Chapter 8
Estimation in Theory and Practice

In Section 7.2.1 we showed how the method of moments may be used to estimate the parameters of a *Gamma*(α, β) distribution, and we immediately stated that the method of maximum likelihood provides a better solution. How do we know this? In general, how should alternative methods of estimation be compared? In this chapter we lay out a series of principles that serve as guides to practice. The main ideas came from Ronald Fisher (1922); they were modified and made more precise by Jerzy Neyman (1937), and have been refined and incorporated into textbooks on statistical theory ever since, beginning notably with Cramér (1946).

Suppose we have a family of probability distributions that depends on a parameter θ, which must be estimated, and we have an estimator T. For now let us assume that θ is a scalar. If we were to say that T is a good estimator of θ, what might we mean? In particular, what might we mean when we say that maximum likelihood produces a good estimator? Clearly, for T to be a good estimator it must be "close" to θ, but because T is a random variable the notion of closeness must be stated probabilistically. For example, if we consider the mean $\bar{X}$ of a random sample $X_1, \ldots, X_n$ from a $N(\theta, 1)$ distribution, we might want to say that the mean $\bar{X}$ is close to θ when $|\bar{X} - \theta| < .1$. Because $\bar{X} \sim N(\theta, 1/n)$, even if n is large it is *possible* that $|\bar{X} - \theta| > .1$. We can not say that $|\bar{X} - \theta| < .1$. All we can say is the probability that $|\bar{X} - \theta| < .1$ is large, meaning close to one or, equivalently, the probability that $|\bar{X} - \theta| > .1$ is small, meaning close to zero.

For a general estimator T we can use the same approach and say that T is a good estimator of θ when it is *highly probable* that T is close to θ. Specifically, we introduce a tolerance ϵ, understanding that ϵ will be some small positive number, and then we require that $P(|\bar{X} - \theta| < \epsilon)$ is close to one or, equivalently, $P(|\bar{X} - \theta| > \epsilon)$ is close to zero. It is, in general, rather difficult to provide guarantees on the size of $P(|\bar{X} - \theta| > \epsilon)$ for fixed sample sizes. In most realistically complicated problems computer simulation studies must be used (as in Section 8.1.2) and they are based on specific cases so they do not provide general assurances. On the other hand, general results may be obtained asymptotically, letting the sample size grow indefinitely large. To take a concrete case, because the mean $\bar{X}$ of a random sample from a

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_8,

$N(\theta, 1)$ distribution follows a $N(\theta, 1/n)$ distribution, if we take n=10,000, from the normal cdf we find $P(|\bar{X} - \theta| > .1) = 1.5 \cdot 10^{-23}$. Indeed, no matter how small we take ϵ we have $P(|\bar{X} - \theta| > \epsilon) \to 0$ as $n \to \infty$. This is simply a restatement of the law of large numbers (p. 143)

$$\bar{X} \xrightarrow{P} \theta.$$

We discuss asymptotic results in Sections 8.2.1–8.3.1.

When we examine what happens as $n \to \infty$ it is helpful to write the generic estimator in the form $T_n = T(X_1, \ldots, X_n)$ to emphasize its dependence on n as we did in Section 7.3.5. One of the most important of the large-sample findings considers estimators that are *asymptotically normal*, as in Eq. (7.23),

$$\frac{T_n - \theta}{\sigma_{T_n}} \xrightarrow{D} N(0, 1). \tag{8.1}$$

For such estimators, in large samples, the probabilistic closeness of T_n to θ depends entirely on σ_{T_n} and we seek estimators that make σ_{T_n} as small as possible. In Sections 8.2.2–8.3.1 we go over the remarkable discovery by Fisher that σ_{T_n} can be minimized, and the minimum is obtained by the MLE. There has been a lot of theoretical work on the general subject of large-sample optimality, all of which leads to the conclusion that in well-behaved parametric problems, the method of maximum likelihood is essentially unbeatable. This, coupled with its very wide applicability (which began to be appreciated with the development of generalized linear models, see Section 14.1.6), has made maximum likelihood an essential tool in data analysis.

Fisher's theoretical insight seems to have been based on geometrical intuitions, which were elaborated in a mathematically rigorous framework by Bradley Efron in the 1970s and early 1980s. For details and references on the asymptotic arguments and their geometrical origins see Kass and Vos (1997). For a rigorous treatment in a more general context see van der Vaart (1998).

While asymptotic results are important, they have an inherent weakness: they apply when the sample size is large, but they do not say what "large" means in practice. In some cases $n = 20$ is more than adequate while in others n =20,000 is not large enough. One approach to coping with this problem is to evaluate a measure of likely deviation for specific cases, with specified sample sizes. The most common assessment of deviation of T from θ is the *mean squared error* (*MSE*) defined by

$$MSE(T) = E((T - \theta)^2). \tag{8.2}$$

In Chapter 4, p. 80, and 89, we considered the mean squared error in predicting one random variable from another. We discuss mean squared error in estimation in Section 8.1. In Section 8.4 we describe some of the practical considerations in applying ML estimation.

The most important points about ML estimation are the following:

1. ML estimation is applicable when the statistical model depends on an unknown parameter vector.[1] See Sections 7.2.2 and 8.4.1.
2. Together with ML estimates it is possible to get large-sample confidence intervals (Sections 8.2.2, 8.3.2, and 8.4.3).
3. In large samples, ML estimation is optimal (Section 8.3.1).
4. In large samples ML estimation agrees with Bayesian estimation (Section 8.3.3).

8.1 Mean Squared Error

The mean squared error criterion defined in (8.2) uses the squared magnitude of the deviation $T-\theta$ rather than its absolute value $|T-\theta|$ because it is easier to work with mathematically, and because it has a very nice decomposition given in Section 8.1.1. Intuitively, because $MSE(T)$ is an average of the values $(T-\theta)^2$, when $MSE(T)$ is small, large values of $(T-\theta)^2$ (and thus also large values of $|T-\theta|$) must be highly improbable. In fact, even more is true: we have

$$P(|T-\theta|>\epsilon) < \frac{E((T-\theta)^2)}{\epsilon^2}. \tag{8.3}$$

Thus, we can make sure it is highly probable for T to be close to θ by instead making sure that $MSE(T)$ is small.

Details: We can use Markov's inequality, which appeared as a lemma in Section 6.2.1, to guarantee that $P(|T-\theta|>\epsilon)$ will be small if $MSE(T)$ is small. First, we have

$$P(|T-\theta|>\epsilon) = P((T-\theta)^2>\epsilon^2).$$

Now, assuming $E((T-\theta)^2)<\infty$, Markov's inequality gives (8.3). □

In some cases $MSE(T)$ may be evaluated by analytical calculation, but in most practical situations computer simulation studies are used. We give two examples of such studies in Section 8.1.2.

8.1.1 Mean squared error is bias squared plus variance.

Two ways an estimator can perform poorly need to be distinguished. The first involves the systematic tendency for the estimator T to miss its target value θ. An estimator's

[1] The parameter must be finite-dimensional; in nonparametric inference the parameter is, instead, infinite-dimensional. Also, there are regularity conditions that make ML estimation work properly. See Bickel and Doksum (2001).

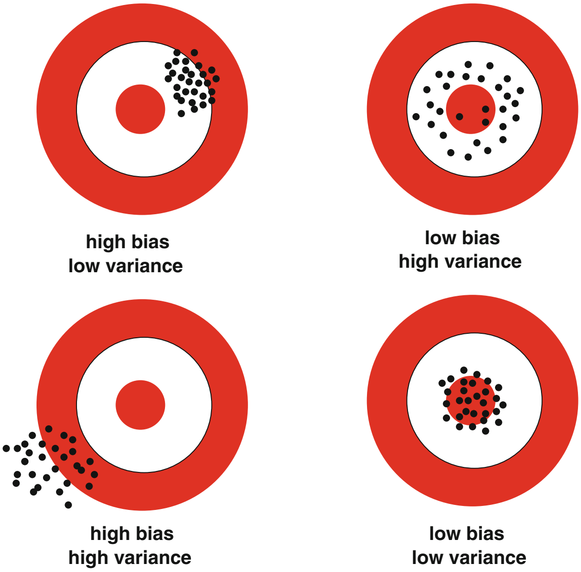

Fig. 8.1 Drawing of *shots* aimed at a target to illustrate the way estimates can miss their "target." They may be systematically biased, or they may have high variability, or both. The best situation, of course, is when there is little systematic bias and little variability.

bias is Bias(T) $= \mathrm{E(T)} - \theta$. When the bias is large, T will not be close to θ *on average*. The second is the variance $V(T)$. If $V(T)$ is large then T will rarely be close to θ. Figure 8.1 illustrates, by analogy with shooting at a bullseye target, the situations in which only the bias is large, only the variance is large, both are large (the worst case) and, finally, both are small (the best case). Part of the appeal of mean squared error is that it combines bias and variance in a beautifully simple way.

Theorem Suppose $E((T-\theta)^2) < \infty$. Then

$$E((T-\theta)^2) = (E(T-\theta))^2 + V(T).$$

That is,

$$MSE(T) = \mathrm{Bias(T)}^2 + \mathrm{Variance(T)}.$$

Proof: Let us write $\mu_T = E(T)$ and $T - \theta = (T - \mu_T) + (\mu_T - \theta)$, and then square both sides to get

$$(T-\theta)^2 = (T-\mu_T)^2 + 2(T-\mu_T)(\mu_T-\theta) + (\mu_T-\theta)^2.$$

Now consider taking the expectation of the cross-product term on the right-hand side. The quantity $\mu_T - \theta$ is a constant (it is not a random variable), while because $E(T) = \mu_T$, we have $E(T - \mu_T) = 0$ and, therefore, $E(2(T-\mu_T)(\mu_T-\theta))=0$. Thus, we have

$$E((T-\theta)^2)=E((T-\mu_T)^2)+(E(\mu_T-\theta))^2$$

and, since $V(T)=E((T-\mu_T)^2)$, we have proven the theorem. □

This decomposition of MSE into squared bias and variance terms is used in various contexts to "tune" estimators in an attempt to decrease MSE. This typically involves some increase in one term, either the squared bias term or the variance term, in order to gain a larger decrease in the other term. Thus, reduction of MSE is often said to involve a *bias variance trade-off*. For an example, see p. 434.

Before we present an illustration of a *MSE* calculation, let us mention a property of the sample mean and sample variance. Assuming they are computed from a random sample $X_1,\ldots,X_n$, we have $E(\bar{X})=\mu_X$ which may be written

$$E(\bar{X})-\mu_X=0.$$

This says that, as an estimator of the theoretical mean, the sample mean has zero bias. When an estimator has zero bias it is called *unbiased*. If an estimator T is unbiased we have $MSE(T)=V(T)$ so that consideration of its performance may be based on a study of its variance.

