner-city hospital before and after an educational program on the importance of wearing gloves. They recorded whether or not the nurses wore gloves during procedures in which they might come in contact with bodily fluids. Before the educational program the nurses were observed during 51 procedures, and they wore gloves in only 13 of them. Let θ be the probability that a nurse will wear gloves two months after the educational program. We might be interested in how θ compares to 13/51, the observed proportion before the program.

We shall consider two different prior distributions for θ in order to see how sensitive the posterior distribution of θ is to the choice of prior distribution. The first prior distribution will be uniform on the interval [0, 1], which is also the beta distribution with parameters 1 and 1. The second prior distribution will be the beta distribution with parameters 13 and 38. This second prior distribution has much smaller variance than the first and has its mean at 13/51. Someone holding the second prior distribution believes fairly strongly that the educational program will have no noticeable effect.

Two months after the educational program, 56 procedures were observed with the nurses wearing gloves in 50 of them. The posterior distribution of θ, based on the first prior, would then be the beta distribution with parameters $1 + 50 = 51$ and $1 + 6 = 7$. In particular, the posterior mean of θ is $51/(51 + 7) = 0.88$, and the posterior probability that $\theta > 2 \times 13/51$ is essentially 1. Based on the second prior, the posterior distribution would be the beta distribution with parameters $13 + 50 = 63$ and $38 + 6 = 44$. The posterior mean would be 0.59, and the posterior probability that $\theta > 2 \times 13/51$ is 0.95. So, even to someone who was initially skeptical, the educational program seems to have been quite effective. The probability is quite high that nurses are at least twice as likely to wear gloves after the program as they were before.

Figure 7.3 shows the p.d.f.'s of both of the posterior distributions computed above. The distributions are clearly very different. For example, the first posterior gives probability greater than 0.99 that $\theta > 0.7$, while the second gives probability less than 0.001 to $\theta > 0.7$. However, since we are only interested in the probability that $\theta > 2 \times 13/51 = 0.5098$, we see that both posteriors agree that this probability is quite large. ◀

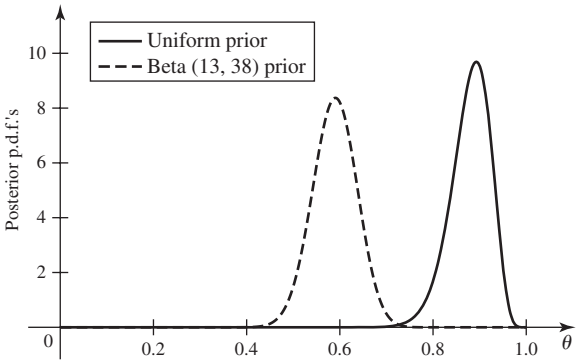

Figure 7.3 Posterior p.d.f.'s in Example 7.2.6. The curves are labeled by the prior that led to the corresponding posterior.

Sampling from a Poisson Distribution

Example 7.3.4 Customer Arrivals. A store owner models customer arrivals as a Poisson process with unknown rate θ per hour. She assigns θ a gamma prior distribution with parameters 3 and 2. Let X be the number of customers that arrive in a specific one-hour period. If $X = 3$ is observed, the store owner wants to update the distribution of θ. ◀

When samples are taken from a Poisson distribution, the family of gamma distributions is a conjugate family of prior distributions. This relationship is shown in the next theorem.

Theorem 7.3.2 Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with mean $\theta > 0$, and θ is unknown. Suppose also that the prior distribution of θ is the gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ, given that $X_i = x_i$ $(i = 1, \ldots, n)$, is the gamma distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$.

Proof Let $y = \sum_{i=1}^{n} x_i$. Then the likelihood function $f_n(\boldsymbol{x}|\theta)$ satisfies the relation

$$f_n(\boldsymbol{x}|\theta) \propto e^{-n\theta}\theta^y.$$

In this relation, a factor that involves $\boldsymbol{x}$ but does not depend on θ has been dropped from the right side. Furthermore, the prior p.d.f. of θ has the form

$$\xi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta} \quad \text{for } \theta > 0.$$

Since the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ is proportional to $f_n(\boldsymbol{x}|\theta)\xi(\theta)$, it follows that

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{\alpha+y-1}e^{-(\beta+n)\theta} \quad \text{for } \theta > 0.$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the gamma distribution with parameters $\alpha + y$ and $\beta + n$. Therefore, the posterior distribution of θ is as specified in the theorem. ■

In Theorem 7.3.2, the numbers α and β are the prior hyperparameters, while $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$ are the posterior hyperparameters. Note that the statistic $Y = \sum_{i=1}^{n} X_i$ is used to compute the posterior distribution of θ, and hence it will be part of any inference based on the posterior.

Example 7.3.5 Customer Arrivals. In Example 7.3.4, we can apply Theorem 7.3.2 with $n = 1$, $\alpha = 3$, $\beta = 2$, and $x_1 = 3$. The posterior distribution of θ given $X = 3$ is the gamma distribution with parameters 6 and 3. ◀

Example 7.3.6 The Variance of the Posterior Gamma Distribution. Consider a Poisson distribution for which the mean θ is unknown, and suppose that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} 2e^{-2\theta} & \text{for } \theta > 0, \\ 0 & \text{for } \theta \le 0. \end{cases}$$

Suppose also that observations are to be taken at random from the given Poisson distribution until the variance of the posterior distribution of θ has been reduced to the value 0.01 or less. We shall determine the number of observations that must be taken before the sampling process is stopped.

The given prior p.d.f. $\xi(\theta)$ is the p.d.f. of the gamma distribution with prior hyperparameters $\alpha = 1$ and $\beta = 2$. Therefore, after we have obtained n observed values $x_1, \ldots, x_n$, the sum of which is $y = \sum_{i=1}^{n} x_i$, the posterior distribution of θ will be the gamma distribution with posterior hyperparameters $y + 1$ and $n + 2$. It was shown in Theorem 5.4.2 that the variance of the gamma distribution with parameters α and β is α/β^2. Therefore, the variance V of the posterior distribution of θ will be

$$V = \frac{y + 1}{(n + 2)^2}.$$

Sampling is to stop as soon as the sequence of observed values $x_1, \ldots, x_n$ is such that $V \le 0.01$. Unlike Example 7.3.2, there is no uniform bound on how large n needs to be because y can be arbitrarily large no matter what n is. Clearly, it takes at least $n = 8$ observations before $V \le 0.01$. ◀

Sampling from a Normal Distribution

Example 7.3.7 Automobile Emissions. Consider again the sampling of automobile emissions, in particular oxides of nitrogen, described in Example 5.6.1 on page 302. Prior to observing the data, suppose that an engineer believed that each emissions measurement had the normal distribution with mean θ and standard deviation 0.5 but that θ was unknown. The engineer's uncertainty about θ might be described by another normal distribution with mean 2.0 and standard deviation 1.0. After seeing the data in Fig. 5.1, how would this engineer describe her uncertainty about θ? ◀

When samples are taken from a normal distribution for which the value of the mean θ is unknown but the value of the variance σ^2 is known, the family of normal distributions is itself a conjugate family of prior distributions, as is shown in the next theorem.

Theorem 7.3.3 Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the value of the mean θ is unknown and the value of the variance $\sigma^2 > 0$ is known. Suppose also that the prior distribution of θ is the normal distribution with mean μ_0 and variance v_0^2. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the normal distribution with mean μ_1 and variance v_1^2 where

$$\mu_1 = \frac{\sigma^2 \mu_0 + n v_0^2 \bar{x}_n}{\sigma^2 + n v_0^2} \tag{7.3.1}$$

and

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2}. \tag{7.3.2}$$

Proof The likelihood function. $f_n(\boldsymbol{x}|\theta)$ has the form

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right].$$

Here a constant factor has been dropped from the right side. The method of completing the square (see Exercise 24 in Sec. 5.6) tells us that

$$\sum_{i=1}^{n}(x_i-\theta)^2 = n(\theta-\overline{x}_n)^2 + \sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

By omitting a factor that involves $x_1, \ldots, x_n$ but does not depend on θ, we may rewrite $f_n(\boldsymbol{x}|\theta)$ in the following form:

$$f_n(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta-\overline{x}_n)^2\right].$$

Since the prior p.d.f. $\xi(\theta)$ has the form

$$\xi(\theta) \propto \exp\left[-\frac{1}{2v_0^2}(\theta-\mu_0)^2\right],$$

it follows that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ satisfies the relation

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\theta-\overline{x}_n)^2 + \frac{1}{v_0^2}(\theta-\mu_0)^2\right]\right\}.$$

If μ_1 and v_1^2 are as specified in Eqs. (7.3.1) and (7.3.2), completing the square again establishes the following identity:

$$\frac{n}{\sigma^2}(\theta-\overline{x}_n)^2 + \frac{1}{v_0^2}(\theta-\mu_0)^2 = \frac{1}{v_1^2}(\theta-\mu_1)^2 + \frac{n}{\sigma^2+nv_0^2}(\overline{x}_n-\mu_0)^2.$$

Since the final term on the right side of this equation does not involve θ, it can be absorbed in the proportionality factor, and we obtain the relation

$$\xi(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{1}{2v_1^2}(\theta-\mu_1)^2\right].$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the normal distribution with mean μ_1 and variance v_1^2. Therefore, the posterior distribution of θ is as specified in the theorem. ■

In Theorem 7.3.3, the numbers μ_0 and v_0^2 are the prior hyperparameters, while μ_1 and v_1^2 are the posterior hyperparameters. Notice that the statistic $\overline{X}_n$ is used in the construction of the posterior distribution, and hence will play a role in any inference based on the posterior.

Example 7.3.8 Automobile Emissions. We can apply Theorem 7.3.3 to answer the question at the end of Example 7.3.7. In the notation of the theorem, we have $n = 46$, $\sigma^2 = 0.5^2 = 0.25$,

$\mu_0 = 2$, and $v^2 = 1.0$. The average of the 46 measurements is $\bar{x}_n = 1.329$. The posterior distribution of θ is then the normal distribution with mean and variance given by

$$\mu_1 = \frac{0.25 \times 2 + 46 \times 1 \times 1.329}{0.25 + 46 \times 1} = 1.333,$$

$$v_1^2 = \frac{0.25 \times 1}{0.25 + 46 \times 1} = 0.0054. \quad \blacktriangleleft$$

The mean μ_1 of the posterior distribution of θ, as given in Eq. (7.3.1), can be rewritten as follows:

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + nv_0^2}\mu_0 + \frac{nv_0^2}{\sigma^2 + nv_0^2}\bar{x}_n. \tag{7.3.3}$$

It can be seen from Eq. (7.3.3) that μ_1 is a weighted average of the mean μ_0 of the prior distribution and the sample mean $\bar{x}_n$. Furthermore, it can be seen that the relative weight given to $\bar{x}_n$ satisfies the following three properties: (1) For fixed values of v_0^2 and σ^2, the larger the sample size n, the greater will be the relative weight that is given to $\bar{x}_n$. (2) For fixed values of v_0^2 and n, the larger the variance σ^2 of each observation in the sample, the smaller will be the relative weight that is given to $\bar{x}_n$. (3) For fixed values of σ^2 and n, the larger the variance v_0^2 of the prior distribution, the larger will be the relative weight that is given to $\bar{x}_n$.

Moreover, it can be seen from Eq. (7.3.2) that the variance v_1^2 of the posterior distribution of θ depends on the number n of observations that have been taken but does not depend on the magnitudes of the observed values. Suppose, therefore, that a random sample of n observations is to be taken from a normal distribution for which the value of the mean θ is unknown, the value of the variance is known, and the prior distribution of θ is a specified normal distribution. Then, before any observations have been taken, we can use Eq. (7.3.2) to calculate the actual value of the variance v_1^2 of the posterior distribution. However, the value of the mean μ_1 of the posterior distribution will depend on the observed values that are obtained in the sample. The fact that the variance of the posterior distribution depends only on the number of observations is due to the assumption that the variance σ^2 of the individual observations is known. In Sec. 8.6, we shall relax this assumption.

Example 7.3.9 The Variance of the Posterior Normal Distribution. Suppose that observations are to be taken at random from the normal distribution with mean θ and variance 1, and that θ is unknown. Assume that the prior distribution of θ is a normal distribution with variance 4. Also, observations are to be taken until the variance of the posterior distribution of θ has been reduced to the value 0.01 or less. We shall determine the number of observations that must be taken before the sampling process is stopped.

It follows from Eq. (7.3.2) that after n observations have been taken, the variance v_1^2 of the posterior distribution of θ will be

$$v_1^2 = \frac{4}{4n + 1}.$$

Therefore, the relation $v_1^2 \leq 0.01$ will be satisfied if and only if $n \geq 99.75$. Hence, the relation $v_1^2 \leq 0.01$ will be satisfied after 100 observations have been taken and not before then. $\blacktriangleleft$

Example 7.3.10 Calorie Counts on Food Labels. Allison, Heshka, Sepulveda, and Heymsfield (1993) sampled 20 nationally prepared foods and compared the stated calorie contents per

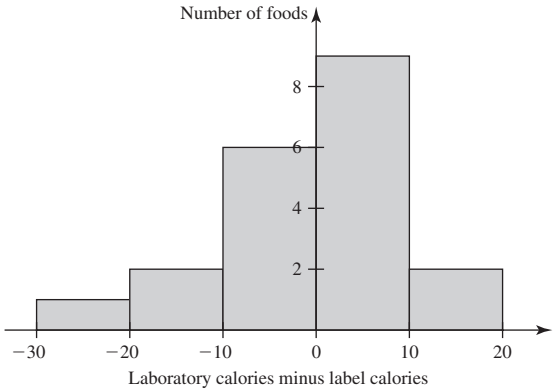

Figure 7.4 Histogram of percentage differences between observed and advertised calories in Example 7.3.10.

gram from the labels to calorie contents determined in the laboratory. Figure 7.4 is a histogram of the percentage differences between the observed laboratory calorie measurements and the advertised calorie contents on the labels of the foods. Suppose that we model the conditional distribution of the differences given θ as the normal distribution with mean θ and variance 100. (In this section, we assume that the variance is known. In Sec. 8.6, we will be able to deal with the case in which the mean and the variance are treated as random variables with a joint distribution.) We will use a prior distribution for θ that is the normal distribution with mean 0 and a variance of 60. The data $\boldsymbol{X}$ comprise the collection of 20 differences in Fig. 7.4, whose average is 0.125. The posterior distribution of θ would then be the normal distribution with mean

$$\mu_1 = \frac{100 \times 0 + 20 \times 60 \times 0.125}{100 + 20 \times 60} = 0.1154,$$

and variance

$$v_1^2 = \frac{100 \times 60}{100 + 20 \times 60} = 4.62.$$

For example, we might be interested in whether or not the packagers are systematically understating the calories in their food by at least 1 percent. This would correspond to $\theta > 1$. Using Theorem 5.6.6, we can find

$$\Pr(\theta > 1|\boldsymbol{x}) = 1 - \Phi\left(\frac{1 - 0.1154)}{\sqrt{4.62}}\right) = 1 - \Phi(1.12) = 0.3403.$$

There is a nonnegligible, but not overwhelming, chance that the packagers are shaving a percent or more off of their labels. ◀

Sampling from an Exponential Distribution

Example 7.3.11

Lifetimes of Electronic Components. In Example 7.2.1, suppose that we observe the lifetimes of three components, $X_1 = 3$, $X_2 = 1.5$, and $X_3 = 2.1$. These were modeled as i.i.d. exponential random variables given θ. Our prior distribution for θ was the gamma distribution with parameters 1 and 2. What is the posterior distribution of θ given these observed lifetimes? ◀

When sampling from an exponential distribution for which the value of the parameter θ is unknown, the family of gamma distributions serves as a conjugate family of prior distributions, as shown in the next theorem.

Theorem 7.3.4 Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter $\theta > 0$ that is unknown. Suppose also that the prior distribution of θ is the gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. Then the posterior distribution of θ given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the gamma distribution with parameters $\alpha + n$ and $\beta + \sum_{i=1}^{n} x_i$.

Proof Again, let $y = \sum_{i=1}^{n} x_i$. Then the likelihood function $f_n(\boldsymbol{x}|\theta)$ is

$$f_n(\boldsymbol{x}|\theta) = \theta^n e^{-\theta y}.$$

Also, the prior p.d.f. $\xi(\theta)$ has the form

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for } \theta > 0.$$

It follows, therefore, that the posterior p.d.f. $\xi(\theta|\boldsymbol{x})$ has the form

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{\alpha+n-1} e^{-(\beta+y)\theta} \quad \text{for } \theta > 0.$$

The right side of this relation can be recognized as being, except for a constant factor, the p.d.f. of the gamma distribution with parameters $\alpha + n$ and $\beta + y$. Therefore, the posterior distribution of θ is as specified in the theorem. ■

The posterior distribution of θ in Theorem 7.3.4 depends on the observed value of the statistic $Y = \sum_{i=1}^{n} X_i$; hence, every inference about θ based on the posterior distribution will depend on the observed value of Y.

Example 7.3.12 Lifetimes of Electronic Components. In Example 7.3.11, we can apply Theorem 7.3.4 to find the posterior distribution. In the notation of the theorem and its proof, we have $n = 3$, $\alpha = 1$, $\beta = 2$, and

$$y = \sum_{i=1}^{n} x_i = 3 + 1.5 + 2.1 = 6.6.$$

The posterior distribution of θ is then the gamma distribution with parameters $\alpha = 1 + 3 = 4$ and $\beta = 2 + 6.6 = 8.6$. ◀

The reader should note that Theorem 7.3.4 would have greatly shortened the derivation of the posterior distribution in Example 7.2.6.

Improper Prior Distributions

In Sec. 7.2, we mentioned improper priors as expedients that try to capture the idea that there is much more information in the data than is captured in our prior distribution. Each of the conjugate families that we have seen in this section has an improper prior as a limiting case.