In addition to the sample mean being unbiased as an estimator of the theoretical mean, it also happens that the *sample variance*, defined by

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

is unbiased as an estimator of the theoretical variance:

$$E(S^2)=\sigma_X^2. \tag{8.4}$$

Details: We wish to evaluate

$$E(S^2)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right).$$

We write $X_i-\bar{X}=(X_i-\mu_X)+(\mu_X-\bar{X})$ and expand the square

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2&=\sum_{i=1}^{n}\left((X_i-\mu_X)+(\mu_X-\bar{X})\right)^2\\&=\sum_{i=1}^{n}(X_i-\mu_X)^2+\sum_{i=1}^{n}2(X_i-\mu_X)(\mu_X-\bar{X})\end{aligned}$$

$$+\sum_{i=1}^{n}(\mu_X - \bar{X})^2.$$

We now rewrite the three terms in the last expression above. Because $E(X_i - \mu_X)^2 = \sigma_X^2$, and the expectation of a sum is the sum of the expectations, the first term has expectation

$$E\left(\sum_{i=1}^{n}(X_i - \mu_X)^2\right) = n\sigma_X^2. \tag{8.5}$$

Next, the second term may be rewritten

$$\begin{aligned}\sum_{i=1}^{n}2(X_i - \mu_X)(\mu_X - \bar{X}) &= 2(\mu_X - \bar{X})\sum_{i=1}^{n}(X_i - \mu_X)\\ &= -2(\bar{X} - \mu_X)\sum_{i=1}^{n}(X_i - \mu_X)\\ &= -2n(\bar{X} - \mu_X)^2,\end{aligned}$$

where the last equality uses $\sum_{i=1}^{n}(X_i - \mu_X) = n(\bar{X} - \mu_X)$, and then, because $E((\bar{X} - \mu_X)^2) = V(\bar{X}) = \sigma_X^2/n$, the expectation of the second term becomes

$$E\left(\sum_{i=1}^{n}2(X_i - \mu_X)(\mu_X - \bar{X})\right) = -2\sigma_X^2. \tag{8.6}$$

Finally, because again, $E((\bar{X} - \mu_X)^2) = \sigma_X^2/n$, the expectation of the third term is

$$E\left(\sum_{i=1}^{n}(\mu_X - \bar{X})^2)\right) = \sigma_X^2 \tag{8.7}$$

and, combining (8.5), (8.6), and (8.7) we get

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = (n-1)\sigma_X^2$$

which gives (8.4). □

We use the unbiasedness of the sample mean and sample variance in the following illustration of the way two estimators may be compared theoretically.

Illustration: Poisson Spike Counts On p. 164 we considered 60 spike counts from a motor cortical neuron and found an approximate 95 % CI for the resulting firing rate using the sample mean. The justification for that approximate CI involved the CLT, and the practical implication was that as long as the sample size is fairly large, and the distribution not too far from normal, the CI would have approximately .95 probability of covering the theoretical mean. In this case, the spike counts do, indeed, appear not too far from normal. Sometimes they are assumed to be Poisson distributed. This is questionable because careful examination of spike trains almost always indicates some departure from the Poisson. On the other hand, the departure is sometimes not large enough to make a practical difference to results. In any case, for the sake of illustrating the *MSE* calculation, let us now *assume* the counts follow a Poisson distribution with mean λ. The sample mean $\bar{X}$ is a reasonable estimator of λ, but one might dream up alternatives. For example, a property of the Poisson distribution is that its variance is also equal to λ; therefore, the sample variance S^2 could also be used to estimate the theoretical variance λ. This may seem odd, and potentially inferior, on intuitive grounds because the whole point is to estimate the mean firing rate, not the variance of the firing rate. On the other hand, once we take the Poisson model seriously the theoretical mean and variance become equal and, from a statistical point of view, it is reasonable to ask whether it is better to estimate one rather than the other from their sample analogues. Our purpose here is to present a simple analysis that demonstrates the inferiority of the sample variance compared with the sample mean as an estimator of the Poisson mean λ. We are going through this exercise so that we can draw an analogy to it later on.

Now, because, as we mentioned immediately before beginning this illustration, $\bar{X}$ and S^2 are unbiased for the theoretical mean and variance they are, in this case, both unbiased as estimators of λ. As a consequence, $MSE(T) = V(T)$ for both $T = \bar{X}$ and $T = S^2$. Analytical calculation of the variance of these estimators (which we omit here) gives

$$V(\bar{X}) = \frac{\lambda}{n}$$
$$V(S^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$$

where n is the number of counts (the number of trials). Therefore, the *MSE* of S^2 is always larger than that of $\bar{X}$ so that S^2 tends to be further from the correct value of λ than $\bar{X}$. For example, if we take $n = 100$ trials and $\lambda = 10$, we find $V(\bar{X}) = .10$ while $V(S^2) = 2.12$. The estimator S^2 has about 21 times the variability as $\bar{X}$, so that estimating λ using S^2 would require about 2,100 trials of data to gain the same accuracy as using $\bar{X}$ with 100 trials. Figure 8.2 shows a pair of histograms of $\bar{X}$ and S^2 values calculated from 1,000 randomly-generated samples of size $n = 100$ when the true Poisson mean was $\lambda = 10$. The distribution represented by the histogram on the right is much wider. □

This illustration nicely shows how one method of estimation can be very much better than another, but it is admittedly somewhat artificial; because the distribution of real spike counts may well depart from Poisson, a careful comparison of $\bar{X}$ versus

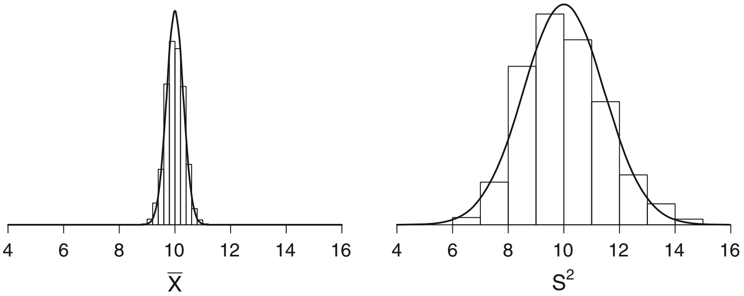

Fig. 8.2 Histograms displaying distributions of $\bar{X}$ and S^2 based on 1,000 randomly-generated samples of size $n = 100$ from a Poisson distribution with mean parameter $\mu = 10$. In these repeated samples both $\bar{X}$ and S^2 have distributions that are approximately normal. Both distributions are centered at 10 (both estimators are unbiased) but the values of S^2 fluctuate much more than do the values of $\bar{X}$.

S^2 should consider their behavior also under alternative assumptions. In this regard, the sample mean remains a reasonably good estimator of the theoretical mean in large samples regardless of the probability distribution of the spike counts. The sample variance, on the other hand, does so only if the theoretical variance is truly equal to the theoretical mean; otherwise, as the sample size increases it will converge to the wrong value. This is likely to be an important consideration. However, even if one were convinced that counts truly followed a Poisson distribution, the analysis above would be compelling. It would be grossly inefficient to use S^2 instead of $\bar{X}$ in estimating λ.

Another thing to notice in Fig. 8.2 is the approximately normal shape of the two histograms. Asymptotic normality of estimators is very common, and we have already relied on it in Section 7.3.5.

8.1.2 Mean squared error may be evaluated by computer simulation of pseudo-data.

In the Poisson spike count illustration on p. 184 we were able to compute the *MSE* exactly. In more complicated situations this is often impossible. Instead we rely on either large-sample arguments, such as those in Section 8.2.2, or numerical simulations. The numerical method uses computer-generated *pseudo-data*, by which we mean numbers or vectors that are generated from known probability distributions in order to mimic the behavior of data. Because the distribution is known, there is a known correct value of θ to which T may be compared.

Suppose we wish to compute $MSE(T)$ in estimating θ under the assumption that a random sample comes from a particular probability distribution having cdf $F(x)$. Assuming we know how to generate random samples from $F(x)$ on the computer, we may use this algorithm:

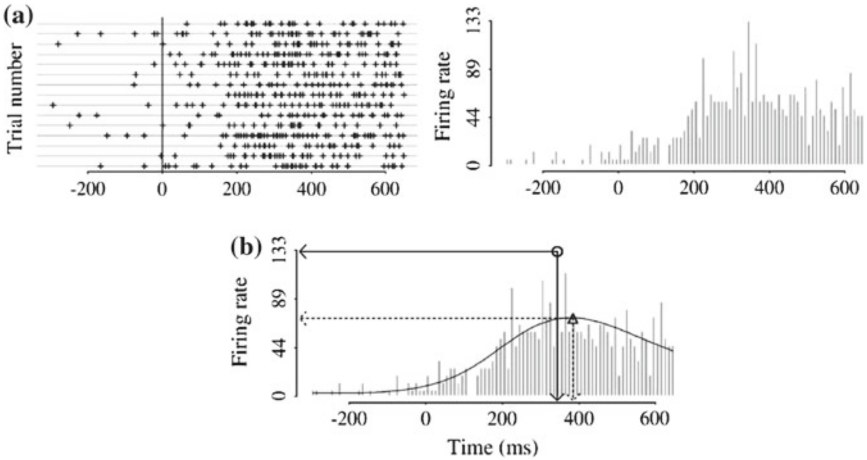

Fig. 8.3 Time of maximal firing rate. **a** displays a raster plot and Peri-Stimulus Time Histogram (PSTH). As explained in Chapter 1, the PSTH represents the firing rate as a function of time. **b** displays the time at which the maximal firing rate occurs, estimated (i) using the PSTH and (ii) using instead a smooth curve. Adapted from Kass et al. (2003).

1. Take G to be a large integer (such as 1,000) and for $g = 1, \ldots, G$ do the following:
   (i) Generate a random sample $X_1^{(g)}, \ldots, X_n^{(g)}$ from $F(x)$.
   (ii) Compute $T^{(g)} = T(X_1^{(1)}, \ldots, X_n^{(g)})$, which is the value of the estimator T based on the gth random sample.
   (iii) Let $Y_g = (T^{(g)} - \theta)^2$.
2. Compute

$$\bar{Y} = \frac{1}{G}\sum_{g=1}^{G} Y_g. \tag{8.8}$$

By the LLN, we have that $\bar{Y}$ converges to the desired $MSE = E((T - \theta)^2)$ in probability. Thus, we take $\bar{Y}$ as our MSE.

This kind of computation is used in the following illustration. It involves the statistical efficiency of smoothing, a topic we take up in Chapter 15. In presenting this now we omit details about the method.

Example 1.1 (continued, see p. 3) In Chapter 1 we discussed a study by Olson et al. (2000), in which neuronal spike trains were recorded from the supplementary eye field (SEF) under two different experimental conditions. As is usually the case in stimulus-response studies, the neuronal response—in this case, the firing rate—varied as a function time. For a particular neuron in one of the conditions, the PSTH in Fig. 8.3 displays the way the firing rate changes across time. The data analytic challenge in the Olson et al. study was to characterize the distinctions between the firing rate functions under the two experimental conditions. One of the distinctions, evident in some of the plots, was that the maximal firing rate occurred somewhat later in one condition than in the other. How should this time of maximal firing rate be computed? One possibility is to use the PSTH, by finding the time bin for which the PSTH is maximized. Panel b of Fig. 8.3 displays the resulting solution: according to the PSTH shown there, the maximal firing rate of about 133 spikes/s occurs at a time marked by the arrow on the left along the time axis. However, this is clearly a noisy estimate. Slight variations in location of time bin, or width, would change this, as would consideration of new data from the same neuron. On the other hand, a second method based on first fitting a smooth curve to the PSTH and then finding its maximum, yields a different answer: the maximum firing rate of about 75 spikes/s (seconds) occurs at a time indicated by the arrow on the right along the time axis. This value is less subject to fluctuations in the data. If we assume that the theoretical firing rate is, in fact, slowly varying in time, then the smooth curve should provide a better estimate. Kass et al. (2003) used MSE to evaluate the extent to which smoothing improves estimation.

Kass et al. (2003) evaluated MSE for the true firing rate function shown in panel a of Fig. 8.4. To do so, they simulated, repeatedly, 16 trials of pseudo-data and then constructed histograms and also fit smooth curves (there are 16 trials in the SEF data shown in Fig. 8.3a). The PSTH and smooth curve from one sample of 16 trials of pseudo-data are shown in panel b of Fig. 8.4. The smoothing method used by Kass et al. (2003) involved regression splines, as discussed in Section 15.2.3. Note that the smooth curve ("estimated rate") is close to the true rate from the simulation, but it misses by a small amount due to the small number of trials we used in the simulation.

To quantify the deviation of both the PSTH and the smooth curve at any one point in time t the *MSE* could be used. That is, we would regard the true firing rate at time t as the value $\theta = \theta_t$ to be estimated, and we would compute $MSE(T) = MSE_t(T)$ when T is based on the PSTH and when T is based on the smooth curve. Here the subscript t is a reminder that we have chosen a particular time point. If $MSE_t(T)$ is evaluated for every time value t the total of all the mean squared errors may be found by integrating across time. This defines what is called the *integrated mean squared error* or *mean integrated squared error* (*MISE*),

$$MISE(T) = \int MSE_t(T)dt$$

where the integration is performed over the time interval of interest. The integral may be evaluated easily simply by calculating the *MSE* along a grid of time values separated by some increment Δt

$$\int MSE_t(T)dt \approx \Delta t \sum_t MSE_t(T).$$

In order to compute the *MSE* at each time value t Kass et al. (2003) used computer simulation: They generated data repeatedly, each time finding both the PSTH and the smooth curve. They simulated 1,000 data sets, each involving 16 randomly-generated

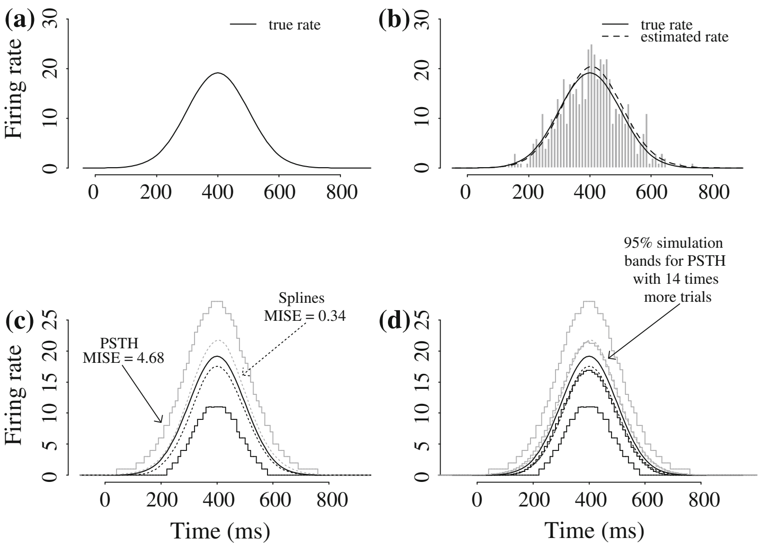

Fig. 8.4 **a** True rate from which 16 trials are simulated; their PSTH is shown in (**b**), with true and estimated firing rates overlaid. **c** shows the true rate and 95 % simulation bands obtained from smoothed and unsmoothed PSTHs. **d** shows the same curves as (**c**), as well as 95 % simulation bands obtained from unsmoothed PSTHs with 16 × 14 trials instead of 16. Adapted from Kass et al. (2003).

spike trains based on the true firing rate curve shown in Fig. 8.4a, and from these 1,000 data sets they computed the *MISE*. They also computed 95 % bands, within which fall 95 % of the estimated curves. Figure 8.4c shows the two pairs of bands, now labeled with the two values of MISE: the spline-based estimate has a MISE of .34 (in spikes/s squared) while the PSTH has a MISE of 4.68, which is 14 times larger. This means that when the PSTH is used to estimate firing rate, 14 times as much data are needed to achieve the same level of accuracy. Similarly, the 95 % bands for the PSTH are much further from the true firing-rate curve than the bands for the spline-based estimate. Figure 8.4d includes a pair of 95 % bands obtained from the PSTH when 224 trials are used rather than 16 (because 224 = 14 × 16). This is another way of showing that the accuracy in estimating the firing rate using spline smoothing based on 16 trials is the same as the accuracy using the PSTH based on 224 trials. Clearly it is very much better to use smoothing when estimating the instantaneous firing rate. □

A detail: One issue that arises in numerical simulation is the accuracy of the computational results, because the value $\bar{Y}$ in (8.8) is itself an estimate of the *MSE*. However, if G is large, the standard error of $\bar{Y}$ will be small. Furthermore, because $\bar{Y}$ is a sample mean, we can apply the method of Section 7.3.4 and use $s/\sqrt{G}$ as its standard error, where $s^2 = \frac{1}{G-1}\sum_{g=1}^{G}(Y_g - \bar{Y})^2$. The standard error lets us determine whether G is adequately large. For instance, if we wish the MSE to be computed with accuracy δ, we can take G big enough to satisfy

$$\frac{s}{\sqrt{G}} < \frac{\delta}{2}.$$

By the result in Section 7.3.4, an approximate 95 % confidence interval for MSE would be $(\theta - \delta, \theta + \delta)$. Thus, we would have 95 % confidence that the desired accuracy was obtained. □

8.1.3 *In estimating a theoretical mean from observations having differing variances a weighted mean should be used, with weights inversely proportional to the variances.*

In the illustration on Poisson spike counts, p. 184, we used the *MSE* criterion to evaluate alternative estimators, based on an analytical expression. In that case both estimators were unbiased and the comparison was based on variance. Another illustration of this type arises when data are considered collectively across many similarly measured objects, such as neurons or subjects, with the observations from the different individuals contributing varying amounts of information; specifically, with the individual observations having different variances. In combining such discrepant observations, it is preferable not to use the sample mean, but instead to weight each observation according to the amount of information it contributes. Here we provide a theoretical analysis of this problem, and give the basic result.

Suppose we have two independent random variables X_i for $i = 1, 2$, with $E(X_1) = E(X_2) = \mu$ but $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, with the two variances possibly being different. After analyzing the two-observation case, we will present analogous results for n observations. Let us assume that σ_1 and σ_2 are known and ask how best to combine X_1 and X_2 linearly in order to estimate μ. We write a general weighted combination as

$$Y_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.9}$$

where $w_1 + w_2 = 1$. It turns out that the optimal special case is

$$\bar{X}_w = w_1 \cdot X_1 + w_2 \cdot X_2 \tag{8.10}$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \tag{8.11}$$

for $i = 1, 2$.

Theorem Suppose X_1 and X_2 are independent random variables with $E(X_1) = E(X_2) = \mu$ and $V(X_1) = \sigma_1^2$ and $V(X_2) = \sigma_2^2$, and let Y_w be defined as in (8.9). Then Y_w is unbiased, so that $MSE(Y_w) = V(Y_w)$, and this quantity is minimized among possible weighting pairs by taking $Y_w = \bar{X}_w$, i.e.,

$$V(\bar{X}_w) \leq V(Y_w)$$

or, equivalently,

$$MSE(\bar{X}_w) \leq MSE(Y_w)$$

with equality holding in both cases only if $Y_w = \bar{X}_w$ defined by (8.10) and (8.11).

Proof of Theorem: First, we have

$$\begin{aligned} E(Y_w) &= w_1 \cdot \mu + w_2 \cdot \mu \\ &= (w_1 + w_2)\mu \\ &= \mu. \end{aligned}$$

Thus, Y_w is unbiased and $MSE(Y_w) = V(Y_w)$. To derive the variance result we start with

$$V(w_1 \cdot X_1 + w_2 \cdot X_2) = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2.$$

Now we use $w_1 + w_2 = 1$ and replace w_2 with $1 - w_1$ to get

$$\begin{aligned} V(w_1 \cdot X_1 + w_2 \cdot X_2) &= w_1^2 \cdot \sigma_1^2 + (1 - w_1)^2 \cdot \sigma_2^2 \\ &= \sigma_1^2 w_1^2 + \sigma_2^2 - 2\sigma_2^2 w_1 + \sigma_2^2 w_1^2 \\ &= (\sigma_1^2 + \sigma_2^2) w_1^2 - 2\sigma_2^2 w_1 + \sigma_2^2. \end{aligned}$$

We now minimize this quantity by differentiating with respect to w_1, and setting the derivative equal to zero. We get

$$0 = 2(\sigma_1^2 + \sigma_2^2) w_1 - 2\sigma_2^2$$

and therefore

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Dividing the numerator and denominator of this fraction by $\sigma_1^2 \sigma_2^2$ gives

$$w_1 = \frac{\frac{\sigma_2^2}{\sigma_1^2 \sigma_2^2}}{\frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 \sigma_2^2}} = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

which is the desired result. □

As an illustration, suppose we had 100 independent observations $U_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, 100$, and grouped them unequally defining, say, $X_1 = \frac{1}{10}\sum_{i=1}^{10} U_i$ and $X_2 = \frac{1}{90}\sum_{i=11}^{100} U_i$. It would seem strange to use $\frac{1}{2}(X_1 + X_2)$ in this situation and the intuitive thing to do would be to use the weighted mean: here the weights are $w_1 = 10/100$ and $w_2 = 90/100$ (because $\sigma_1^2 = \sigma^2/10$ and $\sigma_2^2 = \sigma^2/90$) so we get $\bar{X}_w = \bar{U}$.