Example 7.3.13 A Clinical Trial. What we illustrate here will apply to all examples in which the data comprise a conditionally i.i.d. sample (given θ) from the Bernoulli distribution with parameter θ. Consider the subjects in the imipramine group in Example 2.1.4. The proportion of successes among all patients who might get imipramine had been called P in earlier examples, but let us call it θ this time in keeping with the general notation

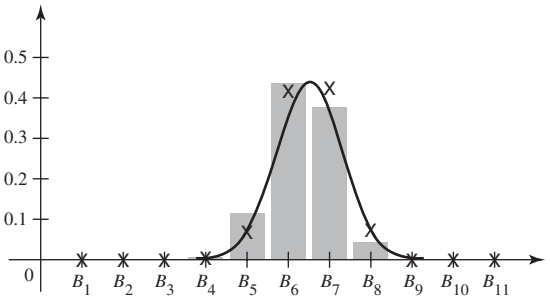

Figure 7.5 The posterior probabilities from Examples 2.3.7 (X) and 2.3.8 (bars) together with the posterior p.d.f. from Example 7.3.13 (solid line).

of this chapter. Suppose that θ has the beta distribution with parameters α and β, a general conjugate prior. There are $n = 40$ patients in the imipramine group, and 22 of them are successes. The posterior distribution of θ is the beta distribution with parameters $\alpha + 22$ and $\beta + 18$, as we saw in Theorem 7.3.1. The mean of the posterior distribution is $(\alpha + 22)/(\alpha + \beta + 40)$. If α and β are small, then the posterior mean is close to 22/40, which is the observed proportion of successes. Indeed, if $\alpha = \beta = 0$, which does not correspond to a real beta distribution, then the posterior mean is exactly 22/40. However, we can look at what happens as α and β get close to 0. The beta p.d.f. (ignoring the constant factor) is $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. We can set $\alpha = \beta = 0$ and pretend that $\xi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is the prior p.d.f. of θ. The likelihood function is $f_{40}(\boldsymbol{x}|\theta) = \binom{40}{22}\theta^{22}(1-\theta)^{18}$. We can ignore the constant factor $\binom{40}{22}$ and obtain the product

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{21}(1-\theta)^{17}, \quad \text{for } 0 < \theta < 1.$$

This is easily recognized as being the same as the p.d.f. of the beta distribution with parameters 22 and 18 except for a constant factor. So, if we use the improper "beta distribution" prior with prior hyperparameters 0 and 0, we get the beta posterior distribution for θ with posterior hyperparameters 22 and 18. Notice that Theorem 7.3.1 yields the correct posterior distribution even in this improper prior case. Figure 7.5 adds the p.d.f. of the posterior beta distribution calculated here to Fig. 2.4 which depicted the posterior probabilities for two different discrete prior distributions. All three posteriors are pretty close. ◀

Definition 7.3.2

Improper Prior. Let ξ be a nonnegative function whose domain includes the parameter space of a statistical model. Suppose that $\int \xi(\theta)d\theta = \infty$. If we pretend as if $\xi(\theta)$ is the prior p.d.f. of θ, then we are using an *improper prior* for θ.

Definition 7.3.2 is not of much use in determining an improper prior to use in a particular application. There are many methods for choosing an improper prior, and the hope is that they all lead to similar posterior distributions so that it does not much matter which of them one chooses. The most straightforward method for choosing an improper prior is to start with the family of conjugate prior distributions, if there is such a family. In most cases, if the parameterization of the conjugate family (prior hyperparameters) is chosen carefully, the posterior hyperparameters will each equal the corresponding prior hyperparameter plus a statistic. One would then replace each of those prior hyperparameters by 0 in the formula for the prior p.d.f. This generally results in a function that satisfies Definition 7.3.2. In Example 7.3.13, each of the posterior hyperparameters were equal to the corresponding prior hyperparameters plus some statistic. In that example, we replaced both prior hyperparameters by 0 to obtain the improper prior. Here are some more examples. The method just

described needs to be modified if one chooses an "inconvenient" parameterization of the conjugate prior, as in Example 7.3.15 below.

Example 7.3.14

Prussian Army Deaths. Bortkiewicz (1898) counted the numbers of Prussian soldiers killed by horsekick (a more serious problem in the nineteenth century than it is today) in 14 army units for each of 20 years, a total of 280 counts. The 280 counts have the following values: 144 counts are 0, 91 counts are 1, 32 counts are 2, 11 counts are 3, and 2 counts are 4. No unit suffered more than four deaths by horsekick during any one year. (These data were reported and analyzed by Winsor, 1947.) Suppose that we were going to model the 280 counts as a random sample of Poisson random variables $X_1, \ldots, X_{280}$ with mean θ conditional on the parameter θ. A conjugate prior would be a member of the gamma family with prior hyperparameters α and β. Theorem 7.3.2 says that the posterior distribution of θ would be the gamma distribution with posterior hyperparameters $\alpha + 196$ and $\beta + 280$, since the sum of the 280 counts equals 196. Unless either α or β is very large, the posterior gamma distribution is nearly the same as the gamma distribution with posterior hyperparameters 196 and 280. This posterior distribution would seem to be the result of using a conjugate prior with prior hyperparameters 0 and 0. Ignoring the constant factor, the p.d.f. of the gamma distribution with parameters α and β is $\theta^{\alpha-1}e^{\beta\theta}$ for $\theta > 0$. If we let $\alpha = 0$ and $\beta = 0$ in this formula, we get the improper prior "p.d.f." $\xi(\theta) = \theta^{-1}$ for $\theta > 0$. Pretending as if this really were a prior p.d.f. and applying Bayes' theorem for random variables (Theorem 3.6.4) would yield

$$\xi(\theta|\boldsymbol{x}) \propto \theta^{195}e^{-280\theta}, \quad \text{for } \theta > 0.$$

This is easily recognized as being the p.d.f. of the gamma distribution with parameters 196 and 280, except for a constant factor. The result in this example applies to all cases in which we model data with Poisson distributions. The improper "gamma distribution" with prior hyperparameters 0 and 0 can be used in Theorem 7.3.2, and the conclusion will still hold. ◀

Example 7.3.15

Failure Times of Ball Bearings. Suppose that we model the 23 logarithms of failure times of ball bearings from Example 5.6.9 as normal random variables $X_1, \ldots, X_{23}$ with mean θ and variance 0.25. A conjugate prior for θ would be the normal distribution with mean μ_0 and variance v_0^2 for some μ_0 and v_0^2. The average of the 23 log-failure times is 4.15, so the posterior distribution of θ would be the normal distribution with mean $\mu_1 = (0.25\mu_0 + 23 \times 4.15v_0^2)/(0.25 + 23v_0^2)$ and variance $v_1^2 = (0.25v_0^2)/(0.25 + 23v_0^2)$. If we let $v_0^2 \to \infty$ in the formulas for μ_1 and v_1^2, we get $\mu_1 \to 4.15$ and $v_1^2 \to 0.25/23$. Having infinite variance for the prior distribution of θ is like saying that θ is equally likely to be anywhere on the real number line. This same thing happens in every example in which we model data $X_1, \ldots, X_n$ as a random sample from the normal distribution with mean θ and known variance σ^2 conditional on θ. If we use an improper "normal distribution" prior with variance ∞ (the prior mean does not matter), the calculation in Theorem 7.3.3 would yield a posterior distribution that is the normal distribution with mean $\overline{x}_n$ and variance σ^2/n. The improper prior "p.d.f." in this case is $\xi(\theta)$ equal to a constant.

This example would be an application of the method described after Definition 7.3.2 if we had described the conjugate prior distribution in terms of the following "more convenient" hyperparameters: 1 over the variance $u_0 = 1/v_0^2$ and the mean over the variance $t_0 = \mu_0/v_0^2$. In terms of these hyperparameters, the posterior distribution has 1 over its variance equal to $u_1 = u_0 + n/0.25$ and mean over variance equal to $t_1 = \mu_1/v_1^2 = t_0 + 23 \times 4.15/0.25$. Each of u_1 and t_1 has the form of the cor-

responding prior hyperparamter plus a statistic. The improper prior with $u_0 = t_0 = 0$ also has $\xi(\theta)$ equal to a constant. ◀

There are improper priors for other sampling models, also. The reader can verify (in Exercise 21) that the "gamma distribution" with parameters 0 and 0 leads to results similar to those in Example 7.3.14 when the data are a random sample from an exponential distribution. Exercises 23 and 24 introduce a general collection of p.d.f.'s $f(x|\theta)$ for which it is easy to construct improper priors.

Improper priors were introduced for cases in which the observed data contain much more information than is represented by our prior distribution. Implicitly, we are assuming that the data are rather informative. When the data do not contain much information, improper priors may be higly inappropriate.

Example 7.3.16

Very Rare Events. In Example 5.4.7, we discussed a drinking water contaminant known as cryptosporidium that generally occurs in very low concentrations. Suppose that a water authority models the oocysts of cryptosporidium in the water supply as a Poisson process with rate of θ oocysts per liter. They decide to sample 25 liters of water to learn about θ. Suppose that they use the improper gamma prior with "p.d.f." θ^{-1}. (This is the same improper prior used in Example 7.3.14.) If the 25-liter sample contains no oocysts, the water authority would be led to a posterior distribution for θ that was the gamma distribution with parameters 0 and 5, which is not a real distribution. No matter how many liters are sampled, the posterior distribution will not be a real distribution until at least one oocyst is observed. When sampling for rare events, one might be forced to quantify prior information in the form a proper prior distribution in order to be able to make inferences based on the posterior distribution. ◀

Summary

For each of several different statistical models for data given the parameter, we found a conjugate family of distributions for the parameter. These families have the property that if the prior distribution is chosen from the family, then the posterior distribution is a member of the family. For data with distributions related to the Bernoulli, such as binomial, geometric, and negative binomial, the conjugate family for the success probability parameter is the family of beta distributions. For data with distributions related to the Poisson process, such as Poisson, gamma (with known first parameter), and exponential, the conjugate family for the rate parameter is the family of gamma distributions. For data having a normal distribution with known variance, the conjugate family for the mean is the normal family. We also described the use of improper priors. Improper priors are not true probability distributions, but if we pretend that they are, we will compute posterior distributions that approximate the posteriors that we would have obtained using proper conjugate priors with extreme values of the prior hyperparameters.

Exercises

1. Consider again the situation described in Example 7.3.10. Once again, suppose that the prior distribution of θ is a normal distribution with mean 0, but this time let the prior variance be $v^2 > 0$. If the posterior mean of θ is 0.12, what value of v^2 was used?

2. Show that in Example 7.3.2 it must be true that $V \leq 0.01$ after 22 items have been selected. Also show that $V > 0.01$ until at least seven items have been selected.

3. Suppose that the proportion θ of defective items in a large shipment is unknown and that the prior distribution

of θ is the beta distribution with parameters 2 and 200. If 100 items are selected at random from the shipment and if three of these items are found to be defective, what is the posterior distribution of θ?

4. Consider again the conditions of Exercise 3. Suppose that after a certain statistician has observed that there were three defective items among the 100 items selected at random, the posterior distribution that she assigns to θ is a beta distribution for which the mean is $2/51$ and the variance is $98/[(51)^2(103)]$. What prior distribution had the statistician assigned to θ?

5. Suppose that the number of defects in a 1200-foot roll of magnetic recording tape has a Poisson distribution for which the value of the mean θ is unknown and that the prior distribution of θ is the gamma distribution with parameters $\alpha = 3$ and $\beta = 1$. When five rolls of this tape are selected at random and inspected, the numbers of defects found on the rolls are 2, 2, 6, 0, and 3. Determine the posterior distribution of θ.

6. Let θ denote the average number of defects per 100 feet of a certain type of magnetic tape. Suppose that the value of θ is unknown and that the prior distribution of θ is the gamma distribution with parameters $\alpha = 2$ and $\beta = 10$. When a 1200-foot roll of this tape is inspected, exactly four defects are found. Determine the posterior distribution of θ.

7. Suppose that the heights of the individuals in a certain population have a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2 inches. Suppose also that the prior distribution of θ is a normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. If 10 people are selected at random from the population, and their average height is found to be 69.5 inches, what is the posterior distribution of θ?

8. Consider again the problem described in Exercise 7.

- **a.** Which interval 1-inch long had the highest prior probability of containing the value of θ?
- **b.** Which interval 1-inch long has the highest posterior probability of containing the value of θ?
- **c.** Find the values of the probabilities in parts (a) and (b).

9. Suppose that a random sample of 20 observations is taken from a normal distribution for which the value of the mean θ is unknown and the variance is 1. After the sample values have been observed, it is found that $\overline{X}_n = 10$, and that the posterior distribution of θ is a normal distribution for which the mean is 8 and the variance is $1/25$. What was the prior distribution of θ?

10. Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution for which the standard deviation is 1. What is the smallest number of observations that must be included in the sample in order to reduce the standard deviation of the posterior distribution of θ to the value 0.1?

11. Suppose that a random sample of 100 observations is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, and the prior distribution of θ is a normal distribution. Show that no matter how large the standard deviation of the prior distribution is, the standard deviation of the posterior distribution will be less than 1/5.

12. Suppose that the time in minutes required to serve a customer at a certain facility has an exponential distribution for which the value of the parameter θ is unknown and that the prior distribution of θ is a gamma distribution for which the mean is 0.2 and the standard deviation is 1. If the average time required to serve a random sample of 20 customers is observed to be 3.8 minutes, what is the posterior distribution of θ?

13. For a distribution with mean $\mu \neq 0$ and standard deviation $\sigma > 0$, the *coefficient of variation* of the distribution is defined as $\sigma/|\mu|$. Consider again the problem described in Exercise 12, and suppose that the coefficient of variation of the prior gamma distribution of θ is 2. What is the smallest number of customers that must be observed in order to reduce the coefficient of variation of the posterior distribution to 0.1?

14. Show that the family of beta distributions is a conjugate family of prior distributions for samples from a negative binomial distribution with a known value of the parameter r and an unknown value of the parameter p $(0 < p < 1)$.

15. Let $\xi(\theta)$ be a p.d.f. that is defined as follows for constants $\alpha > 0$ and $\beta > 0$:

$$\xi(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta} & \text{for } \theta > 0, \\ 0 & \text{for } \theta \leq 0. \end{cases}$$

A distribution with this p.d.f. is called an *inverse gamma distribution*.

- **a.** Verify that $\xi(\theta)$ is actually a p.d.f. by verifying that $\int_0^\infty \xi(\theta)\, d\theta = 1$.
- **b.** Consider the family of probability distributions that can be represented by a p.d.f. $\xi(\theta)$ having the given form for all possible pairs of constants $\alpha > 0$ and $\beta > 0$. Show that this family is a conjugate family of prior distributions for samples from a normal distribution with a known value of the mean μ and an unknown value of the variance θ.

16. Suppose that in Exercise 15 the parameter is taken as the standard deviation of the normal distribution, rather than the variance. Determine a conjugate family of prior distributions for samples from a normal distribution with

a known value of the mean μ and an unknown value of the standard deviation σ.

17. Suppose that the number of minutes a person must wait for a bus each morning has the uniform distribution on the interval $[0, \theta]$, where the value of the endpoint θ is unknown. Suppose also that the prior p.d.f. of θ is as follows:

$$\xi(\theta) = \begin{cases} \frac{192}{\theta^4} & \text{for } \theta \geq 4, \\ 0 & \text{otherwise.} \end{cases}$$

If the observed waiting times on three successive mornings are 5, 3, and 8 minutes, what is the posterior p.d.f. of θ?

18. The Pareto distribution with parameters x_0 and α ($x_0 > 0$ and $\alpha > 0$) is defined in Exercise 16 of Sec. 5.7. Show that the family of Pareto distributions is a conjugate family of prior distributions for samples from a uniform distribution on the interval $[0, \theta]$, where the value of the endpoint θ is unknown.

19. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the value of the parameter θ is unknown ($\theta > 0$), and the prior distribution of θ is the gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$). Determine the mean and the variance of the posterior distribution of θ.

20. Suppose that we model the lifetimes (in months) of electronic components as independent exponential random variables with unknown parameter β. We model β as having the gamma distribution with parameters a and b. We believe that the mean lifetime is four months before we see any data. If we were to observe 10 components with an average observed lifetime of six months, we would then claim that the mean lifetime is five months. Determine a and b. *Hint:* Use Exercise 21 in Sec. 5.7.

21. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter θ. Let the prior distribution of θ be improper with "p.d.f." $1/\theta$ for $\theta > 0$. Find the posterior distribution of θ and show that the posterior mean of θ is $1/\overline{x}_n$.

22. Consider the data in Example 7.3.10. This time, suppose that we use the improper prior "p.d.f." $\xi(\theta) = 1$ (for all θ). Find the posterior distribution of θ and the posterior probability that $\theta > 1$.

23. Consider a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where θ belongs to some parameter space Ω. It is said that the family of distributions obtained by letting θ vary over all values in Ω is an *exponential family*, or a *Koopman-Darmois family*, if $f(x|\theta)$ can be written as follows for $\theta \in \Omega$ and all values of x:

$$f(x|\theta) = a(\theta)b(x)\exp[c(\theta)\, d(x)].$$

Here $a(\theta)$ and $c(\theta)$ are arbitrary functions of θ, and $b(x)$ and $d(x)$ are arbitrary functions of x. Let

$$H = \left\{ (\alpha, \beta) : \int_\Omega a(\theta)^\alpha \exp[c(\theta)\, \beta] d\theta < \infty \right\}.$$

For each $(\alpha, \beta) \in H$, let

$$\xi_{\alpha,\beta}(\theta) = \frac{a(\theta)^\alpha \exp[c(\theta)\, \beta]}{\int_\Omega a(\eta)^\alpha \exp[c(\eta)\, \beta] d\eta},$$

and let Ψ be the set of all probability distributions that have p.d.f.'s of the form $\xi_{\alpha,\beta}(\theta)$ for some $(\alpha, \beta) \in H$.

a. Show that Ψ is a conjugate family of prior distributions for samples from $f(x|\theta)$.

b. Suppose that we observe a random sample of size n from the distribution with p.d.f. $f(x|\theta)$. If the prior p.d.f. of θ is ξ_{α_0,β_0}, show that the posterior hyperparameters are

$$\alpha_1 = \alpha_0 + n, \;\; \beta_1 = \beta_0 + \sum_{i=1}^{n} d(x_i).$$

24. Show that each of the following families of distributions is an exponential family, as defined in Exercise 23:

a. The family of Bernoulli distributions with an unknown value of the parameter p

b. The family of Poisson distributions with an unknown mean

c. The family of negative binomial distributions for which the value of r is known and the value of p is unknown

d. The family of normal distributions with an unknown mean and a known variance

e. The family of normal distributions with an unknown variance and a known mean

f. The family of gamma distributions for which the value of α is unknown and the value of β is known

g. The family of gamma distributions for which the value of α is known and the value of β is unknown

h. The family of beta distributions for which the value of α is unknown and the value of β is known

i. The family of beta distributions for which the value of α is known and the value of β is unknown

25. Show that the family of uniform distributions on the intervals $[0, \theta]$ for $\theta > 0$ is *not* an exponential family as defined in Exercise 23. *Hint:* Look at the support of each uniform distribution.