One way to interpret this is to say that using $\bar{X}$ instead of $\bar{X}_w$ is like throwing away a fraction of the data. For example, suppose X_1 and X_2 both represent means of counts from n trials. If σ_1 is half the size of σ_2 then, from the formula above, the ratio of variances is 1.56. This means that to achieve the same accuracy in the estimator, n would have to be 56 % larger if we used the sample mean instead of the weighted mean. When σ_1 is one-third the size of σ_2 we would have to increase n by a factor of 2.78 (instead of 50 trials, say, we would need 139). In these cases we might say that the weighted mean is, respectively, 1.56 and 2.78 times more efficient than the ordinary sample mean.

Example 8.1 Optimal integration of sensory information Ernst and Banks (2002) considered whether humans might combine two kinds of sensory input optimally, according to (8.10) and (8.11). Subjects were presented with raised bars either visually or by touch (known as haptic input) and had to judge the height of each bar in comparison with a "standard" bar. The experimental apparatus was set up to allow visual or haptic noise to be added to the height of each bar. Subjects were also presented with both visual and haptic input simultaneously. The authors reported evidence that when presented with the simultaneous visual and haptic input, subjects judged heights by combining the two sensory modalities consistently with (8.10) and (8.11). In other words, this was evidence that humans can integrate distinct sensory inputs optimally. □

Here is the result for combining n observations. We have also included here the formula for the standard error of the weighted mean.

Theorem Suppose $X_1, \ldots, X_n$ are independent random variables with $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$ and $V(X_i) = \sigma_i^2$ for $i = 1, \ldots, n$. Let

$$\bar{X}_w = \sum_{i=1}^{n} w_i \cdot X_i \tag{8.12}$$

where, in (8.12),

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

and for any set of weights $w_1, \ldots, w_n$ for which $\sum_{i=1}^{n} w_i = 1$ define

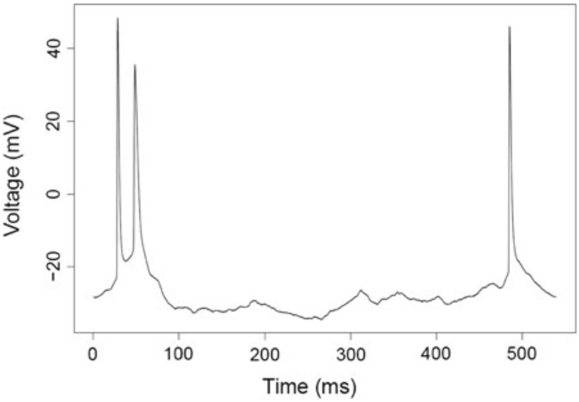

Fig. 8.5 When an action potential follows closely a previous action potential (with small ISI), the second action potential is broader than the first. When a long ISI intervenes, however, the second action potential is very similar to the first.

$$Y_w = \sum_{i=1}^{n} w_i \cdot X_i.$$

Then we have

$$V(\bar{X}_w) \leq V(Y_w)$$

with equality holding if and only if $Y_w = \bar{X}_w$. Furthermore we have

$$SE(\bar{X}_w) = \sqrt{V(\bar{X}_w)} \tag{8.13}$$

where

$$V(\bar{X}_w) = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1}.$$

Proof: The proof is analogous to that for the case $n = 2$. □

Example 8.2 Action potential width and the preceding inter-spike interval As part of a study on the effects of seizure-induced neural activity (Shruti et al. 2008) spike trains were recorded from barrel cortex neurons in slice preparation. One of the interesting findings[2] involved the relationship between the width of each action potential (spike) and its preceding ISI. As is well known, when a spike follows closely on a preceding spike, so that the ISI is relatively short, then the second spike will tend to be wider than the first. If, however, the ISI is sufficiently long, there will not be any effect of the first spike on the second, and the spike widths should be roughly

[2] The results here were obtained by Judy Xi.

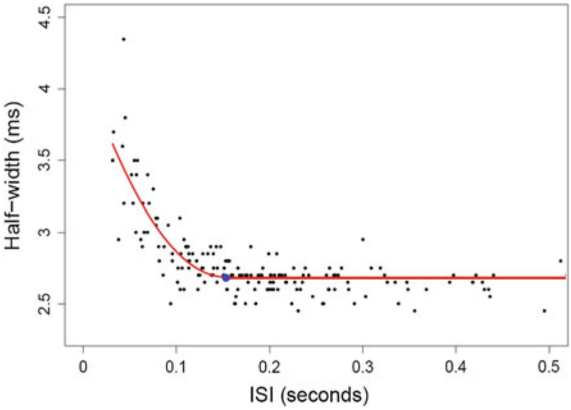

Fig. 8.6 Action potential width varies as function of previous ISI. The data are from many action potentials recorded for a single neuron. A *fitted curve* with a change point is also shown, the change point being indicated as a *large blue dot*.

equal. See (Fig. 8.5). How long is "sufficiently long?" This turns out to be dependent on previous neuronal activity.

Let Y be the spike width and x the preceding ISI length, and let us assume there is an ISI length τ such that, on average, Y is constant for all $x > \tau$. Among neurons taken from animals that had seizures, τ tended to be smaller than its value among control animals. Figure 8.6 displays some of the data, together with a fitted curve. The statistical model used for this curve assumes that, on average, Y decreases with x for $x < \tau$ but remains constant for $x \geq \tau$. In statistical jargon, τ is called a *change point*, because the relationship between Y and x changes at $x = \tau$. The relationship between y and x was assumed to be quadratic for $x < \tau$ (see Section 12.5.4) and constant for $x \geq \tau$. The model was fit using nonlinear least squares. Additional details are given on p. 408 in Section 14.2.1. The parametric bootstrap (Section 9.2.2) was then applied to obtain the $SE(\hat{\tau})$. The method was repeated for neurons from seizure and control animals to see whether there were systematic differences across the two treatment conditions. Figure 8.7 shows results for both groups. Note the very different standard errors across neurons. This suggests that in comparing the two groups it is advisable to use weighted means, as in Eq. (8.12), together with standard errors given by Eq. (8.13). The results were that the control group had weighted mean change point of 190 (±32) ms and the seizure group reset earlier, with weighted mean change point 108 (±.012) ms. □

Example 8.3 Neural response to selective perturbation of a brain-machine interface In order to study learning-related changes in a network of neurons, Jarosiewicz et al. (2008) introduced a paradigm in which the output of a cortical network can be perturbed directly and the neural basis of the compensatory changes studied in detail. Using a brain-computer interface (BCI), dozens of simultaneously recorded neurons in the motor cortex of awake, behaving monkeys were used to control the movement of a cursor in a three-dimensional virtual-reality environment.

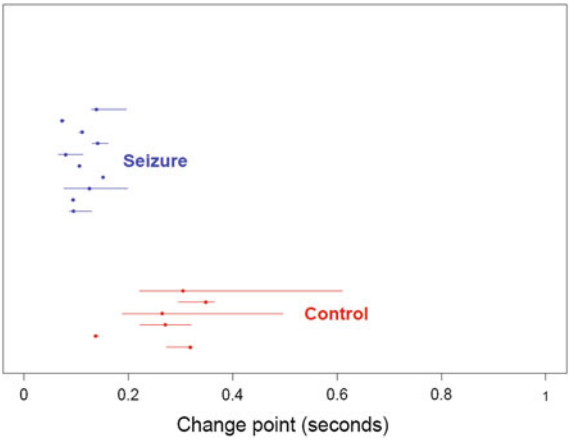

Fig. 8.7 Change points and SEs for neurons of both seizure and control groups. The results for the seizure group appear above those for the control group. The seizure group has change points that occur earlier and they tend to be less variable.

This device creates a precise, well-defined mapping between the firing of the recorded neurons and an expressed behavior (cursor movement). In a series of experiments, they forced the animal to relearn the association between neural firing and cursor movement in a subset of neurons and assessed how the network changes to compensate. Their main finding was that changes in neural activity reflect not only an alteration of behavioral strategy but also the relative contributions of individual neurons to the population error signal. As part of their study the authors compared firing rate modulation among neurons whose BCI signals had been artificially perturbed with that among neurons whose BCI signals remained as determined from their control responses. Because the uncertainties varied substantially across neurons, these comparisons among groups of neurons were carried out using weighted means. □

8.1.4 Decision theory often uses mean squared error to represent risk.

At the end of Section 4.3.4, on p. 102, we mentioned that optimal classification may be considered a problem in decision theory where, in general, the expected loss or *risk* is minimized. In the context of estimation we may consider a decision rule d to be a mapping from each possible vector of observations to a parameter value: we may write $d(X_1, \ldots, X_n) = T$. If we use *squared-error loss* defined by

$$L(d(x_1, \ldots, x_n), \theta) = (d(x_1, \ldots, x_n) - \theta)^2 ,$$

then *MSE* is the risk function

$$MSE(T) = E\left(L(d(X_1, \ldots, X_n), \theta)\right).$$

This terminology, viewing MSE as "risk under squared-error loss," is quite common.

8.2 Estimation in Large Samples

8.2.1 In large samples, an estimator should be very likely to be close to its estimand.

In the introduction to this chapter we offered the reminder that the sample mean satisfies

$$\bar{X} \xrightarrow{P} \theta$$

which is the law of large numbers. Suppose T_n is an estimator of θ. If, as $n \to \infty$, we have

$$T_n \xrightarrow{P} \theta \tag{8.14}$$

then T_n is said to be a *consistent* estimator of θ. This means that for every positive ϵ, as $n \to \infty$ we have

$$P(|T_n - \theta| > \epsilon) \to 0.$$

Note that, by (8.3), if $MSE(T_n) \to 0$ then T_n is consistent. Also, if T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ then T_n is consistent.

> *Details:* Multiplying the left-hand side of (8.1) by σ_{T_n} and applying Slutsky's theorem we have $T_n - \theta \xrightarrow{P} 0$, which is equivalent to $T_n \xrightarrow{P} \theta$. □

In words, to say that an estimator is consistent is to say that, for sufficiently large samples, it will be very likely to be close to the quantity it is estimating. This is clearly a desirable property. When T_n satisfies (8.1) and $\sigma_{T_n} \to 0$ we will call T_n *consistent and asymptotically normal.*

8.2.2 In large samples, the precision with which a parameter may be estimated is bounded by Fisher information.

Let us consider all estimators of θ that are consistent and asymptotically normal in the sense of Section 8.2.1. For such an estimator $T = T_n$ we may say that its distribution is approximately normal, and we write

$$T \,\dot{\sim}\, N(\theta, \sigma_T^2), \tag{8.15}$$

where the symbol $\dot{\sim}$ means "is approximately distributed as." The expression (8.15) is a convenient way to think of the more explicit Eq. (8.1). From (8.15), σ_T may be considered[3] the standard error of T, and an approximate 95 % CI for θ based on T would be $(T - 2\sigma_T, T + 2\sigma_T)$.