26. Show that the family of discrete uniform distributions on the sets of integers $\{0, 1, \ldots, \theta\}$ for θ a nonnegative integer is *not* an exponential family as defined in Exercise 23.

7.4 Bayes Estimators

An estimator of a parameter is some function of the data that we hope is close to the parameter. A Bayes estimator is an estimator that is chosen to minimize the posterior mean of some measure of how far the estimator is from the parameter, such as squared error or absolute error.

Nature of an Estimation Problem

Example 7.4.1 Calorie Counts on Food Labels. In Example 7.3.10, we found the posterior distribution of θ, the mean percentage difference between measured and advertised calorie counts. A consumer group might wish to report a single number as an estimate of θ without specifying the entire distribution for θ. How to choose such a single-number estimate in general is the subject of this section. ◀

We begin with a definition that is appropriate for a real-valued parameter such as in Example 7.4.1. A more general definition will follow after we become more familiar with the concept of estimation.

Definition 7.4.1 Estimator/Estimate. Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of the real line. An *estimator* of the parameter θ is a real-valued function $\delta(X_1, \ldots, X_n)$. If $X_1 = x_1, \ldots, X_n = x_n$ are observed, then $\delta(x_1, \ldots, x_n)$ is called the *estimate* of θ.

Notice that every estimator is, by nature of being a function of data, a statistic in the sense of Definition 7.1.4.

Because the value of θ must belong to the set Ω, it might seem reasonable to require that every possible value of an estimator $\delta(X_1, \ldots, X_n)$ must also belong to Ω. We shall not require this restriction, however. If an estimator can take values outside of the parameter space Ω, the experimenter will need to decide in the specific problem whether that seems appropriate or not. It may turn out that every estimator that takes values only inside Ω has other even less desirable properties.

In Definition 7.4.1, we distinguished between the terms *estimator* and *estimate*. Because an estimator $\delta(X_1, \ldots, X_n)$ is a function of the random variables $X_1, \ldots, X_n$, the estimator itself is a random variable, and its probability distribution can be derived from the joint distribution of $X_1, \ldots, X_n$, if desired. On the other hand, an *estimate* is a specific value $\delta(x_1, \ldots, x_n)$ of the estimator that is determined by using specific observed values $x_1, \ldots, x_n$. If we use the vector notation $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$, then an estimator is a function $\delta(\boldsymbol{X})$ of the random vector $\boldsymbol{X}$, and an estimate is a specific value $\delta(\boldsymbol{x})$. It will often be convenient to denote an estimator $\delta(\boldsymbol{X})$ simply by the symbol δ.

Loss Functions

Example 7.4.2 Calorie Counts on Food Labels. In Example 7.4.1, the consumer group may feel that the farther their estimate $\delta(\boldsymbol{x})$ is from the true mean difference θ, the more embarassment and possible legal action they will encounter. Ideally, they would like to quantify the amount of negative repercussions as a function of θ and the estimate $\delta(\boldsymbol{x})$. Then they could have some idea how likely it is that they will encounter various levels of hassle as a result of their estimation. ◀

The foremost requirement of a good estimator δ is that it yield an estimate of θ that is close to the actual value of θ. In other words, a good estimator is one for which it is highly probable that the error $\delta(\boldsymbol{X}) - \theta$ will be close to 0. We shall assume that for each possible value of $\theta \in \Omega$ and each possible estimate a, there is a number $L(\theta, a)$ that measures the loss or cost to the statistician when the true value of the parameter is θ and her estimate is a. Typically, the greater the distance between a and θ, the larger will be the value of $L(\theta, a)$.

Definition 7.4.2 Loss Function. A *loss function* is a real-valued function of two variables, $L(\theta, a)$, where $\theta \in \Omega$ and a is a real number. The interpretation is that the statistician loses $L(\theta, a)$ if the parameter equals θ and the estimate equals a.

As before, let $\xi(\theta)$ denote the prior p.d.f. of θ on the set Ω, and consider a problem in which the statistician must estimate the value of θ without being able to observe the values in a random sample. If the statistician chooses a particular estimate a, then her expected loss will be

$$E[L(\theta, a)] = \int_{\Omega} L(\theta, a)\xi(\theta)\, d\theta. \tag{7.4.1}$$

We shall assume that the statistician wishes to choose an estimate a for which the expected loss in Eq. (7.4.1) is a minimum.

Definition of a Bayes Estimator

Suppose now that the statistician can observe the value $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$ before estimating θ, and let $\xi(\theta|\boldsymbol{x})$ denote the posterior p.d.f. of θ on Ω. (The case of a discrete parameter can be handled in similar fashion.) For each estimate a that the statistician might use, her expected loss in this case will be

$$E[L(\theta, a)|\boldsymbol{x}] = \int_{\Omega} L(\theta, a)\xi(\theta|\boldsymbol{x})\, d\theta. \tag{7.4.2}$$

Hence, the statistician should now choose an estimate a for which the expectation in Eq. (7.4.2) is a minimum.

For each possible value $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$, let $\delta^*(\boldsymbol{x})$ denote a value of the estimate a for which the expected loss in Eq. (7.4.2) is a minimum. Then the function $\delta^*(\boldsymbol{X})$ for which the values are specified in this way will be an estimator of θ.

Definition 7.4.3 Bayes Estimator/Estimate. Let $L(\theta, a)$ be a loss function. For each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, let $\delta^*(\boldsymbol{x})$ be a value of a such that $E[L(\theta, a)|\boldsymbol{x}]$ is minimized. Then δ^* is called a *Bayes estimator* of θ. Once $\boldsymbol{X} = \boldsymbol{x}$ is observed, $\delta^*(\boldsymbol{x})$ is called a *Bayes estimate* of θ.

Another way to describe a Bayes estimator δ^* is to note that, for each possible value $\boldsymbol{x}$ of $\boldsymbol{X}$, the value $\delta^*(\boldsymbol{x})$ is chosen so that

$$E[L(\theta, \delta^*(\boldsymbol{x}))|\boldsymbol{x}] = \min_{\text{All } a} E[L(\theta, a)|\boldsymbol{x}]. \tag{7.4.3}$$

In summary, we have considered an estimation problem in which a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ is to be taken from a distribution involving a parameter θ that has an unknown value in some specified set Ω. For every given loss function $L(\theta, a)$ and every prior p.d.f. $\xi(\theta)$, the Bayes estimator of θ is the estimator $\delta^*(\boldsymbol{X})$ for which Eq. (7.4.3) is satisfied for every possible value $\boldsymbol{x}$ of $\boldsymbol{X}$. It should be emphasized that the form of the Bayes estimator will depend on both the loss function that is used

in the problem and the prior distribution that is assigned to θ. In the problems described in this text, Bayes estimators will exist. However, there are more complicated situations in which no function δ^* satisfies (7.4.3).

Different Loss Functions

By far, the most commonly used loss function in estimation problems is the squared error loss function.

Definition 7.4.4 Squared Error Loss Function. The loss function

$$L(\theta, a) = (\theta - a)^2 \tag{7.4.4}$$

is called *squared error loss*.

When the squared error loss function is used, the Bayes estimate $\delta^*(\boldsymbol{x})$ for each observed value of $\boldsymbol{x}$ will be the value of a for which the expectation $E[(\theta - a)^2|\boldsymbol{x}]$ is a minimum. Theorem 4.7.3 states that, when the expectation of $(\theta - a)^2$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be equal to the mean $E(\theta|\boldsymbol{x})$ of the posterior distribution, if that posterior mean is finite. If the posterior mean of θ is not finite, then the expected loss is infinite for every possible estimate a. Hence, we have the following corollary to Theorem 4.7.3.

Corollary 7.4.1 Let θ be a real-valued parameter. Suppose that the squared error loss function (7.4.4) is used and that the posterior mean of θ, $E(\theta|\boldsymbol{X})$, is finite. Then, a Bayes estimator of θ is $\delta^*(\boldsymbol{X}) = E(\theta|\boldsymbol{X})$. ■

Example 7.4.3 Estimating the Parameter of a Bernoulli Distribution. Let the random sample $X_1, \ldots, X_n$ be taken from the Bernoulli distribution with parameter θ, which is unknown and must be estimated. Let the prior distribution of θ be the beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Suppose that the squared error loss function is used, as specified by Eq. (7.4.4), for $0 < \theta < 1$ and $0 < a < 1$. We shall determine the Bayes estimator of θ.

For observed values $x_1, \ldots, x_n$, let $y = \sum_{i=1}^n x_i$. Then it follows from Theorem 7.3.1 that the posterior distribution of θ will be the beta distribution with parameters $\alpha_1 = \alpha + y$ and $\beta_1 = \beta + n - y$. Since the mean of the beta distribution with parameters α_1 and β_1 is $\alpha_1/(\alpha_1 + \beta_1)$, the mean of this posterior distribution of θ will be $(\alpha + y)/(\alpha + \beta + n)$. The Bayes estimate $\delta(\boldsymbol{x})$ will be equal to this value for each observed vector $\boldsymbol{x}$. Therefore, the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified as follows:

$$\delta^*(\boldsymbol{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}. \tag{7.4.5}$$

◀

Example 7.4.4 Estimating the Mean of a Normal Distribution. Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from a normal distribution for which the value of the mean θ is unknown and the value of the variance σ^2 is known. Suppose also that the prior distribution of θ is the normal distribution with mean μ_0 and variance v_0^2. Suppose, finally, that the squared error loss function is to be used, as specified in Eq. (7.4.4), for $-\infty < \theta < \infty$ and $-\infty < a < \infty$. We shall determine the Bayes estimator of θ.

It follows from Theorem 7.3.3 that for all observed values $x_1, \ldots, x_n$, the posterior distribution of θ will be a normal distribution with mean μ_1 specified by

Eq. (7.3.1). Therefore, the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified as follows:

$$\delta^*(\boldsymbol{X}) = \frac{\sigma^2\mu_0 + nv_0^2\overline{X}_n}{\sigma^2 + nv_0^2}. \tag{7.4.6}$$

The posterior variance of θ does not enter into this calculation. ◀

Another commonly used loss function in estimation problems is the absolute error loss function.

Definition 7.4.5 *Absolute Error Loss Function.* The loss function

$$L(\theta, a) = |\theta - a| \tag{7.4.7}$$

is called *absolute error loss*.

For every observed value of $\boldsymbol{x}$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will now be the value of a for which the expectation $E(|\theta - a||\boldsymbol{x})$ is a minimum. It was shown in Theorem 4.5.3 that for every given probability distribution of θ, the expectation of $|\theta - a|$ will be a minimum when a is chosen to be equal to a median of the distribution of θ. Therefore, when the expectation of $|\theta - a|$ is calculated with respect to the posterior distribution of θ, this expectation will be a minimum when a is chosen to be a median of the posterior distribution of θ.

Corollary 7.4.2 When the absolute error loss function (7.4.7) is used, a Bayes estimator of a real-valued parameter is $\delta^*(\boldsymbol{X})$ equal to a median of the posterior distribution of θ.

We shall now reconsider Examples 7.4.3 and 7.4.4, but we shall use the absolute error loss function instead of the squared error loss function.

Example 7.4.5 *Estimating the Parameter of a Bernoulli Distribution.* Consider again the conditions of Example 7.4.3, but suppose now that the absolute error loss function is used, as specified by Eq. (7.4.7). For all observed values $x_1, \ldots, x_n$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will be equal to the median of the posterior distribution of θ, which is the beta distribution with parameters $\alpha + y$ and $\beta + n - y$. There is no simple expression for this median. It must be determined by numerical approximations for each given set of observed values. Most statistical computer software can compute the median of an arbitrary beta distribution.

As a specific example, consider the situation described in Example 7.3.13 in which an improper prior was used. The posterior distribution of θ in that example was the beta distribution with parameters 22 and 18. The mean of this beta distribution is $22/40 = 0.55$. The median is 0.5508. ◀

Example 7.4.6 *Estimating the Mean of a Normal Distribution.* Consider again the conditions of Example 7.4.4, but suppose now that the absolute error loss function is used, as specified by Eq. (7.4.7). For all observed values $x_1, \ldots, x_n$, the Bayes estimate $\delta^*(\boldsymbol{x})$ will be equal to the median of the posterior normal distribution of θ. However, since the mean and the median of each normal distribution are equal, $\delta^*(\boldsymbol{x})$ is also equal to the mean of the posterior distribution. Therefore, the Bayes estimator with respect to the absolute error loss function is the same as the Bayes estimator with respect to the squared error loss function, and it is again given by Eq. (7.4.6). ◀

Other Loss Functions Although the squared error loss function and, to a lesser extent, the absolute error loss function are the most commonly used ones in estimation problems, neither of these loss functions may be appropriate in a particular problem. In some problems, it might be appropriate to use a loss function having the form $L(\theta, a) = |\theta - a|^k$, where k is some positive number other than 1 or 2. In other problems, the loss that results when the error $|\theta - a|$ has a given magnitude might depend on the actual value of θ. In such a problem, it might be appropriate to use a loss function having the form $L(\theta, a) = \lambda(\theta)(\theta - a)^2$ or $L(\theta, a) = \lambda(\theta)|\theta - a|$, where $\lambda(\theta)$ is a given positive function of θ. In still other problems, it might be more costly to overestimate the value of θ by a certain amount than to underestimate it by the same amount. One specific loss function that reflects this property is as follows:

$$L(\theta, a) = \begin{cases} 3(\theta - a)^2 & \text{for } \theta \leq a, \\ (\theta - a)^2 & \text{for } \theta > a. \end{cases}$$

Various other types of loss functions might be relevant in specific estimation problems. However, in this book we shall give most of our attention to the squared error and absolute error loss functions.

The Bayes Estimate for Large Samples

Effect of Different Prior Distributions Suppose that the proportion θ of defective items in a large shipment is unknown and that the prior distribution of θ is the uniform distribution on the interval [0, 1]. Suppose also that the value of θ must be estimated, and that the squared error loss function is used. Suppose, finally, that in a random sample of 100 items from the shipment, exactly 10 items are found to be defective. Since the uniform distribution is the beta distribution with parameters $\alpha = 1$ and $\beta = 1$, and since $n = 100$ and $y = 10$ for the given sample, it follows from Eq. (7.4.5) that the Bayes estimate is $\delta^*(\boldsymbol{x}) = 11/102 = 0.108$.

Next, suppose that the prior p.d.f. of θ has the form $\xi(\theta) = 2(1 - \theta)$ for $0 < \theta < 1$, instead of being a uniform distribution, and that again in a random sample of 100 items, exactly 10 items are found to be defective. Since $\xi(\theta)$ is the p.d.f. of the beta distribution with parameters $\alpha = 1$ and $\beta = 2$, it follows from Eq. (7.4.5) that in this case the Bayes estimate of θ is $\delta(\boldsymbol{x}) = 11/103 = 0.107$.

The two prior distributions considered here are quite different. The mean of the uniform prior distribution is 1/2, and the mean of the other beta prior distribution is 1/3. Nevertheless, because the number of observations in the sample is so large ($n = 100$), the Bayes estimates with respect to the two different prior distributions are almost the same. Furthermore, the values of both estimates are very close to the observed proportion of defective items in the sample, which is $\overline{x}_n = 0.1$.

Example 7.4.7

Chest Measurements of Scottish Soldiers. Quetelet (1846) reported (with some errors) data on the chest measurements (in inches) of 5732 Scottish militiamen. These data appeared earlier in an 1817 medical journal and are discussed by Stigler (1986). Figure 7.6 shows a histogram of the data. Suppose that we were to model the individual chest measurements as a random sample (given θ) of normal random variables with mean θ and variance 4. The average chest measurement is $\overline{x}_n = 39.85$. If θ had the normal prior distribution with mean μ_0 and variance v_0^2, then using Eq. (7.3.1) the posterior distribution of θ would be normal with mean

$$\mu_1 = \frac{4\mu_0 + 5732 \times v_0^2 \times 39.85}{4 + 5732 \times v_0^2},$$

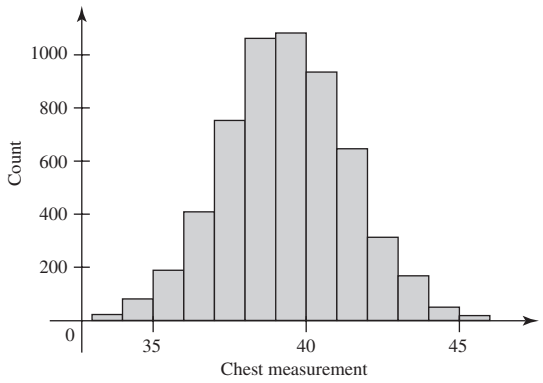

Figure 7.6 Histogram of chest measurements of Scottish militiamen in Example 7.4.7.

and variance

$$v_1^2 = \frac{4v_0^2}{4 + 5732v_0^2}.$$

The Bayes estimate will then be $\delta(\boldsymbol{x}) = \mu_1$. Notice that, unless μ_0 is incredibly large or v_0^2 is very small, we will have μ_1 nearly equal to 39.85 and v_1^2 nearly equal to $4/5732$. Indeed, if the prior p.d.f. of θ is any continuous function that is positive around $\theta = 39.85$ and is not extremely large when θ is far from 39.85, then the posterior p.d.f. of θ will very nearly be the normal p.d.f. with mean 39.85 and variance $4/5732$. The mean and median of the posterior distribution are nearly $\overline{x}_n$ regardless of the prior distribution. ◀

Consistency of the Bayes Estimator Let $X_1, \ldots, X_n$ be a random sample (given θ) from the Bernoulli distribution with parameter θ. Suppose that we use a conjugate prior for θ. Since θ is the mean of the distribution from which the sample is being taken, it follows from the law of large numbers discussed in Sec. 6.2 that $\overline{X}_n$ converges in probability to θ as $n \to \infty$. Since the difference between the Bayes estimator $\delta^*(\boldsymbol{X})$ and $\overline{X}_n$ converges in probability to 0 as $n \to \infty$, it can also be concluded that $\delta^*(\boldsymbol{X})$ converges in probability to the unknown value of θ as $n \to \infty$.

Definition 7.4.6 Consistent Estimator. A sequence of estimators that converges in probability to the unknown value of the parameter being estimated, as $n \to \infty$, is called a *consistent sequence of estimators*.

Thus, we have shown that the Bayes estimators $\delta^*(\boldsymbol{X})$ form a consistent sequence of estimators in the problem considered here. The practical interpretation of this result is as follows: When large numbers of observations are taken, there is high probability that the Bayes estimator will be very close to the unknown value of θ.

The results that have just been presented for estimating the parameter of a Bernoulli distribution are also true for other estimation problems. Under fairly general conditions and for a wide class of loss functions, the Bayes estimators of some parameters θ will form a consistent sequence of estimators as the sample size $n \to \infty$. In particular, for random samples from any one of the various families of distributions discussed in Sec. 7.3, if a conjugate prior distribution is assigned to the parameters and the squared error loss function is used, the Bayes estimators will form a consistent sequence of estimators.

For example, consider again the conditions of Example 7.4.4. In that example, a random sample is taken from a normal distribution for which the value of the mean

θ is unknown, and the Bayes estimator $\delta^*(\boldsymbol{X})$ is specified by Eq. (7.4.6). By the law of large numbers, $\overline{X}_n$ will converge to the unknown value of the mean θ as $n \to \infty$. It can now be seen from Eq. (7.4.6) that $\delta^*(\boldsymbol{X})$ will also converge to θ as $n \to \infty$. Thus, the Bayes estimators again form a consistent sequence of estimators. Other examples are given in Exercises 7 and 11 at the end of this section.

More General Parameters and Estimators

So far in this section, we have considered only real-valued parameters and estimators of those parameters. There are two very common generalizations of this situation that are easy to handle with the same techniques described above. The first generalization is to multidimensional parameters such as the two-dimensional parameter of a normal distribution with unknown mean and variance. The second generalization is to functions of the parameter rather than the parameter itself. For example, if θ is the failure rate in Example 7.1.1, we might be interested in estimating $1/\theta$, the mean time to failure. As another example, if our data arise from a normal distribution with unknown mean and variance, we might wish to estimate the mean only rather than the entire parameter.

The necessary changes to Definition 7.4.1 in order to handle both of the generalizations just mentioned are given in Definition 7.4.7.

Definition 7.4.7 Estimator/Estimate. Let $X_1, \ldots, X_n$ be observable data whose joint distribution is indexed by a parameter θ taking values in a subset Ω of k-dimensional space. Let h be a function from Ω into d-dimensional space. Define $\psi = h(\theta)$. An *estimator* of ψ is a function $\delta(X_1, \ldots, X_n)$ that takes values in d-dimensional space. If $X_1 = x_1, \ldots, X_n = x_n$ are observed, then $\delta(x_1, \ldots, x_n)$ is called the *estimate* of ψ.

When h in Definition 7.4.7 is the identity function $h(\theta) = \theta$, then $\psi = \theta$ and we are estimating the original parameter θ. When $h(\theta)$ is one coordinate of θ, then the ψ that we are estimating is just that one coordinate.

There will be a number of examples of multidimensional parameters in later sections and chapters of this book. Here is an example of estimating a function of a parameter.

Example 7.4.8 Lifetimes of Electronic Components. In Example 7.3.12, suppose that we want to estimate $\psi = 1/\theta$, the mean time to failure of the electronic components. The posterior distribution of θ is the gamma distribution with parameters 4 and 8.6. If we use the squared error loss $L(\theta, a) = (\psi - a)^2$, Theorem 4.7.3 says that the Bayes estimate is the mean of the posterior distribution of ψ. That is,

$$\begin{aligned}
\delta^*(\boldsymbol{x}) &= E(\psi|\boldsymbol{x}) = E\left(\frac{1}{\theta}\middle|\boldsymbol{x}\right) \\
&= \int_0^\infty \frac{1}{\theta}\xi(\theta|\boldsymbol{x})d\theta \\
&= \int_0^\infty \frac{1}{\theta}\frac{8.6^4}{6}\theta^3 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\int_0^\infty \theta^2 e^{-8.6\theta}d\theta \\
&= \frac{8.6^4}{6}\frac{2}{8.6^3} = 2.867,
\end{aligned}$$

where the final equality follows from Theorem 5.7.3. The mean of $1/\theta$ is slightly higher than $1/E(\theta|\boldsymbol{x}) = 8.6/4 = 2.15$. ◀

Note: Loss Functions and Utility. In Sec. 4.8, we introduced the concept of utility to measure the values to a decision maker of various random outcomes. The concept of loss function is closely related to that of utility. In a sense, a loss function is like the negative of a utility. Indeed, Example 4.8.8 shows how to convert absolute error loss into a utility. In that example, Y plays the role of the parameter and $d(W)$ plays the role of the estimator. In a similar manner, one can convert other loss functions into utilities. Hence, it is not surprising that the goal of maximizing expected utility in Sec. 4.8 has been replaced by the goal of minimizing expected loss in the present section.

Limitations of Bayes Estimators

The theory of Bayes estimators, as described in this section, provides a satisfactory and coherent theory for the estimation of parameters. Indeed, according to statisticians who adhere to the Bayesian philosophy, it provides the only coherent theory of estimation that can possibly be developed. Nevertheless, there are certain limitations to the applicability of this theory in practical statistical problems. To apply the theory, it is necessary to specify a particular loss function, such as the squared error or absolute error function, and also a prior distribution for the parameter. Meaningful specifications may exist, in principle, but it may be very difficult and time-consuming to determine them. In some problems, the statistician must determine the specifications that would be appropriate for clients or employers who are unavailable or otherwise unable to communicate their preferences and knowledge. In other problems, it may be necessary for an estimate to be made jointly by members of a group or committee, and it may be difficult for the members of the group to reach agreement about an appropriate loss function and prior distribution.

Another possible difficulty is that in a particular problem the parameter θ may actually be a vector of real-valued parameters for which all the values are unknown. The theory of Bayes estimation, which has been developed in the preceding sections, can easily be generalized to include the estimation of a vector parameter θ. However, to apply this theory in such a problem it is necessary to specify a multivariate prior distribution for the vector θ and also to specify a loss function $L(\theta, \boldsymbol{a})$ that is a function of the vector θ and the vector $\boldsymbol{a}$, which will be used to estimate θ. Even though the statistician may be interested in estimating only one or two components of the vector θ in a given problem, he must still assign a multivariate prior distribution to the entire vector θ. In many important statistical problems, some of which will be discussed later in this book, θ may have a large number of components. In such a problem, it is especially difficult to specify a meaningful prior distribution on the multidimensional parameter space Ω.

It should be emphasized that there is no simple way to resolve these difficulties. Other methods of estimation that are not based on prior distributions and loss functions typically have practical limitations, also. These other methods also typically have serious defects in their theoretical structure as well.

Summary

An estimator of a parameter θ is a function δ of the data $\boldsymbol{X}$. If $\boldsymbol{X} = \boldsymbol{x}$ is observed, the value $\delta(\boldsymbol{x})$ is called our estimate, the observed value of the estimator $\delta(\boldsymbol{X})$. A loss

function $L(\theta, a)$ is designed to measure how costly it is to use the value a to estimate θ. A Bayes estimator $\delta^*(\boldsymbol{X})$ is chosen so that $a = \delta^*(\boldsymbol{x})$ provides the minimum value of the posterior mean of $L(\theta, a)$. That is,

$$E[L(\theta, \delta^*(\boldsymbol{x}))|\boldsymbol{x}] = \min_a E[L(\theta, a)|\boldsymbol{x}].$$

If the loss is squared error, $L(\theta, a) = (\theta - a)^2$, then $\delta^*(\boldsymbol{x})$ is the posterior mean of θ, $E(\theta|\boldsymbol{x})$. If the loss is absolute error, $L(\theta, a) = |\theta - a|$, then $\delta^*(\boldsymbol{x})$ is a median of the posterior distribution of θ. For other loss functions, locating the minimum might have to be done numerically.

Exercises

1. In a clinical trial, let the probability of successful outcome θ have a prior distribution that is the uniform distribution on the interval [0, 1], which is also the beta distribution with parameters 1 and 1. Suppose that the first patient has a successful outcome. Find the Bayes estimates of θ that would be obtained for both the squared error and absolute error loss functions.

2. Suppose that the proportion θ of defective items in a large shipment is unknown, and the prior distribution of θ is the beta distribution for which the parameters are $\alpha = 5$ and $\beta = 10$. Suppose also that 20 items are selected at random from the shipment, and that exactly one of these items is found to be defective. If the squared error loss function is used, what is the Bayes estimate of θ?

3. Consider again the conditions of Exercise 2. Suppose that the prior distribution of θ is as given in Exercise 2, and suppose again that 20 items are selected at random from the shipment.

a. For what number of defective items in the sample will the mean squared error of the Bayes estimate be a maximum?

b. For what number will the mean squared error of the Bayes estimate be a minimum?

4. Suppose that a random sample of size n is taken from the Bernoulli distribution with parameter θ, which is unknown, and that the prior distribution of θ is a beta distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\gamma_n \to 1$ as $n \to \infty$.

5. Suppose that the number of defects in a 1200-foot roll of magnetic recording tape has a Poisson distribution for which the value of the mean θ is unknown, and the prior distribution of θ is the gamma distribution with parameters $\alpha = 3$ and $\beta = 1$. When five rolls of this tape are selected at random and inspected, the numbers of defects found on the rolls are 2, 2, 6, 0, and 3. If the squared error loss function is used, what is the Bayes estimate of θ? (See Exercise 5 of Sec. 7.3.)

6. Suppose that a random sample of size n is taken from a Poisson distribution for which the value of the mean θ is unknown, and the prior distribution of θ is a gamma distribution for which the mean is μ_0. Show that the mean of the posterior distribution of θ will be a weighted average having the form $\gamma_n \overline{X}_n + (1 - \gamma_n)\mu_0$, and show that $\gamma_n \to 1$ as $n \to \infty$.

7. Consider again the conditions of Exercise 6, and suppose that the value of θ must be estimated by using the squared error loss function. Show that the Bayes estimators, for $n = 1, 2, \ldots$, form a consistent sequence of estimators of θ.

8. Suppose that the heights of the individuals in a certain population have a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2 inches. Suppose also that the prior distribution of θ is a normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. Suppose finally that 10 people are selected at random from the population, and their average height is found to be 69.5 inches.

a. If the squared error loss function is used, what is the Bayes estimate of θ?

b. If the absolute error loss function is used, what is the Bayes estimate of θ? (See Exercise 7 of Sec. 7.3).

9. Suppose that a random sample is to be taken from a normal distribution for which the value of the mean θ is unknown and the standard deviation is 2, the prior distribution of θ is a normal distribution for which the standard deviation is 1, and the value of θ must be estimated by using the squared error loss function. What is the smallest random sample that must be taken in order for the mean squared error of the Bayes estimator of θ to be 0.01 or less? (See Exercise 10 of Sec. 7.3.)

10. Suppose that the time in minutes required to serve a customer at a certain facility has an exponential distribution for which the value of the parameter θ is unknown,

the prior distribution of θ is a gamma distribution for which the mean is 0.2 and the standard deviation is 1, and the average time required to serve a random sample of 20 customers is observed to be 3.8 minutes. If the squared error loss function is used, what is the Bayes estimate of θ? (See Exercise 12 of Sec. 7.3.)

11. Suppose that a random sample of size n is taken from an exponential distribution for which the value of the parameter θ is unknown, the prior distribution of θ is a specified gamma distribution, and the value of θ must be estimated by using the squared error loss function. Show that the Bayes estimators, for $n = 1, 2, \ldots$, form a consistent sequence of estimators of θ.

12. Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior p.d.f.'s $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\xi_A(\theta) = 2\theta \quad \text{for } 0 < \theta < 1,$$
$$\xi_B(\theta) = 4\theta^3 \quad \text{for } 0 < \theta < 1.$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

- **a.** Find the posterior distribution that each statistician assigns to θ.
- **b.** Find the Bayes estimate for each statistician based on the squared error loss function.
- **c.** Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown. Suppose also that the prior distribution of θ is the Pareto distribution with parameters x_0 and α ($x_0 > 0$ and $\alpha > 0$), as defined in Exercise 16 of Sec. 5.7. If the value of θ is to be estimated by using the squared error loss function, what is the Bayes estimator of θ? (See Exercise 18 of Sec. 7.3.)

14. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter θ is unknown ($\theta > 0$). Let $\xi(\theta)$ denote the prior p.d.f. of θ, and let $\hat{\theta}$ denote the Bayes estimator of θ with respect to the prior p.d.f. $\xi(\theta)$ when the squared error loss function is used. Let $\psi = \theta^2$, and suppose that instead of estimating θ, it is desired to estimate the value of ψ subject to the following squared error loss function:

$$L(\psi, a) = (\psi - a)^2 \quad \text{for } \psi > 0 \text{ and } a > 0.$$

Let $\hat{\psi}$ denote the Bayes estimator of ψ. Explain why $\hat{\psi} > \hat{\theta}^2$. *Hint:* Look at Exercise 4 in Sec. 4.4.

15. Let $c > 0$ and consider the loss function

$$L(\theta, a) = \begin{cases} c|\theta - a| & \text{if } \theta < a, \\ |\theta - a| & \text{if } \theta \geq a. \end{cases}$$

Assume that θ has a continuous distribution. Prove that a Bayes estimator of θ will be any $1/(1 + c)$ quantile of the posterior distribution of θ. *Hint:* The proof is a lot like the proof of Theorem 4.5.3. The result holds even if θ does not have a continuous distribution, but the proof is more cumbersome.

7.5 Maximum Likelihood Estimators

Maximum likelihood estimation is a method for choosing estimators of parameters that avoids using prior distributions and loss functions. It chooses as the estimate of θ the value of θ that provides the largest value of the likelihood function.

Introduction

Example 7.5.1

Lifetimes of Electronic Components. Suppose that we observe the data in Example 7.3.11 consisting of the lifetimes of three electronic components. Is there a method for estimating the failure rate θ without first constructing a prior distribution and a loss function? ◀

In this section, we shall develop a relatively simple method of constructing an estimator without having to specify a loss function and a prior distribution. It is called the method of *maximum likelihood*, and it was introduced by R. A. Fisher in 1912. Maximum likelihood estimation can be applied in most problems, it has a strong

intuitive appeal, and it will often yield a reasonable estimator of θ. Furthermore, if the sample is large, the method will typically yield an excellent estimator of θ. For these reasons, the method of maximum likelihood is probably the most widely used method of estimation in statistics.

Note: Terminology. Because maximum likelihood estimation, as well as many other procedures to be introduced later in the text, do not involve the specification of a prior distribution of the parameter, some different terminology is often used in describing the statistical models to which these procedures are applied. Rather than saying that $X_1, \ldots, X_n$ are i.i.d. with p.f. or p.d.f. $f(x|\theta)$ conditional on θ, we might say that $X_1, \ldots, X_n$ form a random sample from a distribution with p.f. or p.d.f. $f(x|\theta)$ where θ is unknown. More specifically, in Example 7.5.1, we could say that the lifetimes form a random sample from the exponential distribution with unknown parameter θ.

Definition of a Maximum Likelihood Estimator

Let the random variables $X_1, \ldots, X_n$ form a random sample from a discrete distribution or a continuous distribution for which the p.f. or the p.d.f. is $f(x|\theta)$, where the parameter θ belongs to some parameter space Ω. Here, θ can be either a real-valued parameter or a vector. For every observed vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ in the sample, the value of the joint p.f. or joint p.d.f. will, as usual, be denoted by $f_n(\boldsymbol{x}|\theta)$. Because of its importance in this section, we repeat Definition 7.2.3.

Definition 7.5.1 Likelihood Function. When the joint p.d.f. or the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of the observations in a random sample is regarded as a function of θ for given values of $x_1, \ldots, x_n$, it is called the *likelihood function*.

Consider first, the case in which the observed vector $\boldsymbol{x}$ came from a discrete distribution. If an estimate of θ must be selected, we would certainly not consider any value of $\theta \in \Omega$ for which it would be impossible to obtain the vector $\boldsymbol{x}$ that was actually observed. Furthermore, suppose that the probability $f_n(\boldsymbol{x}|\theta)$ of obtaining the actual observed vector $\boldsymbol{x}$ is very high when θ has a particular value, say, $\theta = \theta_0$, and is very small for every other value of $\theta \in \Omega$. Then we would naturally estimate the value of θ to be θ_0 (unless we had strong prior information that outweighed the evidence in the sample and pointed toward some other value). When the sample comes from a continuous distribution, it would again be natural to try to find a value of θ for which the probability density $f_n(\boldsymbol{x}|\theta)$ is large and to use this value as an estimate of θ. For each possible observed vector $\boldsymbol{x}$, we are led by this reasoning to consider a value of θ for which the likelihood function $f_n(\boldsymbol{x}|\theta)$ is a maximum and to use this value as an estimate of θ. This concept is formalized in the following definition.