Now, suppose we had two such estimators T^A and T^B that both satisfy (8.15). We would say that T^A is asymptotically more accurate than T^B if $\sigma_{T^A} < \sigma_{T^B}$. An extreme case of this was displayed in Fig. 8.2, where $T^A = \bar{X}$ and $T^B = S^2$, with both histograms being approximately normal in shape and σ_{T^B} being more than four times larger than σ_{T^A}. In general, we would prefer to use an estimator with a small σ_T because it would tend to be closer to θ than an estimator with a larger value of σ_T. In addition, a small σ_T would produce comparatively narrow CIs, indicating improved knowledge about θ. Ideally, we would like to find an estimator T for which σ_T would be as small as possible. Fisher (1922) discovered that this is a soluble problem: there is a minimum value of σ_T and, furthermore, this minimum value is achieved by the method of maximum likelihood.

To understand how this works, we may use some rough heuristics[4] based on the normality in (8.15) to get an expression for σ_T. Let us first note an important fact about normal distributions. Suppose $X \sim N(\mu, \sigma^2)$ with σ known, and consider the loglikelihood function

$$\ell(\mu) = \log f_X(x|\mu).$$

We have

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

so that

$$\ell(\mu) = -\frac{(x-\mu)^2}{2\sigma^2}, \tag{8.16}$$

and when we differentiate twice we get

$$\ell'(\mu) = \frac{x-\mu}{\sigma^2}$$

and

$$\ell''(\mu) = -\frac{1}{\sigma^2}$$

[3] In practice, σ_T may depend on the value of θ, which is unknown, so that a data-based version $\hat{\sigma}_T$ would have to be substituted in forming a confidence interval.

[4] For a rigorous treatment along the lines of the argument here see Kass and Vos (1997, Chapter 2). See also Bickel and Doksum (2001, Chapter 5).

which gives

$$\sigma^2 = \frac{1}{-\ell''(\mu)}. \tag{8.17}$$

That is, the standard deviation of a normal pdf is determined by the second derivative of the loglikelihood function $\ell(\mu)$.

The result (8.17) suggests that when a pdf of an estimator is approximately normal, its standard error may be found in terms of the second derivative of the corresponding loglikelihood function. We now apply this idea to the approximate normal pdf based on (8.15). We write the pdf of the estimator T as $f_T(t|\theta)$ and define its loglikelihood function to be

$$\ell_T(\theta) = \log f_T(t|\theta). \tag{8.18}$$

Using the approximate normality in (8.15) and applying (8.17) we get

$$\sigma_T^2 = \frac{1}{-\ell_T''(\theta)}. \tag{8.19}$$

Equation (8.19) implies that minimizing σ_T is the same as maximizing $-\ell_T''(\theta)$. However, there is an important distinction between (8.19) and (8.17). In (8.17), $\ell''(\mu)$ is a constant whereas, because T is a random variable, $-\ell_T''(\theta)$ is also random (it does not reduce to a constant except when T is exactly normally distributed, so that its loglikelihood becomes exactly quadratic). Thus, regardless of how we were to choose the estimator T, we could not guarantee that $-\ell_T''(\theta)$ would be large because there would be some probability that it might be small. We therefore work with its average value, i.e., its expectation, for which we use the following notation:

$$I^T(\theta) = E\left(-\frac{d^2}{d\theta^2}\log f_T(t|\theta)\right). \tag{8.20}$$

If we replace $-\ell_T''(\theta)$ in (8.19) by its expectation, using (8.20), we get

$$\sigma_T^2 = \frac{1}{I^T(\theta)}. \tag{8.21}$$

The quantity $I^T(\theta)$ is called the *information* about θ contained in the estimator T. Thus, an optimal estimator would be one that makes the information as large as possible.

How large can the information $I^T(\theta)$ be? Fisher's insight was that the information in the estimator can not exceed the analogous quantity derived from the whole sample, which is now known as the *Fisher information*. For a parametric family of distributions having pdf $f(x|\theta)$ the Fisher information is given by

$$I_F(\theta) = E\left(-\frac{d^2}{d\theta^2}\log f(X|\theta)\right).$$

To be clear, for a continuous random variable on (A,B) this expectation is

$$I_F(\theta) = -\int_A^B \left(\frac{d^2}{d\theta^2}\log f(x|\theta)\right) f(x|\theta)dx.$$

For a random sample drawn from this distribution the Fisher information is given by[5]

$$\begin{aligned} I(\theta) &= E\left(-\frac{d^2}{d\theta^2}\log\prod_{i=1}^n f(X_i|\theta)\right) \\ &= E\left(-\frac{d^2}{d\theta^2}\sum_{i=1}^n \log f(X_i|\theta)\right) \\ &= \sum_{i=1}^n E\left(-\frac{d^2}{d\theta^2}\log f(X_i|\theta)\right) \end{aligned}$$

and, because the sample involves identically distributed random variables, all of the expected values in this final expression are the same, and equal to $I_F(\theta)$. Therefore, we have

$$I(\theta) = nI_F(\theta).$$

Result Under certain general conditions, the information in an estimator T satisfies

$$I^T(\theta) \le I(\theta). \tag{8.22}$$

Therefore, the large-sample variance σ_T^2 of a consistent and asymptotically normal estimator satisfies

$$\sigma_T^2 \ge \frac{1}{I(\theta)}. \tag{8.23}$$

In words, (8.22) says that the information in an estimator can not exceed the information in the whole sample. In Section 8.3.1 we add that the MLE attains this upper bound asymptotically, as $n \to \infty$ and, therefore, has the smallest possible asymptotic variance.

> *A detail:* It is possible for an estimator T to achieve the information bound exactly, in finite samples, i.e.,
>
> $$I^T(\theta) = I(\theta)$$

[5] Because the expectation is used in defining $I(\theta)$, it is often called the *expected information* to distinguish it from the *observed information* which we discuss in Section 8.3.2.

for all n. When this happens the estimator contains all of the information about θ that is available in the data, and it is called a *sufficient statistic*. For instance, if we have a sample from a $N(\mu, \sigma^2)$ distribution with σ known, then the sample mean $\bar{X}$ is sufficient for estimating μ. Sufficiency may be characterized in many ways. If T is a sufficient statistic, then the likelihood function based on T is the same as the likelihood function based on the entire sample. For example, it is not hard to verify that the likelihood function based on a sample $(x_1, \ldots, x_n)$ from a $N(\mu, \sigma^2)$ distribution with σ known is the same as the likelihood function based on $\bar{X}$. This property is sometimes known as *Bayesian sufficiency* (see Bickel and Doksum 2001). In addition, if θ is given a prior distribution as in Section 7.3.9, then T is sufficient when the mutual information between θ and T is equal to the mutual information between θ and the whole sample (see Cover and Thomas 1991). Parametrized families of distributions for which it is possible to find a sufficient statistic with the same dimension as the parameter vector are called *exponential families*. See Section 14.1.6. □

A related result is the following. If we let $\psi(\theta) = E(T)$, where the expectation is based on a random sample from the distribution with pdf $f(x|\theta)$, it may be shown[6] that

$$V(T) \geq \frac{\psi'(\theta)^2}{I(\theta)}.$$

Therefore, if T is an unbiased estimator of θ based on a random sample from the distribution with pdf $f(x|\theta)$ we have $\psi'(\theta) = 1$ and

$$V(T) \geq \frac{1}{I(\theta)}. \tag{8.24}$$

This is usually called the *Cramér-Rao lower bound*. Although Eq. (8.24) is of less practical importance than the asymptotic result (8.23), authors often speak of the bound in (8.23) as a Cramér-Rao lower bound.

Fisher information also arises in theoretical neuroscience, particularly in discussion of neural decoding and optimal properties of tuning curves (see Dayan and Abbott 2001).

8.2.3 Estimators that minimize large-sample variance are called efficient.

A consistent and asymptotically normal estimator T satisfies (8.1) and it also satisfies (8.22). In (8.1) we suppressed the dependence of T and σ_T on n. The information $I^T(\theta)$ also depends on n, as does $I(\theta)$. We now consider what happens as $n \to \infty$.

[6] See Bickel and Doksum (2001, Chapter 3).

Suppose we have a consistent and asymptotically normal estimator T which, by definition, satisfies (8.1). If we find a sequence of numbers $c_1, c_2, \dots, c_n, \dots$ such that

$$\frac{\sigma_{T_n}}{c_n} \to 1 \tag{8.25}$$

then we have

$$\frac{T_n - \theta}{c_n} \xrightarrow{D} N(0, 1). \tag{8.26}$$

Details: We write
$$\frac{T_n - \theta}{c_n} = \frac{T_n - \theta}{\sigma_{T_n}} \frac{\sigma_{T_n}}{c_n}$$

and apply Slutsky's Theorem (p. 163) using (8.25). □

Equation (8.26) says that c_n can also serve as the large-sample standard error of T. If we have two consistent and asymptotically normal estimators T^A and T^B what matters is the limiting ratio η defined by

$$\frac{\sigma_{T^A}}{\sigma_{T^B}} \to \eta$$

as $n \to \infty$. If $\eta < 1$ then, in large samples, T^A is more accurate than T^B, while if $\eta = 1$ the two estimators are equally accurate. This, together with (8.22), leads us to conclude that the large-sample value of σ_T is minimized if

$$\frac{I^T(\theta)}{I(\theta)} \to 1 \tag{8.27}$$

$n \to \infty$. In this case we also have

$$\sqrt{I(\theta)}(T - \theta) \xrightarrow{D} N(0, 1). \tag{8.28}$$

When an estimator attains (8.27), and therefore (8.28), it is said to be *efficient.*

Details: In general, if $a_1, \dots, a_n, \dots$ and $b_1, \dots, b_n, \dots$ are positive sequences that satisfy
$$\frac{a_n}{b_n} \to 1$$

then
$$\sqrt{\frac{a_n}{b_n}} \to 1.$$

Applying this to (8.27) we get

$$\sqrt{\frac{I^T(\theta)}{I(\theta)}} \to 1. \tag{8.29}$$

as $n \to \infty$. Let us rewrite $1/\sigma_T$ as

$$\frac{1}{\sigma_T} = \sqrt{I^T(\theta)} = \sqrt{\frac{I^T(\theta)}{I(\theta)}}\sqrt{I(\theta)}. \tag{8.30}$$

Putting (8.30) in (8.1) we get

$$\sqrt{\frac{I^T(\theta)}{I(\theta)}}\sqrt{I(\theta)}(T_n - \theta) \overset{D}{\to} N(0, 1). \tag{8.31}$$

Therefore, by Slutsky's Theorem (p. 163), if (8.27) holds for some estimator T then (8.28) also holds. □

Fisher (1922) described efficient estimators by saying they contain the maximal amount of information supplied by the data about the value of a parameter, and there are rigorous mathematical results that justify Fisher's use of these words. Roughly speaking, the information in the data pertaining to the parameter value may be used well (or poorly) to make an estimator more (or less) accurate; in using as much information about the parameter as is possible, an efficient estimator uses the data most efficiently and reduces to a minimum the uncertainty attached to it. Other definitions of efficiency are sometimes used in statistical theory, but the one based on Fisher information remains most immediately relevant to data analysis, and supports Fisher's observations about maximum likelihood.

8.3 Properties of ML Estimators

8.3.1 In large samples, ML estimation is optimal.

We now state Fisher's main discovery about ML estimation.

Result Under certain general conditions, if T is the MLE then (8.27) and (8.28) hold. That is, ML estimators are consistent, asymptotically normal, and efficient:

$$\sqrt{I(\theta)}(\hat{\theta} - \theta) \overset{D}{\to} N(0, 1). \tag{8.32}$$

In other words, when we consider what happens as $n \to \infty$, among all those "nice" estimators that are consistent and asymptotically normal, ML estimators are the best in the sense of having the smallest possible limiting standard deviation.

Results may also be derived[7] in terms of *MSE*. Under certain conditions, an estimator T_n must satisfy

$$I(\theta) \cdot MSE(T_n) \to c$$

where $c \geq 1$ and for the MLE, where $T = \hat{\theta}$, we have

$$I(\theta) \cdot MSE(\hat{\theta}) \to 1.$$

This is a different way of saying that, for large samples, ML estimation is as accurate as possible.

8.3.2 The standard error of the MLE is obtained from the second derivative of the loglikelihood function.

Although we have emphasized the theoretical importance of Eq. (8.28), to be useful for data analysis it must be modified: the quantity $I(\theta)$ depends on the unknown parameter θ, so we must replace $I(\theta)$ with an estimate of it. In other words, when we apply maximum likelihood and want to use (8.32) we must modify it to obtain a confidence interval. One possible such modification is fairly obvious, based on the way we modified initial asymptotic normality results in our discussion of confidence intervals in Section 7.3: we replace θ with the MLE $\hat{\theta}$. Under certain conditions we have

$$\sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \qquad (8.33)$$

> *Details:* Because $\hat{\theta} \to \theta$ in probability (i.e., the MLE is consistent), it may be shown that we also have $\sqrt{I(\hat{\theta})/I(\theta)} \to 1$ in probability, so we can again apply Slutsky's Theorem together with (8.28) to get (8.33). □

It turns out that there is a more convenient version of the result. The difficulty with (8.33) is that in some problems it is hard to compute $I(\theta)$ analytically because of the required expectation. Instead, as a general rule, we replace $I(\theta)$ with the *observed information* given by

$$I_{OBS}(\hat{\theta}) = -\ell''(\hat{\theta}). \qquad (8.34)$$

[7] See the discussion and references in Kass and Vos (1997).

In other words, instead of the expected information evaluated at $\hat{\theta}$ in (8.33), we use the negative second derivative of the loglikelihood, evaluated at $\hat{\theta}$, without[8] any expectation. Again, under certain conditions, we have

$$\sqrt{I_{OBS}(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{D} N(0, 1). \tag{8.35}$$

Details: Note that

$$-\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f(x_i|\theta)$$

and that the expectation of the right-hand side is $I_F(\theta)$. From the LLN we therefore have

$$-\frac{1}{n}\ell''(\theta) \xrightarrow{P} I_F(\theta),$$

and it may also be shown that

$$\sqrt{\frac{I_{OBS}(\hat{\theta})}{I(\hat{\theta})}} \xrightarrow{P} 1,$$

which, again by Slutsky's Theorem, gives (8.35). □

Equation (8.35) provides large-sample standard errors and confidence intervals based on ML estimation, given in the following result.

> **Result** For large samples, under certain general conditions, the MLE $\hat{\theta}$ satisfies (8.35), so that its standard error is given by
>
> $$SE = \frac{1}{\sqrt{-\ell''(\hat{\theta})}} \tag{8.36}$$
>
> and an approximate 95 % CI for θ is given by $(\hat{\theta} - 2SE, \hat{\theta} + 2SE)$.

Additional insight about the observed information can be gained by returning to the derivation of (8.17) and applying it, instead, to the likelihood function based on a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution with σ known, as in Section 7.3.2. There, we found the loglikelihood function to be

[8] For the special class of models known as exponential families, which are used with the generalized linear models discussed in Chapter 14, we have $I(\hat{\theta}) = I_{OBS}(\hat{\theta})$ (see, e.g., Kass and Vos 1997) but this is not true in general.

$$\ell(\theta) = -\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2}$$

which simplified to Eq. (7.2),

$$\ell(\theta) = -\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta).$$

Differentiating this twice we get

$$\ell''(\theta) = -\frac{n}{\sigma^2},$$

so that

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{-\ell''(\theta)}}. \qquad (8.37)$$

In other words, $1/\sqrt{-\ell''(\theta)}$ gives the standard error of the mean in that case.

Quite generally, for large samples, the likelihood function has an approximately normal form and there is a strong analogy with this paradigm case. Specifically, a quadratic approximation to the loglikelihood function (using a second-order Taylor expansion) produces a normal likelihood (because if $Q(\theta)$ is quadratic then $\exp(Q(\theta))$ is proportional to a normal likelihood function) and in this normal likelihood the value of the standard deviation is $1/\sqrt{-\ell''(\hat{\theta})}$. This heuristic helps explain (8.36).

Details: The quadratic approximation to $\ell(\theta)$ at $\hat{\theta}$ is

$$Q(\theta) = \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}\ell''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Using $\ell'(\hat{\theta}) = 0$ and setting $c = \exp(\ell(\hat{\theta}))$ we have

$$\exp(Q(\theta)) = c\exp\left(-\frac{1}{2}(-\ell(\hat{\theta}))(\hat{\theta} - \theta)^2\right). \qquad (8.38)$$

The function on the right-hand side of (8.38) has the form of a likelihood function based on $X \sim N(\theta, \sigma^2)$ where $\hat{\theta}$ plays the role of x and $\sigma = 1/\sqrt{-\ell''(\hat{\theta})}$. □

We now consider two simple illustrations.

Illustration: Exponential distribution Suppose $X_i \sim Exp(\lambda)$ for $i = 1, \ldots, n$, independently. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ &= \lambda^n e^{-\lambda \sum x_i} \\ &= \lambda^n e^{-\lambda n\bar{x}} \end{aligned}$$

and the loglikelihood function is

$$\ell(\lambda) = n \log \lambda - n\lambda\bar{x}.$$

Differentiating this and setting equal to zero gives

$$0 = n(\frac{1}{\lambda} - \bar{x})$$

and solving this for λ yields the MLE

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Continuing, we compute the observed information:

$$\begin{aligned} -\ell''(\hat{\lambda}) &= \frac{n}{\hat{\lambda}^2} \\ &= n\bar{x}^2 \end{aligned}$$

which gives us the large-sample standard error

$$SE(\hat{\lambda}) = \frac{1}{\bar{x}\sqrt{n}}.$$ □

Illustration: Binomial For a $B(n, p)$ random variable it is straightforward to obtain the observed information

$$-\ell''(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}.$$

This gives

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

which is the same as the *SE* found in Section 7.3.5. Therefore, the approximate 95 % CI in (7.22) is an instance of that provided by ML estimation with SE given by (8.36).

□

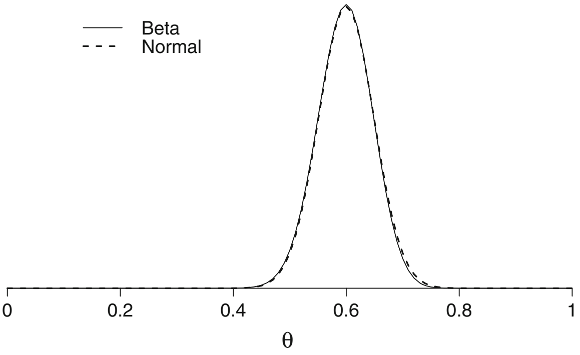

Fig. 8.8 Normal approximation $N(.6, (.049)^2)$ to beta posterior $Beta(61, 41)$.

8.3.3 *In large samples, ML estimation is approximately Bayesian.*

In Section 7.3.9 we said that Bayes' theorem may be used to provide a form of estimation based on the posterior distribution according to (7.28), i.e.,

$$f_{\theta|x}(\theta|x) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}.$$

One of the most important results in theoretical statistics is the approximate large-sample equivalence of inference based on ML and inference using Bayes' theorem.

Result For large samples, under certain general conditions, the posterior distribution of θ is approximately normal with mean given by the MLE $\hat{\theta}$ and standard deviation given by the standard error formula (8.36).

We elaborate in Section 16.1.5 and content ourselves here with a simple illustration.

Illustration: Binomial distribution Suppose $Y \sim B(n, \theta)$ with $n = 100$ and we observe $y = 60$. As we said in Section 7.3.9, if we take the prior distribution on θ to be $U(0, 1)$, which is also the $Beta(1, 1)$ distribution, we obtain a $Beta(61, 41)$ posterior. The observed proportion is the MLE $\hat{\theta} = x/n = .6$. The usual standard error then becomes $SE = \sqrt{\hat{\theta}(1 - \hat{\theta})/n} = .049$. As shown in Fig. 8.8 the normal distribution with mean $\hat{\theta}$ and standard deviation $\sqrt{\hat{\theta}(1 - \hat{\theta})/n}$ is a remarkably good approximation to the posterior. □

For the data from subject P.S. in Example 1.4, which involves a relatively small sample, we already noted (see p. 174) that the approximate 95 % confidence interval (.64, 1.0) found using (8.36) (which is the same as (7.22), see p. 206) differed by only a modest amount from the exact 95 % posterior probability interval we obtained, which was (.59, .94).

8.3.4 MLEs transform along with parameters.

It sometimes happens that we wish to consider an alternative parameterization of a pdf, say γ rather than θ, and then want find the MLE of γ. If $\gamma = g(\theta)$ for a transformation function g having nonzero derivative, then the MLE of the transformation equals the transformation of the MLE:

$$\hat{\gamma} = g(\hat{\theta}).$$

This is often called *invariance* or *equivariance.* The derivation of invariance of ML is perhaps most easily followed in a concrete example. The argument given next for the exponential distribution could be applied to any parametric family.