Definition 7.5.2 Maximum Likelihood Estimator/Estimate. For each possible observed vector $\boldsymbol{x}$, let $\delta(\boldsymbol{x}) \in \Omega$ denote a value of $\theta \in \Omega$ for which the likelihood function $f_n(\boldsymbol{x}|\theta)$ is a maximum, and let $\hat{\theta} = \delta(\boldsymbol{X})$ be the estimator of θ defined in this way. The estimator $\hat{\theta}$ is called a *maximum likelihood estimator* of θ. After $\boldsymbol{X} = \boldsymbol{x}$ is observed, the value $\delta(\boldsymbol{x})$ is called a *maximum likelihood estimate* of θ.

The expressions *maximum likelihood estimator* and *maximum likelihood estimate* are abbreviated M.L.E. One must rely on context to determine whether the abbreviation refers to an estimator or to an estimate. Note that the M.L.E. is required to be an element of the parameter space Ω, unlike general estimators/estimates for which no such requirement exists.

Examples of Maximum Likelihood Estimators

Example 7.5.2

Lifetimes of Electronic Components. In Example 7.3.11, the observed data are $X_1 = 3$, $X_2 = 1.5$, and $X_3 = 2.1$. The random variables had been modeled as a random sample of size 3 from the exponential distribution with parameter θ. The likelihood function is, for $\theta > 0$,

$$f_3(\boldsymbol{x}|\theta) = \theta^3 \exp(-6.6\theta),$$

where $\boldsymbol{x} = (2, 1.5, 2.1)$. The value of θ that maximizes the likelihood function $f_3(\boldsymbol{x}|\theta)$ will be the same as the value of θ that maximizes $\log f_3(\boldsymbol{x}|\theta)$, since log is an increasing function. Therefore, it will be convenient to determine the M.L.E. by finding the value of θ that maximizes

$$L(\theta) = \log f_3(\boldsymbol{x}|\theta) = 3\log(\theta) - 6.6\theta.$$

Taking the derivative $dL(\theta)/d\theta$, setting the derivative to 0, and solving for θ yields $\theta = 3/6.6 = 0.455$. The second derivative is negative at this value of θ, so it provides a maximum. The maximum likelihood estimate is then 0.455. ◀

It should be noted that in some problems, for certain observed vectors $\boldsymbol{x}$, the maximum value of $f_n(\boldsymbol{x}|\theta)$ may not actually be attained for any point $\theta \in \Omega$. In such a case, an M.L.E. of θ does not exist. For certain other observed vectors $\boldsymbol{x}$, the maximum value of $f_n(\boldsymbol{x}|\theta)$ may actually be attained at more than one point in the space Ω. In such a case, the M.L.E. is not uniquely defined, and any one of these points can be chosen as the value of the estimator $\hat{\theta}$. In many practical problems, however, the M.L.E. exists and is uniquely defined.

We shall now illustrate the method of maximum likelihood and these various possibilities by considering several examples. In each example, we shall attempt to determine an M.L.E.

Example 7.5.3

Test for a Disease. Suppose that you are walking down the street and notice that the Department of Public Health is giving a free medical test for a certain disease. The test is 90 percent reliable in the following sense: If a person has the disease, there is a probability of 0.9 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of only 0.1 that the test will give a positive response. This same test was considered in Example 2.3.1. We shall let X stand for the result of the test, where $X = 1$ means that the test is positive and $X = 0$ means that the test is negative. Let the parameter space be $\Omega = \{0.1, 0.9\}$, where $\theta = 0.1$ means that the person tested does not have the disease, and $\theta = 0.9$ means that the person has the disease. This parameter space was chosen so that, given θ, X has the Bernoulli distribution with parameter θ. The likelihood function is

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}.$$

If $x = 0$ is observed, then

$$f(0|\theta) = \begin{cases} 0.9 & \text{if } \theta = 0.1, \\ 0.1 & \text{if } \theta = 0.9. \end{cases}$$

Clearly, $\theta = 0.1$ maximizes the likelihood when $x = 0$ is observed. If $x = 1$ is observed, then

$$f(1|\theta) = \begin{cases} 0.1 & \text{if } \theta = 0.1, \\ 0.9 & \text{if } \theta = 0.9. \end{cases}$$

Clearly, $\theta = 0.9$ maximizes the likelihood when $x = 1$ is observed. Hence, we have that the M.L.E. is

$$\hat{\theta} = \begin{cases} 0.1 & \text{if } X = 0, \\ 0.9 & \text{if } X = 1. \end{cases}$$

◀

Example 7.5.4

Sampling from a Bernoulli Distribution. Suppose that the random variables $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown ($0 \le \theta \le 1$). For all observed values $x_1, \ldots, x_n$, where each x_i is either 0 or 1, the likelihood function is

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}. \tag{7.5.1}$$

Instead of maximizing the likelihood function $f_n(\boldsymbol{x}|\theta)$ directly, it is again easier to maximize $\log f_n(\boldsymbol{x}|\theta)$:

$$L(\theta) = \log f_n(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} [x_i \log\theta + (1-x_i)\log(1-\theta)]$$
$$= \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta).$$

Now calculate the derivative $dL(\theta)/d\theta$, set this derivative equal to 0, and solve the resulting equation for θ. If $\sum_{i=1}^{n} x_i \notin \{0, n\}$, we find that the derivative is 0 at $\theta = \overline{x}_n$, and it can be verified (for example, by examining the second derivative) that this value does indeed maximize $L(\theta)$ and the likelihood function defined by Eq. (7.5.1). If $\sum_{i=1}^{n} x_i = 0$, then $L(\theta)$ is a decreasing function of θ for all θ, and hence L achieves its maximum at $\theta = 0$. Similarly, if $\sum_{i=1}^{n} x_i = n$, L is an increasing function, and it achieves its maximum at $\theta = 1$. In these last two cases, note that the maximum of the likelihood occurs at $\theta = \overline{x}_n$. It follows, therefore, that the M.L.E. of θ is $\hat{\theta} = \overline{X}_n$.

◀

It follows from Example 7.5.4 that if $X_1, \ldots, X_n$ are regarded as n Bernoulli trials and if the parameter space is $\Omega = [0, 1]$, then the M.L.E. of the unknown probability of success on any given trial is simply the proportion of successes observed in the n trials. In Example 7.5.3, we have $n = 1$ Bernoulli trial, but the parameter space is $\Omega = \{0.1, 0.9\}$ rather than $[0, 1]$, and the M.L.E. differs from the proportion of successes.

Example 7.5.5

Sampling from a Normal Distribution with Unknown Mean. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known. For all observed values $x_1, \ldots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\mu)$ will be

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \tag{7.5.2}$$

It can be seen from Eq. (7.5.2) that $f_n(\boldsymbol{x}|\mu)$ will be maximized by the value of μ that minimizes

$$Q(\mu) = \sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2.$$

We see that Q is a quadratic in μ with positive coefficient on μ^2. It follows that Q will be minimized where its derivative is 0. If we now calculate the derivative $dQ(\mu)/d\mu$, set this derivative equal to 0, and solve the resulting equation for μ, we find that $\mu = \overline{x}_n$. It follows, therefore, that the M.L.E. of μ is $\hat{\mu} = \overline{X}_n$. ◀

It can be seen in Example 7.5.5 that the estimator $\hat{\mu}$ is not affected by the value of the variance σ^2, which we assumed was known. The M.L.E. of the unknown mean μ is simply the sample mean $\overline{X}_n$, regardless of the value of σ^2. We shall see this again in the next example, in which both μ and σ^2 must be estimated.

Example 7.5.6 Sampling from a Normal Distribution with Unknown Mean and Variance. Suppose again that $X_1, \ldots, X_n$ form a random sample from a normal distribution, but suppose now that both the mean μ and the variance σ^2 are unknown. The parameter is then $\theta = (\mu, \sigma^2)$. For all observed values $x_1, \ldots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ will again be given by the right side of Eq. (7.5.2). This function must now be maximized over all possible values of μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Instead of maximizing the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ directly, it is again easier to maximize $\log f_n(\boldsymbol{x}|\mu, \sigma^2)$. We have

$$\begin{aligned} L(\theta) &= \log f_n(\boldsymbol{x}|\mu, \sigma^2) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \end{aligned} \tag{7.5.3}$$

We shall find the value of $\theta = (\mu, \sigma^2)$ for which $L(\theta)$ is maximum in three stages. First, for each fixed σ^2, we shall find the value $\hat{\mu}(\sigma^2)$ that maximizes the right side of (7.5.3). Second, we shall find the value $\widehat{\sigma^2}$ of σ^2 that maximizes $L(\theta')$ when $\theta' = (\hat{\mu}(\sigma^2), \sigma^2)$. Finally, the M.L.E. of θ will be the random vector whose observed value is $(\hat{\mu}(\widehat{\sigma^2}), \widehat{\sigma^2})$. The first stage has already been solved in Example 7.5.5. There, we obtained $\hat{\mu}(\sigma^2) = \overline{x}_n$. For the second stage, we set $\theta' = (\overline{x}_n, \sigma^2)$ and maximize

$$L(\theta') = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2. \tag{7.5.4}$$

This can be maximized by setting its derivative with respect to σ^2 equal to 0 and solving for σ^2. The derivative is

$$\frac{d}{d\sigma^2}L(\theta') = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Setting this to 0 yields

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2. \tag{7.5.5}$$

The second derivative of (7.5.4) is negative at the value of σ^2 in (7.5.5), so we have found the maximum. Therefore, the M.L.E. of $\theta = (\mu, \sigma^2)$ is

$$\hat{\theta} = (\hat{\mu}, \widehat{\sigma^2}) = \left(\overline{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right). \tag{7.5.6}$$

Notice that the first coordinate of the M.L.E. in Eq. (7.5.6) is called the sample mean of the data. Likewise, we call the second coordinate of this M.L.E. the *sample variance*. It is not difficult to see that the observed value of the sample variance is

the variance of a distribution that assigns probability $1/n$ to each of the n observed values $x_1, \ldots, x_n$ in the sample. (See Exercise 1.) ◀

Example 7.5.7 Sampling from a Uniform Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown ($\theta > 0$). The p.d.f. $f(x|\theta)$ of each observation has the following form:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.7}$$

Therefore, the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \ldots, X_n$ has the form

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le x_i \le \theta \ (i = 1, \ldots, n), \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.8}$$

It can be seen from Eq. (7.5.8) that the M.L.E. of θ must be a value of θ for which $\theta \ge x_i$ for $i = 1, \ldots, n$ and that maximizes $1/\theta^n$ among all such values. Since $1/\theta^n$ is a decreasing function of θ, the estimate will be the smallest value of θ such that $\theta \ge x_i$ for $i = 1, \ldots, n$. Since this value is $\theta = \max\{x_1, \ldots, x_n\}$, the M.L.E. of θ is $\hat{\theta} = \max\{X_1, \ldots, X_n\}$. ◀

Limitations of Maximum Likelihood Estimation

Despite its intuitive appeal, the method of maximum likelihood is not necessarily appropriate in all problems. For instance, in Example 7.5.7, the M.L.E. $\hat{\theta}$ does not seem to be a suitable estimator of θ. Since $\max\{X_1, \ldots, X_n\} < \theta$ with probability 1, it follows that $\hat{\theta}$ surely underestimates the value of θ. Indeed, if any prior distribution is assigned to θ, then the Bayes estimator of θ will surely be greater than $\hat{\theta}$. The actual amount by which the Bayes estimator exceeds $\hat{\theta}$ will, of course, depend on the particular prior distribution that is used and on the observed values of $X_1, \ldots, X_n$. Example 7.5.7 also raises another difficulty with maximum likelihood, as we illustrate in Example 7.5.8.

Example 7.5.8 Nonexistence of an M.L.E. Suppose again that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$. However, suppose now that instead of writing the p.d.f. $f(x|\theta)$ of the uniform distribution in the form given in Eq. (7.5.7), we write it in the following form:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.9}$$

The only difference between Eq. (7.5.7) and Eq. (7.5.9) is that the value of the p.d.f. at each of the two endpoints 0 and θ has been changed by replacing the weak inequalities in Eq. (7.5.7) with strict inequalities in Eq. (7.5.9). Therefore, either equation could be used as the p.d.f. of the uniform distribution. However, if Eq. (7.5.9) is used as the p.d.f, then an M.L.E. of θ will be a value of θ for which $\theta > x_i$ for $i = 1, \ldots, n$ and which maximizes $1/\theta^n$ among all such values. It should be noted that the possible values of θ no longer include the value $\theta = \max\{x_1, \ldots, x_n\}$, because θ must be *strictly* greater than each observed value x_i ($i = 1, \ldots, n$). Because θ can be chosen arbitrarily close to the value $\max\{x_1, \ldots, x_n\}$ but cannot be chosen equal to this value, it follows that the M.L.E. of θ does not exist. ◀

In all of our previous discussions about p.d.f.'s, we emphasized the fact that it is irrelevant whether the p.d.f. of the uniform distribution is chosen to be equal to $1/\theta$

over the open interval $0 < x < \theta$ or over the closed interval $0 \le x \le \theta$. Now, however, we see that the existence of an M.L.E. depends on this irrelevant and unimportant choice. This difficulty is easily avoided in Example 7.5.8 by using the p.d.f. given by Eq. (7.5.7) rather than that given by Eq. (7.5.9). In many other problems as well, a difficulty of this type can be avoided simply by choosing one particular appropriate version of the p.d.f. to represent the given distribution. However, as we shall see in Example 7.5.10, the difficulty cannot always be avoided.

Example 7.5.9 Non-uniqueness of an M.L.E. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta + 1]$, where the value of the parameter θ is unknown ($-\infty < \theta < \infty$). In this example, the joint p.d.f. $f_n(\boldsymbol{x}|\theta)$ has the form

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} 1 & \text{for } \theta \le x_i \le \theta + 1, (i = 1, \ldots, n), \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.10}$$

The condition that $\theta \le x_i$ for $i = 1, \ldots, n$ is equivalent to the condition that $\theta \le \min\{x_1, \ldots, x_n\}$. Similarly, the condition that $x_i \le \theta + 1$ for $i = 1, \ldots, n$ is equivalent to the condition that $\theta \ge \max\{x_1, \ldots, x_n\} - 1$. Therefore, instead of writing $f_n(\boldsymbol{x}|\theta)$ in the form given in Eq. (7.5.10), we can use the following form:

$$f_n(\boldsymbol{x}|\theta) = \begin{cases} 1 & \text{for } \max\{x_1, \ldots, x_n\} - 1 \le \theta \le \min\{x_1, \ldots, x_n\}, \\ 0 & \text{otherwise.} \end{cases} \tag{7.5.11}$$

Thus, it is possible to select as an M.L.E. any value of θ in the interval

$$\max\{x_1, \ldots, x_n\} - 1 \le \theta \le \min\{x_1, \ldots, x_n\}. \tag{7.5.12}$$

In this example, the M.L.E. is not uniquely specified. In fact, the method of maximum likelihood provides very little help in choosing an estimate of θ. The likelihood of every value of θ outside the interval (7.5.12) is actually 0. Therefore, no value θ outside this interval would ever be estimated, and all values inside the interval are M.L.E.'s. ◀

Example 7.5.10 Sampling from a Mixture of Two Distributions. Consider a random variable X that can come with equal probability either from the normal distribution with mean 0 and variance 1 or from another normal distribution with mean μ and variance σ^2, where both μ and σ^2 are unknown. Under these conditions, the p.d.f. $f(x|\mu, \sigma^2)$ of X will be the average of the p.d.f.'s of the two different normal distributions. Thus,

$$f(x|\mu, \sigma^2) = \frac{1}{2}\left\{\frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x^2}{2}\right) + \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right\}. \tag{7.5.13}$$

Suppose now that $X_1, \ldots, X_n$ form a random sample from the distribution for which the p.d.f. is given by Eq. (7.5.13). As usual, the likelihood function $f_n(\boldsymbol{x}|\mu, \sigma^2)$ has the form

$$f_n(\boldsymbol{x}|\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2). \tag{7.5.14}$$

To find the M.L.E. of $\theta = (\mu, \sigma^2)$, we must find values of μ and σ^2 for which $f_n(\boldsymbol{x}|\mu, \sigma^2)$ is maximized.

Let x_k denote any one of the observed values $x_1, \ldots, x_n$. If we let $\mu = x_k$ and let $\sigma^2 \to 0$, then the factor $f(x_k|\mu, \sigma^2)$ on the right side of Eq. (7.5.14) will grow large without bound, while each factor $f(x_i|\mu, \sigma^2)$ for $x_i \ne x_k$ will approach the value

$$\frac{1}{2(2\pi)^{1/2}} \exp\left(-\frac{x_i^2}{2}\right).$$

Hence, when $\mu = x_k$ and $\sigma^2 \to 0$, we find that $f_n(\boldsymbol{x}|\mu, \sigma^2) \to \infty$.

The value 0 is not a permissible estimate of σ^2, because we know in advance that $\sigma^2 > 0$. Since the likelihood function can be made arbitrarily large by choosing $\mu = x_k$ and choosing σ^2 arbitrarily close to 0, it follows that the M.L.E. does not exist.

If we try to correct this difficulty by allowing the value 0 to be a permissible estimate of σ^2, then we find that there are n different M.L.E.'s of μ and σ^2; namely,

$$\hat{\theta}_k = (\hat{\mu}, \widehat{\sigma^2}) = (X_k, 0) \text{ for } k = 1, \ldots, n.$$

None of these estimators seems appropriate. Consider again the description, given at the beginning of this example, of the two normal distributions from which each observation might come. Suppose, for example, that $n = 1000$, and we use the estimator $\hat{\theta}_3 = (X_3, 0)$. Then, we would be estimating the value of the unknown variance to be 0; also, in effect, we would be behaving as if exactly one of the X_i's (namely, X_3) comes from the given unknown normal distribution, whereas all the other 999 observation values come from the normal distribution with mean 0 and variance 1. In fact, however, since each observation was equally likely to come from either of the two distributions, it is much more probable that hundreds of observations, rather than just one, come from the unknown normal distribution. In this example, the method of maximum likelihood is obviously unsatisfactory. A Bayesian solution to this problem is outlined in Exercise 10 in Sec. 12.5. ◀

Finally, we shall mention one point concerning the interpretation of the M.L.E. The M.L.E. is the value of θ that maximizes the conditional p.f. or p.d.f. of the data $\boldsymbol{X}$ given θ. Therefore, the maximum likelihood estimate is the value of θ that assigned the highest probability to seeing the observed data. It is not necessarily the value of the parameter that appears to be most likely given the data. To say how likely are different values of the parameter, one would need a probability distribution for the parameter. Of course, the posterior distribution of the parameter (Sec. 7.2) would serve this purpose, but no posterior distribution is involved in the calculation of the M.L.E. Hence, it is not legitimate to interpret the M.L.E. as the most likely value of the parameter after having seen the data.

For example, consider a situation covered by Example 7.5.4. Suppose that we are going to flip a coin a few times, and we are concerned with whether or not it has a slight bias toward heads or toward tails. Let $X_i = 1$ if the ith flip is heads and $X_i = 0$ if not. If we obtain four heads and one tail in the first five flips, the observed value of the M.L.E. will be 0.8. But it would be difficult to imagine a situation in which we would feel that the most likely value of θ, the probability of heads, is as large as 0.8 based on just five tosses of what appeared a priori to be a typical coin. Treating the M.L.E. as if it were the most likely value of the parameter is very much the same as ignoring the prior information about the rare disease in the medical test of Examples 2.3.1 and 2.3.3. If the test is positive in these examples, we found (in Example 7.5.3) that the M.L.E. takes the value $\hat{\theta} = 0.9$, which corresponds to having the disease. However, if the prior probability that you have the disease is as small as in Example 2.3.1, the posterior probability that you have the disease ($\theta = 0.9$) is still small even after the positive test result. The test is not accurate enough to completely overcome the prior information. So too with our coin tossing; five tosses are not enough information to overcome prior beliefs about the coin being typical. Only when the data contain much more information than is available a priori would

it be approximately correct to think of the M.L.E. as the value that we believe the parameter is most likely to be near. This could happen either when the M.L.E. is based on a lot of data or when there is very little prior information.

Summary

The maximum likelihood estimate of a parameter θ is that value of θ that provides the largest value of the likelihood function $f_n(\boldsymbol{x}|\theta)$ for fixed data $\boldsymbol{x}$. If $\delta(\boldsymbol{x})$ denotes the maximum likelihood estimate, then $\hat{\theta} = \delta(\boldsymbol{X})$ is the maximum likelihood estimator (M.L.E.). We have computed the M.L.E. when the data comprise a random sample from a Bernoulli distribution, a normal distribution with known variance, a normal distribution with both parameters unknown, or the uniform distribution on the interval $[0, \theta]$ or on the interval $[\theta, \theta + 1]$.

Exercises

1. Let $x_1, \ldots, x_n$ be distinct numbers. Let Y be a discrete random variable with the following p.f.:

$$f(y) = \begin{cases} \frac{1}{n} & \text{if } y \in \{x_1, \ldots, x_n\}, \\ 0 & \text{otherwise.} \end{cases}$$

Prove that $\text{Var}(Y)$ is given by Eq. (7.5.5).

2. It is not known what proportion p of the purchases of a certain brand of breakfast cereal are made by women and what proportion are made by men. In a random sample of 70 purchases of this cereal, it was found that 58 were made by women and 12 were made by men. Find the M.L.E. of p.

3. Consider again the conditions in Exercise 2, but suppose also that it is known that $\frac{1}{2} \leq p \leq \frac{2}{3}$. If the observations in the random sample of 70 purchases are as given in Exercise 2, what is the M.L.E. of p?

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter θ, which is unknown, but it is known that θ lies in the open interval $0 < \theta < 1$. Show that the M.L.E. of θ does not exist if every observed value is 0 or if every observed value is 1.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from a Poisson distribution for which the mean θ is unknown, $(\theta > 0)$.

a. Determine the M.L.E. of θ, assuming that at least one of the observed values is different from 0.

b. Show that the M.L.E. of θ does not exist if every observed value is 0.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ is known, but the variance σ^2 is unknown. Find the M.L.E. of σ^2.

7. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown $(\beta > 0)$. Find the M.L.E. of β.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} e^{\theta - x} & \text{for } x > \theta, \\ 0 & \text{for } x \leq \theta. \end{cases}$$

Also, suppose that the value of θ is unknown $(-\infty < \theta < \infty)$.

a. Show that the M.L.E. of θ does not exist.

b. Determine another version of the p.d.f. of this same distribution for which the M.L.E. of θ will exist, and find this estimator.

9. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \begin{cases} \theta x^{\theta - 1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that the value of θ is unknown $(\theta > 0)$. Find the M.L.E. of θ.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. $f(x|\theta)$ is as follows:

$$f(x|\theta) = \frac{1}{2} e^{-|x - \theta|} \quad \text{for } -\infty < x < \infty.$$

Also, suppose that the value of θ is unknown $(-\infty < \theta < \infty)$. Find the M.L.E. of θ. *Hint:* Compare this to the problem of minimizing M.A.E as in Theorem 4.5.3.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta_1, \theta_2]$, where both θ_1 and θ_2 are unknown ($-\infty < \theta_1 < \theta_2 < \infty$). Find the M.L.E.'s of θ_1 and θ_2.

12. Suppose that a certain large population contains k different types of individuals ($k \geq 2$), and let θ_i denote the proportion of individuals of type i, for $i = 1, \ldots, k$. Here, $0 \leq \theta_i \leq 1$ and $\theta_1 + \cdots + \theta_k = 1$. Suppose also that in a random sample of n individuals from this population, exactly n_i individuals are of type i, where $n_1 + \cdots + n_k = n$. Find the M.L.E.'s of $\theta_1, \ldots, \theta_k$.

13. Suppose that the two-dimensional vectors (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ form a random sample from a bivariate normal distribution for which the means of X and Y are unknown but the variances of X and Y and the correlation between X and Y are known. Find the M.L.E.'s of the means.

7.6 Properties of Maximum Likelihood Estimators

In this section, we explore several properties of M.L.E.'s, including:

- *The relationship between the M.L.E. of a parameter and the M.L.E. of a function of that parameter*
- *The need for computational algorithms*
- *The behavior of the M.L.E. as the sample size increases*
- *The lack of dependence of the M.L.E. on the sampling plan*

We also introduce a popular alternative method of estimation (method of moments) that sometimes agrees with maximum likelihood, but can sometimes be computationally simpler.

Invariance

Example 7.6.1 Lifetimes of Electronic Components. In Example 7.1.1, the parameter θ was interpreted as the failure rate of electronic components. In Example 7.4.8, we found a Bayes estimate of $\psi = 1/\theta$, the average lifetime. Is there a corresponding method for computing the M.L.E. of ψ? ◀

Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which either the p.f. or the p.d.f. is $f(x|\theta)$, where the value of the parameter θ is unknown. The parameter may be one-dimensional or a vector of parameters. Let $\hat{\theta}$ denote the M.L.E. of θ. Thus, for all observed values $x_1, \ldots, x_n$, the likelihood function $f_n(\boldsymbol{x}|\theta)$ is maximized when $\theta = \hat{\theta}$.

Suppose now that we change the parameter in the distribution as follows: Instead of expressing the p.f. or the p.d.f. $f(x|\theta)$ in terms of the parameter θ, we shall express it in terms of a new parameter $\psi = g(\theta)$, where g is a one-to-one function of θ. Is there a relationship between the M.L.E. of θ and the M.L.E. of ψ?

Theorem 7.6.1 Invariance Property of M.L.E.'s. If $\hat{\theta}$ is the maximum likelihood estimator of θ and if g is a one-to-one function, then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$.

Proof The new parameter space is Γ, the image of Ω under the function g. We shall let $\theta = h(\psi)$ denote the inverse function. Then, expressed in terms of the new parameter ψ, the p.f. or p.d.f. of each observed value will be $f[x|h(\psi)]$, and the likelihood function will be $f_n[\boldsymbol{x}|h(\psi)]$.

The M.L.E. $\hat{\psi}$ of ψ will be equal to the value of ψ for which $f_n[\boldsymbol{x}|h(\psi)]$ is maximized. Since $f_n(\boldsymbol{x}|\theta)$ is maximized when $\theta = \hat{\theta}$, it follows that $f_n[\boldsymbol{x}|h(\psi)]$ is

maximized when $h(\psi) = \hat{\theta}$. Hence, the M.L.E. $\hat{\psi}$ must satisfy the relation $h(\hat{\psi}) = \hat{\theta}$ or, equivalently, $\hat{\psi} = g(\hat{\theta})$. ■

Example 7.6.2 Lifetimes of Electronic Components. According to Theorem 7.6.1, the M.L.E. of ψ is one over the M.L.E. of θ. In Example 7.5.2, we computed the observed value of $\hat{\theta} = 0.455$. The observed value of $\hat{\psi}$ would then be $1/0.455 = 2.2$. This is a bit smaller than the Bayes estimate using squared error loss of 2.867 found in Example 7.4.8. ◀

The invariance property can be extended to functions that are not one-to-one. For example, suppose that we wish to estimate the mean μ of a normal distribution when both the mean and the variance are unknown. Then μ is not a one-to-one function of the parameter $\theta = (\mu, \sigma^2)$. In this case, the function we wish to estimate is $g(\theta) = \mu$. There is a way to define the M.L.E. of a function of θ that is not necessarily one-to-one. One popular way is the following.

Definition 7.6.1 M.L.E. of a Function. Let $g(\theta)$ be an arbitrary function of the parameter, and let G be the image of Ω under the function g. For each $t \in G$, define $G_t = \{\theta : g(\theta) = t\}$ and define

$$L^*(t) = \max_{\theta \in G_t} \log f_n(\boldsymbol{x}|\theta).$$

Finally, define the M.L.E. of $g(\theta)$ to be $\hat{t}$ where

$$L^*(\hat{t}) = \max_{t \in G} L^*(t). \tag{7.6.1}$$

The following result shows how to find the M.L.E. of $g(\theta)$ based on Definition 7.6.1.

Theorem 7.6.2 Let $\hat{\theta}$ be an M.L.E. of θ, and let $g(\theta)$ be a function of θ. Then an M.L.E. of $g(\theta)$ is $g(\hat{\theta})$.

Proof We shall prove that $\hat{t} = g(\hat{\theta})$ satisfies (7.6.1). Since $L^*(t)$ is the maximum of $\log f_n(\boldsymbol{x}|\theta)$ over θ in a subset of Ω, and since $\log f_n(\boldsymbol{x}|\hat{\theta})$ is the maximum over all θ, we know that $L^*(t) \le \log f_n(\boldsymbol{x}|\hat{\theta})$ for all $t \in G$. Let $\hat{t} = g(\hat{\theta})$. We are done if we can show that $L^*(\hat{t}) = \log f_n(\boldsymbol{x}|\hat{\theta})$. Note that $\hat{\theta} \in G_{\hat{t}}$. Since $\hat{\theta}$ maximizes $f_n(\boldsymbol{x}|\theta)$ over all θ, it also maximizes $f_n(\boldsymbol{x}|\theta)$ over $\theta \in G_{\hat{t}}$. Hence, $L^*(\hat{t}) = \log f_n(\boldsymbol{x}|\hat{\theta})$ and $\hat{t} = g(\hat{\theta})$ is an M.L.E. of $g(\theta)$. ■

Example 7.6.3 Estimating the Standard Deviation and the Second Moment. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. We shall determine the M.L.E. of the standard deviation σ and the M.L.E. of the second moment of the normal distribution $E(X^2)$. It was found in Example 7.5.6 that the M.L.E. of $\theta = (\mu, \sigma^2)$ is $\hat{\theta} = (\hat{\mu}, \widehat{\sigma^2})$. From the invariance property, we can conclude that the M.L.E. $\hat{\sigma}$ of the standard deviation is simply the square root of the sample variance. In symbols, $\hat{\sigma} = (\widehat{\sigma^2})^{1/2}$. Also, since $E(X^2) = \sigma^2 + \mu^2$, the M.L.E. of $E(X^2)$ will be $\widehat{\sigma^2} + \hat{\mu}^2$. ◀

Consistency

Consider an estimation problem in which a random sample is to be taken from a distribution involving a parameter θ. Suppose that for every sufficiently large sample

size n, that is, for every value of n greater than some given minimum number, there exists a unique M.L.E. of θ. Then, under certain conditions, which are typically satisfied in practical problems, the sequence of M.L.E.'s is a consistent sequence of estimators of θ. In other words, in such problems the sequence of M.L.E.'s converges in probability to the unknown value of θ as $n \to \infty$.

We have remarked in Sec. 7.4 that under certain general conditions the sequence of Bayes estimators of a parameter θ is also a consistent sequence of estimators. Therefore, for a given prior distribution and a sufficiently large sample size n, the Bayes estimator and the M.L.E. of θ will typically be very close to each other, and both will be very close to the unknown value of θ.

We shall not present any formal details of the conditions that are needed to prove this result. (Details can be found in chapter 7 of Schervish, 1995.) We shall, however, illustrate the result by considering again a random sample $X_1, \ldots, X_n$ from the Bernoulli distribution with parameter θ, which is unknown ($0 \le \theta \le 1$). It was shown in Sec. 7.4 that if the given prior distribution of θ is a beta distribution, then the difference between the Bayes estimator of θ and the sample mean $\overline{X}_n$ converges to 0 as $n \to \infty$. Furthermore, it was shown in Example 7.5.4 that the M.L.E. of θ is $\overline{X}_n$. Thus, as $n \to \infty$, the difference between the Bayes estimator and the M.L.E. will converge to 0. Finally, the law of large numbers (Theorem 6.2.4) says that the sample mean $\overline{X}_n$ converges in probability to θ as $n \to \infty$. Therefore, both the sequence of Bayes estimators and the sequence of M.L.E.'s are consistent sequences.

Numerical Computation

In many problems there exists a unique M.L.E. $\hat{\theta}$ of a given parameter θ, but this M.L.E. cannot be expressed in closed form as a function of the observations in the sample. In such a problem, for a given set of observed values, it is necessary to determine the value of $\hat{\theta}$ by numerical computation. We shall illustrate this situation by two examples.

Example 7.6.4 Sampling from a Gamma Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution for which the p.d.f. is as follows:

$$f(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \quad \text{for } x > 0. \tag{7.6.2}$$

Suppose also that the value of α is unknown ($\alpha > 0$) and is to be estimated.

The likelihood function is

$$f_n(\boldsymbol{x}|\alpha) = \frac{1}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(- \sum_{i=1}^{n} x_i \right). \tag{7.6.3}$$

The M.L.E. of α will be the value of α that satisfies the equation

$$\frac{\partial \log f_n(\boldsymbol{x}|\alpha)}{\partial \alpha} = 0. \tag{7.6.4}$$

When we apply Eq. (7.6.4) in this example, we obtain the following equation:

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log x_i. \tag{7.6.5}$$

Tables of the function $\Gamma'(\alpha)/\Gamma(\alpha)$, which is called the *digamma function*, are included in various published collections of mathematical tables. The digamma function is also available in several mathematical software packages. For all given values

of $x_1, \ldots, x_n$, the unique value of α that satisfies Eq. (7.6.5) must be determined either by referring to these tables or by carrying out a numerical analysis of the digamma function. This value will be the M.L.E. of α. ◀

Example 7.6.5 Sampling from a Cauchy Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from a Cauchy distribution centered at an unknown point θ $(-\infty < \theta < \infty)$, for which the p.d.f. is as follows:

$$f(x|\theta) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]} \quad \text{for } -\infty < x < \infty. \tag{7.6.6}$$

Suppose also that the value of θ is to be estimated.

The likelihood function is

$$f_n(\boldsymbol{x}|\theta) = \frac{1}{\pi^n \prod_{i=1}^n \left[1 + (x_i - \theta)^2\right]}. \tag{7.6.7}$$

Therefore, the M.L.E. of θ will be the value that minimizes

$$\prod_{i=1}^n \left[1 + (x_i - \theta)^2\right]. \tag{7.6.8}$$

For most values of $x_1, \ldots, x_n$, the value of θ that minimizes the expression (7.6.8) must be determined by a numerical computation. ◀

An alternative to exact solution of Eq. (7.6.4) is to start with a heuristic estimator of α and then apply Newton's method.

Definition 7.6.2 Newton's Method. Let $f(\theta)$ be a real-valued function of a real variable, and suppose that we wish to solve the equation $f(\theta) = 0$. Let θ_0 be an initial guess at the solution. *Newton's method* replaces the initial guess with the updated guess

$$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}.$$

The rationale behind Newton's method is illustrated in Fig. 7.7. The function $f(\theta)$ is the solid curve. Newton's method approximates the curve by a line tangent to the curve, that is, the dashed line passing through the point $(\theta_0, f(\theta_0))$, indicated by the circle. The approximating line crosses the horizontal axis at the revised guess θ_1. Typically, one replaces the intial guess with the revised guess and iterates Newton's method until the results stabilize.

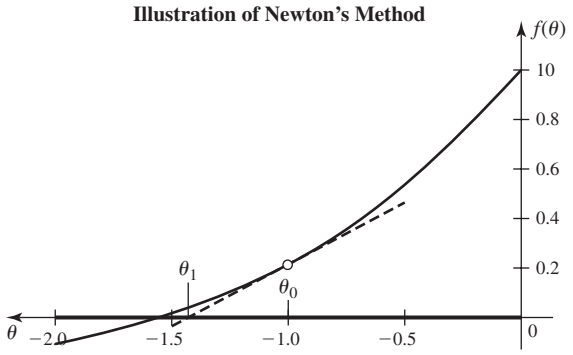

Figure 7.7 Newton's method to approximate the solution to $f(\theta) = 0$. The initial guess is θ_0, and the revised guess is θ_1.