Illustration: Exponential distribution (continued from p. 205) Suppose we parameterize the *Exp*(λ) distribution in terms of the mean $\mu = 1/\lambda$ so that its pdf becomes

$$f(x) = \frac{1}{\mu}e^{-x/\mu}.$$

Previously (see p. 205) we found that the MLE of λ based on a sample from *Exp*(λ) is $\hat{\lambda} = 1/\bar{x}$. The invariance property of ML says that

$$\hat{\mu} = 1/\hat{\lambda} = \bar{x}.$$

To see why this works for the exponential distribution, let us use a subscript on the likelihood function to indicate its argument, $L_\lambda(\lambda)$ vs. $L_\mu(\mu)$. We find $L_\mu(\mu)$ by starting with

$$L_\lambda(\lambda) = \lambda^n e^{-\lambda n\bar{x}}$$

and writing

$$L_\mu(\mu) = L_\lambda(\frac{1}{\mu}) = \frac{1}{\mu^n}e^{-n\bar{x}/\mu}.$$

Thus, when we maximize $L_\mu(\mu)$ over μ, we are maximizing $L_\lambda(1/\mu)$ over μ which is the same thing as maximizing $L_\lambda(\lambda)$ over λ. We therefore must have $\hat{\mu} = 1/\hat{\lambda}$. More generally, the same argument shows that when $\gamma = g(\theta)$ we must have $\hat{\gamma} = g(\hat{\theta})$. □

Invariance is by no means a trivial property: some methods of estimation are *not* invariant to transformations of the parameter.

8.3.5 Under normality, ML produces the weighted mean.

We now return to choosing the weights for a weighted mean, discussed in Section 8.1.3. Previously (p. 190) we found the weights that minimized *MSE*. A different way to solve the problem is to introduce a statistical model, and then apply the method of maximum likelihood. Let us do this.

To apply ML, we assume that X_1 and X_2 are both normally distributed. The loglikelihood is

$$\ell(\mu) = -\frac{(x_1-\mu)^2}{2\sigma_1^2} - \frac{(x_2-\mu)^2}{2\sigma_2^2}$$

and setting its derivative equal to zero gives

$$\begin{aligned} 0 &= -\frac{x_1-\mu}{\sigma_1^2} - \frac{x_2-\mu}{\sigma_2^2} \\ &= -\frac{x_1}{\sigma_1^2} - \frac{x_2}{\sigma_2^2} + \mu\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right). \end{aligned}$$

Therefore, dividing through by $\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$, the MLE is

$$\hat{\mu} = w_1 \cdot X_1 + w_2 \cdot X_2,$$

where

$$w_i = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

for $i = 1, 2$. This is Eq. (8.10).

8.4 Multiparameter Maximum Likelihood

The method of ML estimation was defined for the case of a scalar parameter θ in Section 7.2.2, together with Eqs. (8.35) and (8.36). More generally, when θ is a vector, the definitions of the likelihood function, loglikelihood function, and MLE remain unchanged. The observed information instead becomes a matrix, and the approximate normal distribution mentioned in conjunction with Eq. (8.36) instead becomes an approximate *multivariate* normal distribution.

8.4.1 The MLE solves a set of partial differential equations.

In Section 7.2.2 we computed the MLE by solving the differential equation

$$0 = \ell'(\theta) \tag{8.39}$$

when θ was a scalar. To obtain the MLE of an m-dimensional vector parameter, we must solve precisely the same equation, except that now the derivative in Eq. (8.39) is the vector

$$\ell'(\theta) = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_m} \end{pmatrix}.$$

This means that Eq. (8.39) is really a set of m equations, often called *the likelihood equations*, which need to be solved simultaneously.

Illustration: Normal MLE Let us return to finding the MLE for a sample $x_1, \ldots, x_n$ from a $N(\mu, \sigma^2)$ distribution. Previously we assumed σ was known, but now we consider the joint estimation of μ and σ. The loglikelihood function now must include a term previously omitted that involves σ. The joint pdf is

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

and the loglikelihood function is

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

The partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

Setting the first equation equal to 0 we obtain

$$\hat{\mu} = \bar{x}.$$

Setting the second equation equal to 0 and substituting $\hat{\mu} = \bar{x}$ gives

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

The MLE is thus slightly different than the usual sample standard deviation s, which is defined with the denominator $n - 1$ so that the sample variance becomes unbiased as an estimator of σ^2, as in (8.4). We have

$$\hat{\sigma} = \sqrt{\frac{n-1}{n}} \cdot s.$$

Clearly the distinction is unimportant for substantial sample sizes.[9] □

Illustration: Gamma MLE Let us rewrite the gamma loglikelihood function:

$$\ell(\alpha, \beta) = n\alpha \log \beta + (\alpha - 1)\sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i - n \log \Gamma(\alpha).$$

The partial derivatives are

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$
$$\frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i$$

where $\Gamma'(u)$ is the derivative of the function $\Gamma(u)$ (sometimes called the "digamma function"). Setting the second partial derivative equal to zero we obtain

$$\hat{\beta} = \frac{n\hat{\alpha}}{\sum_{i=1}^{n} x_i}.$$

When we set the first equation equal to zero and substitute this expression for $\hat{\beta}$, we get the nonlinear equation

$$n \log \hat{\alpha} - n \log \bar{x} + \sum_{i=1}^{n} \log x_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

To obtain the MLE $(\hat{\alpha}, \hat{\beta})$ we may proceed iteratively: given a value $\hat{\beta}^{(j)}$ we can solve the first equation for $\hat{\alpha}^{(j+1)}$ and solve the second equation to obtain $\hat{\beta}^{(j+1)}$; we

[9] We may obtain $\hat{\sigma} = s$ if we instead integrate out μ from the likelihood and then maximize the resulting function; this function is sometimes called an *integrated* or *marginal* likelihood, and in some situations maximizing the integrated likelihood yields a preferable estimator.

continue until the results converge. The second equation must be solved numerically, but it is not very difficult to use available software to do so. □

8.4.2 Least squares may be viewed as a special case of ML estimation.

In Example 1.5 we discussed data collected by Hursh (1939), indicating the linear relationship between a neuron's conduction velocity and its axonal diameter. We also briefly described the method of *least-squares regression*, based on the *linear regression model* (1.4), which is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{8.40}$$

where $\epsilon_i \sim N(0, \sigma^2)$, independently. Least-squares regression is discussed at length in Chapter 12. Here we show that the method of least squares may be considered a special case of ML estimation.

Least squares may be derived by assuming that the ϵ error variables in (8.40) are normally distributed, and that the problem is to estimate the parameter vector $\theta = (\beta_0, \beta_1)$. Specifically, we assume $\epsilon_i \sim N(0, \sigma^2)$, independently for all i. Calculation then shows that the ML estimate of θ is the least squares estimate. In other words, in the simple linear regression problem, ML based on the assumption of normal errors reproduces the least-squares solution.

Details: In the illustration on p. 210 we wrote down the loglikelihood function for a sample from a $N(\mu, \sigma^2)$ distribution,

$$\ell(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

and obtained the MLE $\hat{\mu} = \bar{x}$. Notice that, as a function of μ, the loglikelihood is maximized by minimizing the sum of squares $\sum_{i=1}^{n}(x_i - \mu)^2$. Thus, the MLE $\hat{\mu} = \bar{x}$ is also a least-squares estimator in the one-sample problem. For the simple linear regression model (8.40) the loglikelihood function becomes

$$\ell(\beta_0, \beta_1, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

We can maximize $\ell(\beta_0, \beta_1, \sigma)$ by first defining $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ to be the maximum of $\ell(\beta_0, \beta_1, \sigma)$ over (β_0, β_1) for fixed σ, and then maximizing $\ell(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma), \sigma)$ over σ. However, from inspection of the formula above, for every σ the solution $(\hat{\beta}_0(\sigma), \hat{\beta}_1(\sigma))$ (the maximum of $\ell(\beta_0, \beta_1, \sigma)$) is found by minimizing the sum of squares

$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. Therefore, the MLE $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ has the least-squares estimate as its first two components. □

8.4.3 The observed information is the negative of the matrix of second partial derivatives of the loglikelihood function, evaluated at $\hat{\boldsymbol{\theta}}$.

In the multiparameter case the second derivative $\ell''(\theta)$ becomes a matrix,

$$\ell''(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 \ell}{\partial \theta_1 \theta_m} & \frac{\partial^2 \ell}{\partial \theta_2 \theta_m} & \cdots & \frac{\partial^2 \ell}{\partial \theta_m^2} \end{pmatrix}.$$

This second-derivative matrix is often called the *Hessian* of $\ell(\theta)$. The *observed information matrix* is $-\ell''(\hat{\theta})$, which generalizes (8.34).

Result For large samples, under certain general conditions, the MLE $\hat{\theta}$ of the m-dimensional parameter θ is distributed approximately as an m-dimensional multivariate normal random vector with variance matrix

$$\hat{\Sigma} = -\ell''(\hat{\theta})^{-1}, \tag{8.41}$$

i.e.,

$$\hat{\Sigma}^{-1/2}(\hat{\theta} - \theta) \xrightarrow{D} N_m(0, I_m) \tag{8.42}$$

as $n \to \infty$.

Example 5.5 (continued from p. 112) In the Hecht et al. experiments on threshold for visual perception of light, the response variable was an indication of whether or not light was observed by a particular subject ("yes" or "no"), and the explanatory variable was the intensity of the light (in units of average number of light quanta per flash). Several different intensities were used, and for each the experiment was repeated many times. The results for one series of trials in one subject are plotted in Fig. 8.9.

As illustrated in Fig. 8.9, the linear regression model (8.40) does not work very well in this example. The proportions vary between 0 and 1 but a line $y = a + bx$ is unrestricted and can not represent the variation accurately, at least not for proportions that get close to 0 or 1. A solution is to replace the line $y = a + bx$ by a sigmoidal curve, which goes to zero as the explanatory variable x goes to $-\infty$ and increases

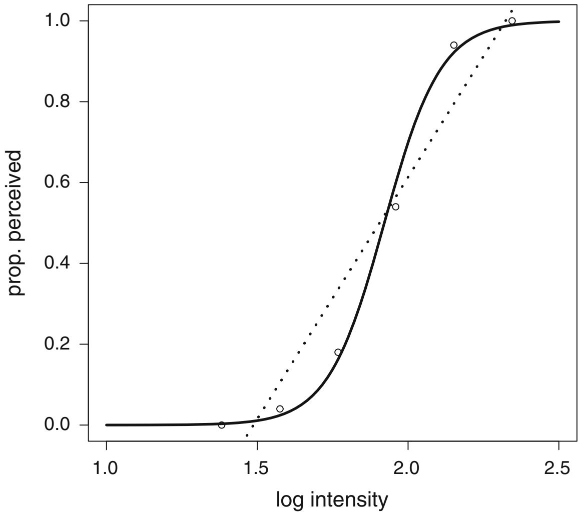

Fig. 8.9 Proportion of trials, out of 50, on which light flashes were perceived by subject S.S. as a function of $\log_{10}$ intensity, together with fits. Data from Hecht et al. (first series of trials) are shown as *circles*. *Dashed line* is the fit obtained by linear regression. *Solid curve* is the fit obtained by logistic regression.

to one as $x \to \infty$. The fitted curve in Fig. 8.9 is based on the following statistical model: for the ith value of light intensity we let Y_i be the number of light flashes on which the subject perceives light and then take

$$Y_i \sim B(n_i, p_i) \tag{8.43}$$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \tag{8.44}$$

This is known as the *logistic regression model*. There are many possible approaches to estimating the parameter vector $\theta = (\beta_0, \beta_1)$ but the usual solution is to apply maximum likelihood. The observed information matrix is then used to get standard errors of the coefficients. These calculations are performed by most statistical software packages. For the data in Fig. 8.9 we obtained $\hat{\beta}_0 = -20.5 \pm 2.4$ and $\hat{\beta}_1 = 10.7 \pm 1.2$. Further discussion of logistic regression, and interpretation of this result, are given in Section 14.1. □

8.4.4 When using numerical methods to implement ML estimation, some care is needed.

There are three issues surrounding the application of numerical maximization to ML estimation. The first is that, while loglikelihood functions are usually well behaved near their maxima, they may be poorly behaved away from the maxima. In particular, a loglikelihood may have multiple smaller peaks, and numerical methods may get stuck in a region away from the actual maximum. Except in cases where the loglikelihood is known to be concave (see Section 14.1.6.), it is essential to begin an iterative algorithm with a good preliminary estimate. Sometimes models may be altered and simplified in some way to get guesses at the parameter values. In some cases the method of moments may be used to get initial values for an iterative maximization algorithm.