Example 7.6.6 Sampling from a Gamma Distribution. In Example 7.6.4, suppose that we observe $n = 20$ gamma random variables $X_1, \ldots, X_{20}$ with parameters α and 1. Suppose that the observed values are such that $\frac{1}{20}\sum_{i=1}^{20} \log(x_i) = 1.220$ and $\frac{1}{20}\sum_{i=1}^{20} x_i = 3.679$. We wish to use Newton's method to approximate the M.L.E. A sensible initial guess is based on the fact that $E(X_i) = \alpha$. This suggests using $\alpha_0 = 3.679$, the sample mean. The function $f(\alpha)$ is $\psi(\alpha) - 1.220$, where ψ is the digamma function. The derivative $f'(\alpha)$ is $\psi'(\alpha)$, which is known as the trigamma function. Newton's method updates the intial guess α_0 to

$$\alpha_1 = \alpha_0 - \frac{\psi(\alpha_0) - 1.220}{\psi'(\alpha_0)} = 3.679 - \frac{1.1607 - 1.220}{0.3120} = 3.871.$$

Here, we have used statistical software that computes both the digamma and trigamma functions. After two more iterations, the approximation stabilizes at 3.876. ◀

Newton's method can fail terribly if $f'(\theta)/f(\theta)$ gets close to 0 between θ_0 and the actual solution to $f(\theta) = 0$. There is a multidimensional version of Newton's method, which we will not present here. There are also many other numerical methods for maximizing functions. Any text on numerical optimization, such as Nocedal and Wright (2006), will describe some of them.

Method of Moments

Example 7.6.7 Sampling from a Gamma Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution with parameters α and β. In Example 7.6.4, we explained how one could find the M.L.E. of α if β were known. The method involved the digamma function, which is unfamiliar to many people. A Bayes estimate would also be difficult to find in this example because we would have to integrate a function that includes a factor of $1/\Gamma(\alpha)^n$. Is there no other way to estimate the vector parameter θ in this example? ◀

The method of moments is an intuitive method for estimating parameters when other, more attractive, methods may be too difficult. It can also be used to obtain an initial guess for applying Newton's method.

Definition 7.6.3 Method of Moments. Assume that $X_1, \ldots, X_n$ form a random sample from a distribution that is indexed by a k-dimensional parameter θ and that has at least k finite moments. For $j = 1, \ldots, k$, let $\mu_j(\theta) = E(X_1^j|\theta)$. Suppose that the function $\mu(\theta) = (\mu_1(\theta), \ldots, \mu_k(\theta))$ is a one-to-one function of θ. Let $M(\mu_1, \ldots, \mu_k)$ denote the inverse function, that is, for all θ,

$$\theta = M(\mu_1(\theta), \ldots, \mu_k(\theta)).$$

Define the *sample moments* by $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ for $j = 1, \ldots, k$. The *method of moments estimator* of θ is $M(m_1, \ldots, m_j)$.

The usual way of implementing the method of moments is to set up the k equations $m_j = \mu_j(\theta)$ and then solve for θ.

Example 7.6.8 Sampling from a Gamma Distribution. In Example 7.6.4, we considered a sample of size n from the gamma distribution with parameters α and 1. The mean of each

such random variable is $\mu_1(\alpha) = \alpha$. The method of moments estimator is then $\hat{\alpha} = m_1$, the sample mean. This was the initial guess used to start Newton's method in Example 7.6.6. ◀

Example 7.6.9 Sampling from a Gamma Distribution with Both Parameters Unknown. Theorem 5.7.5 tells us that the first two moments of the gamma distribution with parameters α and β are

$$\mu_1(\theta) = \frac{\alpha}{\beta},$$
$$\mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2}.$$

The method of moments says to replace the right-hand sides of these equations by the sample moments and then solve for α and β. In this case, we get

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2},$$
$$\hat{\beta} = \frac{m_1}{m_2 - m_1^2}$$

as the method of moments estimators. Note that $m_2 - m_1^2$ is just the sample variance. ◀

Example 7.6.10 Sampling from a Uniform Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta, \theta + 1]$, as in Example 7.5.9. In that example, we found that the M.L.E. is not unique and there is an interval of M.L.E.'s

$$\max\{x_1, \ldots, x_n\} - 1 \le \theta \le \min\{x_1, \ldots, x_n\}. \tag{7.6.9}$$

This interval contains all of the possible values of θ that are consistent with the observed data. We shall now apply the method of moments, which will produce a single estimator. The mean of each X_i is $\theta + 1/2$, so the method of moments estimator is $\overline{X}_n - 1/2$. Typically, one would expect the observed value of the method of moments estimator to be a number in the interval (7.6.9). However, that is not always the case. For example, if $n = 3$ and $X_1 = 0.2$, $X_2 = 0.99$, $X_3 = 0.01$ are observed, then (7.6.9) is the interval $[-0.01, 0.01]$, while $\overline{X}_3 = 0.4$. The method of moments estimate is then -0.1, which could not possibly be the true value of θ. ◀

There are several examples in which method of moments estimators are also M.L.E.'s. Some of these are the subjects of exercises at the end of this section.

Despite occasional problems such as Example 7.6.10, the method of moments estimators will typically be consistent in the sense of Definition 7.4.6.

Theorem 7.6.3 Suppose that $X_1, X_2, \ldots$ are i.i.d. with a distribution indexed by a k-dimensional parameter vector θ. Suppose that the first k moments of that distribution exist and are finite for all θ. Suppose also that the inverse function M in Definition 7.6.3 is continuous. Then the sequence of method of moments estimators based on $X_1, \ldots, X_n$ is a consistent sequence of estimators of θ.

Proof The law of large numbers says that the sample moments converge in probability to the moments $\mu_1(\theta), \ldots, \mu_k(\theta)$. The generalization of Theorem 6.2.5 to

functions of k variables implies that M evaluated at the sample moments (i.e., the method of moments estimator) converges in probability to θ. ■

M.L.E.'s and Bayes Estimators

Bayes estimators and M.L.E.'s depend on the data solely through the likelihood function. They use the likelihood function in different ways, but in many problems they will be very similar. When the function $f(x|\theta)$ satisfies certain smoothness conditions (as a function of θ), it can be shown that the likelihood function will tend to look more and more like a normal p.d.f. as the sample size increases. More specifically, as n increases, the likelihood function starts to look like a constant (not depending on θ, but possibly depending on the data) times

$$\exp\left[-\frac{1}{2V_n(\theta)/n}(\theta-\hat{\theta})^2\right], \tag{7.6.10}$$

where $\hat{\theta}$ is the M.L.E. and $V_n(\theta)$ is a sequence of random variables that typically converges as $n \to \infty$ to a limit that we shall call $v_\infty(\theta)$. When n is large, the function in (7.6.10) rises quickly to its peak as θ approaches $\hat{\theta}$ and then drops just as quickly as θ moves away from $\hat{\theta}$. Under these conditions, so long as the prior p.d.f. of θ is relatively flat compared to the very peaked likelihood function, the posterior p.d.f. will look a lot like the likelihood multiplied by the constant needed to turn it into a p.d.f. The posterior mean of θ will then be approximately $\hat{\theta}$. In fact, the posterior distribution of θ will be approximately the normal distribution with mean $\hat{\theta}$ and variance $V_n(\hat{\theta})/n$. In similar fashion, the distribution of the maximum likelihood estimator (given θ) will be approximately the normal distribution with mean θ and variance $v_\infty(\theta)/n$. The conditions and proofs needed to make these claims precise are beyond the scope of this text but can be found in chapter 7 of Schervish (1995).

Example 7.6.11

Sampling from an Exponential Distribution. Suppose that $X_1, X_2, \ldots$ are i.i.d. having the exponential distribution with parameter θ. Let $T_n = \sum_{i=1}^n X_i$. Then the M.L.E. of θ is $\hat{\theta}_n = n/T_n$. (This was found in Exercise 7 in Sec. 7.5.) Because $1/\hat{\theta}_n$ is an average of i.i.d. random variables with finite variance, the central limit theorem tells us that the distribution of $1/\hat{\theta}_n$ is approximately normal. The mean and variance, in this case, of that approximate normal distribution are, respectively, $1/\theta$ and $1/(\theta^2 n)$. The delta method (Theorem 6.3.2) says that $\hat{\theta}$ then has approximately the normal distribution with mean θ and variance θ^2/n. In the notation above, we have $V_n(\theta) = \theta^2$.

Next, let the prior distribution of θ be the gamma distribution with parameters α and β. Theorem 7.3.4 says that the posterior distribution of θ will be the gamma distribution with parameters $\alpha + n$ and $\beta + t_n$. We conclude by showing that this gamma distribution is approximately a normal distribution. Assume for simplicity that α is an integer. Then the posterior distribution of θ is the same as the distribution of the sum of $\alpha + n$ i.i.d. exponential random variables with parameter $\beta + t_n$. Such a sum has approximately the normal distribution with mean $(\alpha + n)/(\beta + t_n)$ and variance $(\alpha + n)/(\beta + t_n)^2$. If α and β are small, the approximate mean is then nearly $n/t_n = \hat{\theta}$, and the approximate variance is then nearly $n/t_n^2 = \hat{\theta}^2/n = V_n(\hat{\theta})/n$. ◀

Example 7.6.12

Prussian Army Deaths. In Example 7.3.14, we found the posterior distribution of θ, the mean number of deaths per year by horsekick in Prussian army units based on a sample of 280 observations. The posterior distribution was found to be the gamma distribution with parameters 196 and 280. By the same argument used in

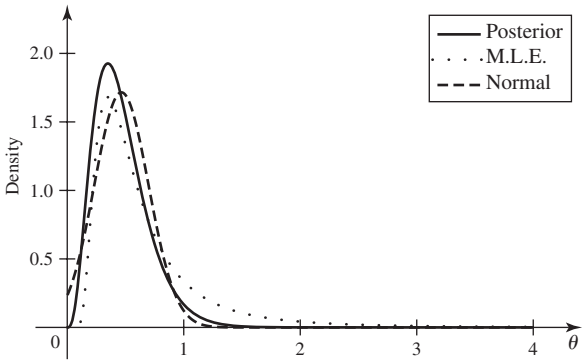

Figure 7.8 Posterior p.d.f. together with p.d.f. of M.L.E. and approximating normal p.d.f. in Example 7.6.13. For the p.d.f of the M.L.E., the value of $\theta = 3/6.6$ is used to make the p.d.f.'s as similar as possible.

Example 7.6.11, this gamma distribution is approximately the distribution of the sum of 196 i.i.d. exponential random variables with parameter 280. The distribution of this sum is approximately the normal distribution with mean $196/280$ and variance $196/280^2$.

Using the same data as in Example 7.3.14, we can find the M.L.E. of θ, which is the average of the 280 observations (according to Exercise 5 in Sec. 7.5). The distribution of the average of 280 i.i.d. Poisson random variables with mean θ is approximately the normal distribution with mean θ and variance $\theta/280$ according to the central limit theorem. We then have $V_n(\theta) = \theta$ in the earlier notation. The maximum likelihood estimate with the observed data is $\hat{\theta} = 196/280$ the mean of the posterior distribution. The variance of the posterior distribution is also $V_n(\hat{\theta})/n = \hat{\theta}/280$. ◀

There are two common situations in which posterior distributions and distributions of M.L.E.'s are not such similar normal distributions as in the preceding discussion. One is when the sample size is not very large, and the other is when the likelihood function is not smooth. An example with small sample size is our electronic components example.

Example 7.6.13

Lifetimes of Electronic Components. In Example 7.3.12, we have a sample of $n = 3$ exponential random variables with parameter θ. The posterior distribution found there was the gamma distribution with parameters 4 and 8.6. The M.L.E. is $\hat{\theta} = 3/(X_1 + X_2 + X_3)$, which has the distribution of 1 over a gamma random variable with parameters 3 and 3θ. Figure 7.8 shows the posterior p.d.f. along with the p.d.f. of the M.L.E. assuming that $\theta = 3/6.6$, the observed value of the M.L.E. The two p.d.f.'s, although similar, are still different. Also, both p.d.f.'s are similar to, but still different from, the normal p.d.f. with the same mean and variance as the posterior, which also appears on the plot. ◀

An example of an unsmooth likelihood function involves the uniform distribution on the interval $[0, \theta]$.

Example 7.6.14

Sampling from a Uniform Distribution. In Example 7.5.7, we found the M.L.E. of θ based on a sample of size n from the uniform distribution on the interval $[0, \theta]$. The M.L.E. is $\hat{\theta} = \max\{X_1, \ldots, X_n\}$. We can find the exact distribution of $\hat{\theta}$ using the result in Example 3.9.6. The p.d.f. of $Y = \hat{\theta}$ is

$$g_n(y|\theta) = n[F(y|\theta)]^{n-1} f(y|\theta), \tag{7.6.11}$$

where $f(\cdot|\theta)$ is the p.d.f. of the uniform distribution on $[0, \theta]$ and $F(\cdot|\theta)$ is the corresponding c.d.f. Substituting these well-known functions into Eq. (7.6.11) yields the p.d.f. of $Y = \hat{\theta}$:

$$g_n(y|\theta) = n\left[\frac{y}{\theta}\right]^{n-1}\frac{1}{\theta} = n\frac{y^{n-1}}{\theta^n},$$

for $0 < y < \theta$. This p.d.f. is not the least bit like a normal p.d.f. It is very asymmetric and has its maximum at the largest possible value of the M.L.E. In fact, one can compute the mean and variance of $\hat{\theta}$, respectively, as

$$E(\hat{\theta}) = \frac{n}{n+1}\theta,$$
$$Var(\hat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

The variance goes down like $1/n^2$ instead of like $1/n$ in the approximately normal examples we saw earlier.

If n is large, the posterior distribution of θ will have a p.d.f. that is approximately the likelihood function times the constant needed to make it into a p.d.f. The likelihood is in Eq. (7.5.8). Integrating that function over θ to obtain the needed constant leads to the following approximate posterior p.d.f. of θ:

$$\xi(\theta|\boldsymbol{x}) \approx \begin{cases} \dfrac{(n-1)\hat{\theta}^{n-1}}{\theta^n} & \text{for } \theta > \hat{\theta}, \\ 0 & \text{otherwise.} \end{cases}$$

The mean and variance of this approximate posterior distribution are, respectively, $(n-1)\hat{\theta}/(n-2)$ and $(n-1)\hat{\theta}^2/[(n-2)^2(n-3)]$. The posterior mean is still nearly equal to the M.L.E. (but a little larger), and the posterior variance decreases at a rate like $1/n^2$, as does the variance of the M.L.E. But the posterior distribution is not the least bit normal, as the p.d.f. has its maximum at the smallest possible value of θ and decreases from there. ◀

The EM Algorithm

There are a number of complicated situations in which it is difficult to compute the M.L.E. Many of these situations involve forms of missing data. The term "missing data" can refer to several different types of information. The most obvious would be observations that we had planned or hoped to observe but were not observed. For example, imagine that we planned to collect both heights and weights for a sample of athletes. For reasons that might be beyond our control, it is possible that we observed both heights and weights for most of the athletes, but only heights for one subset of atheletes and only weights for another subset. If we model the heights and weights as having a bivariate normal distribution, we might want to compute the M.L.E. of the parameters of that distribution. For a complete collection of pairs, Exercise 24 in this section gives formulas for the M.L.E. It is not difficult to see how much more complicated it would be to compute the M.L.E. in the situation described above with missing data.

The *EM algorithm* is an iterative method for approximating M.L.E.'s when missing data are making it difficult to find the M.L.E.'s in closed form. One begins (as in most iterative procedures) at stage 0 with an initial parameter vector $\theta^{(0)}$. To move from stage j to stage $j+1$, one first writes the *full-data log-likelihood*, which is what the logarithm of the likelihood function would be if we had observed the

missing data. The values of the missing data appear in the full-data log-likelihood as random variables rather than as observed values. The "E" step of the EM algorithm is the following: Compute the conditional distribution of the missing data given the observed data as if the parameter θ were equal to $\theta^{(j)}$, and then compute the conditional mean of the full-data log-likelihood treating θ as constant and the missing data as random variables. The E step gets rid of the unobserved random variables from the full-data log-likelihood and leaves θ where it was. For the "M" step, choose $\theta^{(j+1)}$ to maximize the expected value of the full-data log-likelihood that you just computed. The M step takes you to stage $j+1$. Ideally, the maximization step is no harder than it would be if the missing data had actually been observed.

Example 7.6.15 Heights and Weights. Suppose that we try to observe $n = 6$ pairs of heights and weights, but we get only three complete vectors plus one lone weight and two lone heights. We model the pairs as bivariate normal random vectors, and we want to find the M.L.E. of the parameter vector $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. (This example is for illustrative purposes. One cannot expect to get a good estimate of a five-dimensional parameter vector with only nine observed values and no prior information.) The data are in Table 7.1. The missing weights are $X_{4,2}$ and $X_{5,2}$. The missing height is $X_{6,1}$. The full-data log-likelihood is the sum of the logarithms of six expressions of the form Eq. (5.10.2) each with one of the rows of Table 7.1 substituted for the dummy variables (x_1, x_2). For example, the term corresponding to the fourth row of Table 7.1 is

$$-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left[\left(\frac{68-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{68-\mu_1}{\sigma_1}\right)\left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right) + \left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right)^2\right]. \tag{7.6.12}$$

As an initial parameter vector we choose a naïve estimate computed from the observed data:

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}, \rho^{(0)}) = (69.60, 194.75, 2.87, 14.82, 0.1764).$$

This consists of the M.L.E.'s based on the marginal distributions of the two coordinates, together with the sample correlation computed from the three complete observations.

Table 7.1 Heights and weights for Example 7.6.15. The missing values are given random variable names.

Height	Weight
72	197
70	204
73	208
68	$X_{4,2}$
65	$X_{5,2}$
$X_{6,1}$	170

The E step pretends that $\theta = \theta^{(0)}$ and computes the conditional mean of the full-data log-likelihood given the observed data. For the fourth row of Table 7.1, the conditional distribution of $X_{4,2}$ given the observed data and $\theta = \theta^{(0)}$ can be found from Theorem 5.10.4 to be the normal distribution with mean

$$194.75 + 0.1764 \times (14.82)^{1/2} \left(\frac{68 - 69.60}{2.87^{1/2}} \right) = 193.3$$

and variance $(1 - 0.1764^2)14.82^2 = 212.8$. The conditional mean of $(X_{4,2} - \mu_2)^2$ would then be $212.8 + (193.3 - \mu_2)^2$. The conditional mean of the expression in (7.6.12) would then be

$$-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\left[\left(\frac{68-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{68-\mu_1}{\sigma_1}\right)\left(\frac{193.3-\mu_2}{\sigma_2}\right) + \left(\frac{193.3-\mu_2}{\sigma_2}\right)^2 + \frac{212.8}{\sigma_2^2}\right].$$

The point to notice about this last expression is that, except for the last term $212.8/\sigma_2^2$, it is exactly the contribution to the log-likelihood that we would have obtained if $X_{4,2}$ had been observed to equal 193.3, its conditional mean. Similar calculations can be done for the other two observations with missing coordinates. Each will produce a contribution to the log-likelihood that is the conditional variance of the missing coordinate divided by its variance plus what the log-likelihood would have been if the missing value had been observed to equal its conditional mean. This makes the M step almost identical to finding the M.L.E. for a completely observed data set. The only difference from the formulas in Exercise 24 is the following: For each observation that is missing X, add the conditional variance of X given Y to $\sum_{i=1}^n (X_i - \overline{X}_n)^2$ in both the formula for $\widehat{\sigma_1^2}$ and $\hat{\rho}$. Similarly, for each observation that is missing Y, add the conditional variance of Y given X to $\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ in both the formula for $\widehat{\sigma_2^2}$ and $\hat{\rho}$.

We now illustrate the first iteration of the EM algorithm with the data of this example. We already have $\theta^{(0)}$, and we can compute the log-likelihood function from the observed data at $\theta^{(0)}$ as -31.359. To begin the algorithm, we have already computed the conditional mean and variance of the missing second coordinate from the fourth row of Table 7.1. The corresponding conditional means and variances for the fifth and sixth rows are 190.6 and 212.8 for the fifth row and 68.76 and 7.98 for the sixth row. For the E step, we replace the missing observations by their conditional means and add the conditional variances to the sums of squared deviations. For the M step, we insert the values just computed into the formulas of Exercise 24 as described above. The new vector is

$$\theta^{(1)} = (69.46, 193.81, 2.88, 14.83, 0.3742),$$

and the log-likelihood is -31.03. After 32 iterations, the estimate and log-likelihood stop changing. The final estmate is

$$\theta^{(32)} = (68.86, 189.71, 3.15, 15.03, 0.8965),$$

with log-likelihood -29.66. ◀

Example 7.6.16 Mixture of Normal Distributions. A very popular use of the EM algorithm is in fitting mixture distributions. Let $X_1, \ldots, X_n$ be random variables such that each one is

sampled either from the normal distribution with mean μ_1 and variance σ^2 (with probability p) or from the normal distribution with mean μ_2 and variance σ^2 (with probability $1-p$), where $\mu_1 < \mu_2$. The restriction that $\mu_1 < \mu_2$ is to make the model identifiable in the following sense. If $\mu_1 = \mu_2$ is allowed, then every value of p leads to the same joint distribution of the observable data. Also, if neither mean is constrained to be below the other, then switching the two means and changing p to $1-p$ will produce the same joint distribution for the observable data. The restriction $\mu_1 < \mu_2$ ensures that every distinct parameter vector produces a different joint distribution for the observable data.

The data in Fig. 7.4 have the typical appearance of a distribution that is a mixture of two normals with means not very far apart. Because we have assumed that the variances of the two distributions are the same, we will not have the problem that arose in Example 7.5.10.

The likelihood function from observations $X_1 = x_1, \ldots, X_n = x_n$ is

$$\prod_{i=1}^{n}\left[\frac{p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_1)^2}{2\sigma^2}\right)+\frac{1-p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_2)^2}{2\sigma^2}\right)\right]. \quad (7.6.13)$$

The parameter vector is $\theta = (\mu_1, \mu_2, \sigma^2, p)$, and maximizing the likelihood as written is a challenge. However, we can introduce missing observations $Y_1, \ldots, Y_n$ where $Y_i = 1$ if X_i was sampled from the distribution with mean μ_1 and $Y_i = 0$ if X_i was sampled from the distribution with mean μ_2. The full-data log-likelihood can be written as the sum of the logarithm of the marginal p.f. of the missing Y data plus the logarithm of the conditional p.d.f. of the observed X data given the Y data. That is,

$$\begin{aligned}\sum_{i=1}^{n} Y_i \log(p) + \left(n - \sum_{i=1}^{n} Y_i\right)\log(1-p) - \frac{n}{2}\log(2\pi\sigma^2)\\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i(x_i-\mu_1)^2 + (1-Y_i)(x_i-\mu_2)^2\right].\end{aligned} \quad (7.6.14)$$

At stage j with estimate $\theta^{(j)}$ of θ, the E step first finds the conditional distribution of $Y_1, \ldots, Y_n$ given the observed data and $\theta = \theta^{(j)}$. Since $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent pairs, we can find the conditional distribution separately for each pair. The joint distribution of (X_i, Y_i) is a mixed distribution with p.f./p.d.f.

$$f(x_i, y_i|\theta^{(j)}) = \frac{p^{y_i}(1-p)^{1-y_i}}{(2\pi)^{1/2}\sigma^{(j)}}\exp\left(-\frac{1}{\sigma^{2(j)}}\left[y_i(x_i-\mu_1^{(j)})^2 + (1-y_i)(x_i-\mu_2^{(j)})^2\right]\right).$$

The marginal p.d.f. of X_i is the ith factor in (7.6.13). It is straightforward to determine that the conditional distribution of Y_i given the observed data is the Bernoulli distribution with parameter

$$q_i^{(j)} = \frac{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right)}{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right)+(1-p^{(j)})\exp\left(-\frac{(x_i-\mu_2^{(j)})^2}{2\sigma^{2(j)}}\right)}. \quad (7.6.15)$$

Because the full-data log-likelihood is a linear function of the Y_i's, the E step simply replaces each Y_i in (7.6.14) by $q_i^{(j)}$. The result is

$$\sum_{i=1}^{n} q_i^{(j)} \log(p) + \left(n - \sum_{i=1}^{n} q_i^{(j)}\right) \log(1-p) - \frac{n}{2}\log(2\pi\sigma^2) \\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i-\mu_1)^2 + (1-q_i^{(j)})(x_i-\mu_2)^2\right]. \tag{7.6.16}$$

Maximizing (7.6.16) is straightforward. Since p appears in only the first two terms, we see that $p^{(j+1)}$ is just the average of the $q_i^{(j)}$'s. Also, $\mu_1^{(j+1)}$ is the weighted average of the X_i's with weights $q_i^{(j)}$. Similarly, $\mu_2^{(j+1)}$ is the weighted average of the X_i's with weights $1 - q_i^{(j)}$. Finally,

$$\sigma^{2(j+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i-\mu_1^{(j+1)})^2 + (1-q_i^{(j)})(x_i-\mu_2^{(j+1)})^2\right]. \tag{7.6.17}$$

We will illustrate the first E and M steps using the data in Example 7.3.10. For the initial parameter vector $\theta^{(0)}$, we will let $\mu_1^{(0)}$ be the average of the 10 lowest observations and $\mu_2^{(0)}$ be the average of the 10 highest observations. We set $p^{(0)} = 1/2$, and $\sigma^{2(0)}$ is the average of the sample variance of the 10 lowest observations and the sample variance of the 10 highest observations. This makes

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma^{2(0)}, p^{(0)}) = (-7.65, 7.36, 46.28, 0.5).$$

For each of the 20 observed values x_i, we compute $q_i^{(0)}$. For example, $x_{10} = -4.0$. According to (7.6.15),

$$q_{10}^{(0)} = \frac{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right)}{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right) + 0.5\exp\left(-\frac{(-4.0-7.36)^2}{2\times 46.28}\right)} = 0.7774.$$

A similar calculation for $x_8 = 9.0$ yields $q_8^{(0)} = 0.0489$. The initial log-likelihood, calculated as the logarithm of (7.6.13), is -75.98. The average of the 20 $q_i^{(0)}$ values is $p^{(1)} = 0.4402$. The weighted average of the data values using the $q_i^{(0)}$'s as weights is $\mu_1^{(1)} = -7.736$, and the weighted average using the $1 - q_i^{(0)}$'s is $\mu_2^{(1)} = 6.3068$. Using (7.6.17), we get $\sigma^{2(1)} = 56.5491$. The log-likelihood rises to -75.19. After 25 iterations, the results settle on $\theta^{(25)} = (-21.9715, 2.6802, 48.6864, 0.1037)$ with a final log-likelihood of -72.84. The histogram from Fig. 7.4 is reproduced in Fig. 7.9 together with the p.d.f. of an observation from the fitted mixture distribution, namely,

$$f(x) = \frac{0.1037}{(2\pi \times 48.6864)^{1/2}}\exp\left(-\frac{(x+21.9715)^2}{2\times 48.6864}\right) \\ + \frac{1-0.1037}{(2\pi \times 48.6864)^{1/2}}\exp\left(-\frac{(x-2.6802)^2}{2\times 48.6864}\right).$$

In addition, the fitted p.d.f. based on a single normal distribution is also shown in Fig. 7.9. The mean and variance of that single normal distribution are 0.1250 and 110.6809, respectively. ◀

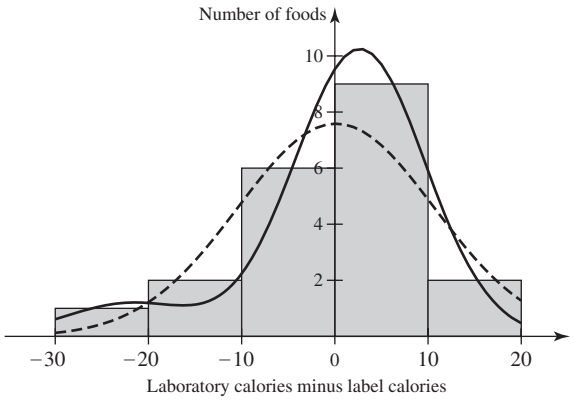

Figure 7.9 Histogram of data from Example 7.3.10 together with fitted p.d.f. from Example 7.6.16 (solid curve). The p.d.f. has been scaled up to match the fact that the histogram gives counts rather than an estimated p.d.f. Also, the dashed curve gives the estimated p.d.f. for a single normal distribution.

One can prove that the log-likelihood increases with each iteration of the EM algorithm and that the algorithm converges to a local maximum of the likelihood function. As with other numerical maximization routines, it is difficult to guarantee convergence to a global maximum.

Sampling Plans

Suppose that an experimenter wishes to take observations from a distribution for which the p.f. or the p.d.f. is $f(x|\theta)$ in order to gain information about the value of the parameter θ. The experimenter could simply take a random sample of a predetermined size from the distribution. Instead, however, he may begin by first observing a few values at random from the distribution and noting the cost and the time spent in taking these observations. He may then decide to observe a few more values at random from the distribution and to study all the values thus far obtained. At some point, the experimenter will decide to stop taking observations and will estimate the value of θ from all the observed values that have been obtained up to that point. He might decide to stop because either he feels that he has enough information to be able to make a good estimate of θ or he cannot afford to spend any more money or time on sampling.

In this experiment, the number n of observations in the sample is not fixed beforehand. It is a random variable whose value may very well depend on the magnitudes of the observations as they are obtained.

Suppose that an experimenter contemplates using a sampling plan in which, for every n, the decision of whether or not to stop sampling after n observations have been collected is a function of the n observations seen so far. Regardless of whether the experimenter chooses such a sampling plan or decides to fix the value of n before

any observations are taken, it can be shown that the likelihood function based on the observed values is proportional (as a function of θ) to

$$f(x_1|\theta) \ldots f(x_n|\theta).$$

In such a situation, the M.L.E. of θ will depend only on the likelihood function and not on what type of sampling plan is used. In other words, the value of $\hat{\theta}$ depends only on the values $x_1, \ldots, x_n$ that are actually observed and does not depend on the plan (if there was one) that was used by the experimenter to decide when to stop sampling.

To illustrate this property, suppose that the intervals of time, in minutes, between arrivals of successive customers at a certain service facility are i.i.d. random variables. Suppose also that each interval has the exponential distribution with parameter θ, and that a set of observed intervals $X_1, \ldots, X_n$ form a random sample from this distribution. It follows from Exercise 7 of Sec. 7.5 that the M.L.E. of θ will be $\hat{\theta} = 1/\overline{X}_n$. Also, since the mean μ of the exponential distribution is $1/\theta$, it follows from the invariance property of M.L.E.'s that $\hat{\mu} = \overline{X}_n$. In other words, the M.L.E. of the mean is the average of the observations in the sample.

Consider now the following three sampling plans:

1. An experimenter decides in advance to take exactly 20 observations, and the average of these 20 observations turns out to be 6. Then the M.L.E. of μ is $\hat{\mu} = 6$.
2. An experimenter decides to take observations $X_1, X_2 \ldots$ until she obtains a value greater than 10. She finds that $X_i < 10$ for $i = 1, \ldots, 19$ and that $X_{20} > 10$. Hence, sampling terminates after 20 observations. If the average of these 20 observations is 6, then the M.L.E. is again $\hat{\mu} = 6$.
3. An experimenter takes observations one at a time, with no particular plan in mind, until either she is forced to stop sampling or she gets tired of sampling. She is certain that neither of these causes (being forced to stop or getting tired) depends in any way on μ. If for either reason she stops as soon as she has taken 20 observations and if the average of the 20 observations is 6, then the M.L.E. is again $\hat{\mu} = 6$.

Sometimes, an experiment of this type must be terminated during an interval when the experimenter is waiting for the next customer to arrive. If a certain amount of time has elapsed since the arrival of the last customer, this time should not be omitted from the sample data, even though the full interval to the arrival of the next customer has not been observed. Suppose, for example, that the average of the first 20 observations is 6, the experimenter waits another 15 minutes but no other customer arrives, and then she terminates the experiment. In this case, we know that the M.L.E. of μ would have to be greater than 6, since the value of the 21st observation must be greater than 15, even though its exact value is unknown. The new M.L.E. can be obtained by multiplying the likelihood function for the first 20 observations by the probability that the 21st observation is greater than 15, namely, $\exp(-15\theta)$, and finding the value of θ that maximizes this new likelihood function (see Exercise 15).

Remember that the M.L.E. is determined by the likelihood function. The only way in which the M.L.E. is allowed to depend on the sampling plan is through the likelihood function. If the decision about when to stop observing data is based solely on the observations seen so far, then this information has already been included in the likelihood function. If the decision to stop is based on something else, one needs

to evaluate the probability of that "something else" given each possible value of θ and include that probability in the likelihood.

Other properties of M.L.E.'s will be discussed later in this chapter and in Chapter 8.

Summary

The M.L.E. of a function $g(\theta)$ is $g(\hat{\theta})$, where $\hat{\theta}$ is the M.L.E. of θ. For example, if θ is the rate at which customers are served in a queue, then $1/\theta$ is the average service time. The M.L.E. of $1/\theta$ is 1 over the M.L.E. of θ. Sometimes we cannot find a closed form expression for the M.L.E. of a parameter and we must resort to numerical methods to find or approximate the M.L.E. In most problems, the sequence of M.L.E.'s, as sample size increases, converges in probability to the parameter. When data are collected in such a way that the decision to stop collecting data is based solely on the data already observed or on other considerations that are not related to the parameter, then the M.L.E. will not depend on the sampling plan. That is, if two different sampling plans lead to proportional likelihood functions, then the value of θ that maximizes one likelihood will also maximize the other.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution with the p.d.f. given in Exercise 10 of Sec. 7.5. Find the M.L.E. of $e^{-1/\theta}$.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from a Poisson distribution for which the mean is unknown. Determine the M.L.E. of the standard deviation of the distribution.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown. Determine the M.L.E. of the median of the distribution.

4. Suppose that the lifetime of a certain type of lamp has an exponential distribution for which the value of the parameter β is unknown. A random sample of n lamps of this type are tested for a period of T hours and the number X of lamps that fail during this period is observed, but the times at which the failures occurred are not noted. Determine the M.L.E. of β based on the observed value of X.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[a, b]$, where both endpoints a and b are unknown. Find the M.L.E. of the mean of the distribution.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which both the mean and the variance are unknown. Find the M.L.E. of the 0.95 quantile of the distribution, that is, of the point θ such that $\Pr(X < \theta) = 0.95$.

7. For the conditions of Exercise 6, find the M.L.E. of $\nu = \Pr(X > 2)$.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from a gamma distribution for which the p.d.f. is given by Eq. (7.6.2). Find the M.L.E. of $\Gamma'(\alpha)/\Gamma(\alpha)$.

9. Suppose that $X_1, \ldots, X_n$ form a random sample from a gamma distribution for which both parameters α and β are unknown. Find the M.L.E. of α/β.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from a beta distribution for which both parameters α and β are unknown. Show that the M.L.E.'s of α and β satisfy the following equation:

$$\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \frac{\Gamma'(\hat{\beta})}{\Gamma(\hat{\beta})} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{X_i}{1 - X_i}.$$

11. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown. Show that the sequence of M.L.E.'s of θ is a consistent sequence.

12. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential distribution for which the value of the parameter β is unknown. Show that the sequence of M.L.E.'s of β is a consistent sequence.