Illustration: Gamma distribution On p. 153 we found the method of moments estimator for the Gamma distribution,

$$\beta^* = \frac{\bar{x}}{s^2}$$
$$\alpha^* = \frac{\bar{x}^2}{s^2}.$$

In order to obtain the MLE of (α, β) we may use an iterative maximization algorithm beginning with $(\hat{\alpha}^{(1)}, \hat{\beta}^{(1)}) = (\alpha^*, \beta^*)$. □

With good initial values, iterative maximization software usually only needs to run for a few iterations, after which the estimates don't change by more than a small fraction of the statistical uncertainty (represented by standard errors). In fact, it may be shown, theoretically, that from any consistent estimator for which the *MSE* vanishes at the rate $1/n$, a single iteration of Newton's method for maximizing the loglikelihood function will produce an efficient estimator (see Lehmann, 1983).

A second important implementation issue is that the second derivatives used in numerical maximization software are often themselves estimated numerically, and they may be estimated rather poorly (because they do not need to be estimated accurately to obtain the maximum). Thus, for the purpose of finding a variance matrix, one should either evaluate second derivatives separately (from an analytical formula, or from special-purpose software), or one should apply the parametric bootstrap (see Section 9.2).

The third issue is that parameterization can be important. Numerical maximization procedures tend to work well when the loglikelihood function is roughly quadratic, which means that the likelihood function is approximately normal. Transformations of parameters can improve this approximation. For example, before running maximization software it is often helpful to transform variance parameters by taking logs.

8.4.5 MLEs are sometimes obtained with the EM algorithm.

Certain statistical models have a structure that lends itself to a special method of likelihood maximization known as the *expectation-maximization (EM) algorithm*. We describe it in one special case.

Illustration: Mixture of Two Gaussians Suppose a random variable X follows either a $N(\mu_1, \sigma_1^2)$ distribution or a $N(\mu_2, \sigma_2^2)$ distribution, and that the selection of the distribution is determined probabilistically: with probability π we have $X \sim N(\mu_1, \sigma_1^2)$ and with probability $1 - \pi$ we have $X \sim N(\mu_2, \sigma_2^2)$. The pdf of X is

$$f_X(x) = \pi f(x; \mu_1, \sigma_1^2) + (1 - \pi) f(x; \mu_2, \sigma_2^2) \tag{8.45}$$

where $f(x; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ pdf. If we consider a large sample of values $x_1, \ldots, x_n$ from the distribution of X, some proportion of x_i values (approximately $n\pi$ of them) would be from the $N(\mu_1, \sigma_1^2)$ distribution, while the rest would be from the $N(\mu_2, \sigma_2^2)$ distribution. Such a sample would thus blend the $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions and (8.45) defines a *mixture* of two normal distributions, often called a *mixture of Gaussians* model. Based on a sample of data the problem is to estimate the parameter vector $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)$.

Let us introduce a random variable W_i to represent the selected distribution for X_i in the sense that $W_i = 1$ with probability π and if $W_i = 1$ a value U is drawn from $N(\mu_1, \sigma_1^2)$ and we set $X_i = U$, while $W_i = 0$ with probability $1 - \pi$ and if $W_i = 0$ a value V is drawn from $N(\mu_2, \sigma_2^2)$ and we set $X_i = V$. The variables $W_1, \ldots, W_n$ are not observed. If they were known, however, the problem would be much simpler: we could collect the values of W_i for which $W_i = 1$ and take the sample mean and variance of those as estimates[10] of μ_1 and σ_1^2 and then collect the values of W_i for which $W_i = 0$ and take the sample mean and variance of those as estimates of μ_2 and σ_2^2. Because the W_is are unobserved, they are often called *latent variables* (see Section 16.2). The data $(x_1, \ldots, x_n)$ are said to be *augmented* by $(w_1, \ldots, w_n)$. Let us write $Y = (X_1, \ldots, X_n)$ and $Z = (W_1, \ldots, W_n)$ and then write the loglikelihood function based on the original data y as $\ell_y(\theta)$ and that based on the augmented data (y, z) as $\ell_{(y,z)}(\theta)$. We have

$$\begin{aligned}
\ell_{(y,z)}(\theta) &= \sum_{i=1}^{n} w_i \log f(x_i; \mu_1, \sigma_1^2) + \sum_{i=1}^{n} (1 - w_i) \log f(x_i; \mu_2, \sigma_2^2) \\
&\quad + \sum_{i=1}^{n} w_i \log \pi + \sum_{i=1}^{n} (1 - w_i) \log(1 - \pi) \\
&= \sum_{\{i: w_i = 1\}} \log f(x_i; \mu_1, \sigma_1^2) + \sum_{\{i: w_i = 0\}} \log f(x_i; \mu_2, \sigma_2^2) \\
&\quad + \sum_{\{i: w_i = 1\}} \log \pi + \sum_{\{i: w_i = 0\}} \log(1 - \pi)
\end{aligned} \tag{8.46}$$

and maximizing this with respect to (μ_1, σ_1^2) is the same as maximizing the likelihood for a sample a $N(\mu_1, \sigma_1^2)$ pdf made up of the values x_i for which $w_i = 1$ (and similarly

[10] As we said in Section 8.4.1 (see p. 210), the MLE of the variance has denominator n rather than $n - 1$ but the sample variance is usually preferred.

for (μ_2, σ_2^2)). Thus, the introduction of the latent variables W_i has greatly simplified the problem. However, because these latent variables have not been observed we must get estimates that do not rely on them. To do this we may integrate out the W_i variables (marginalize over them), as we next explain.

If we think of π as a prior probability that $W_i = 1$ then, after observing $X_i = x_i$ we may compute the posterior probability from Bayes' Theorem as

$$P(W_i = 1 | X_i = x_i) = \frac{\pi f(x_i; \mu_1, \sigma_1^2)}{\pi f(x_i; \mu_1, \sigma_1^2) + (1 - \pi) f(x_i; \mu_2, \sigma_2^2)}. \tag{8.47}$$

We use the notation

$$\gamma_i = P(W_i = 1 | X_i = x_i). \tag{8.48}$$

Note that, because W_i is a binary variable γ_i may also be written

$$\gamma_i = E(W_i | X_i = x_i)$$

and, for later purposes, we make the dependence on θ explicit by writing

$$\gamma_i = E(W_i | X_i = x_i, \theta). \tag{8.49}$$

With this framework in hand, the EM algorithm for this problem may be defined. It produces an iterative sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ that, with good initial values, will converge to the MLE $\hat{\theta}$. Here is the algorithm.

1. Find an initial value $\theta^{(1)}$ for θ and set $j = 1$.
2. Given a current value $\theta^{(j)}$ compute $\gamma_i^{(j)}$ for $i = 1, \ldots, n$ by applying (8.48) using (8.47) where $\theta = \theta^{(j)}$.
3. Using $\gamma_1^{(j)}, \ldots, \gamma_n^{(j)}$ from Step 2 compute the components of $\theta^{(j+1)}$ as follows:

$$\begin{aligned}
\mu_1^{(j+1)} &= \frac{\sum_{i=1}^n \gamma_i^{(j)} x_i}{\sum_{i=1}^n \gamma_i^{(j)}} \\
\mu_2^{(j+1)} &= \frac{\sum_{i=1}^n (1 - \gamma_i^{(j)}) x_i}{\sum_{i=1}^n (1 - \gamma_i^{(j)})} \\
\sigma_1^{2(j+1)} &= \frac{\sum_{i=1}^n \gamma_i^{(j)} (x_i - \mu_1^{(j+1)})^2}{\sum_{i=1}^n \gamma_i^{(j)}} \\
\sigma_2^{2(j+1)} &= \frac{\sum_{i=1}^n (1 - \gamma_i^{(j)})(x_i - \mu_1^{(j+1)})^2}{\sum_{i=1}^n (1 - \gamma_i^{(j)})} \\
\pi^{(j+1)} &= \frac{1}{n} \sum_{i=1}^n \gamma_i^{(j)}.
\end{aligned}$$

4. Increment j and return to Step 2.
5. Repeat Steps 2–4 until convergence. □

A key step in formulating the EM algorithm in the mixture of two Gaussians model, above, was the introduction of the random variables W_i. In order to maximize the loglikelihood $\ell_Y(\theta)$ defined by the pdf $f_Y(y|\theta)$ we effectively introduced the loglikelihood $\ell_{(Y,Z)}(\theta)$ in (8.46) based on the augmented data pdf $f_{(Y,Z)}(y,z|\theta)$. Step 2 of the algorithm, known as *the expectation step*, is based on the expectation $E(\ell_{(Y,Z)}(\theta)|Y = y, \theta = \theta^{(j)})$. In Step 2 the conditional expectation in (8.49) was evaluated for $\theta = \theta^{(j)}$. In Step 3 the loglikelihood was maximized in terms of the expectations computed in Step 2.

In general, if $Y = y$ is the data vector augmented by $Z = z$ we define

$$Q(\theta, \theta^{(j)}) = E(\ell_{(Y,Z)}(\theta)|Y = y, \theta = \theta^{(j)}). \tag{8.50}$$

Beginning with an initial guess $\theta^{(1)}$, for each j the EM algorithm computes $Q(\theta, \theta^{(j)})$ and sets $\theta^{(j+1)}$ equal to the maximizer of $Q(\theta, \theta^{(j)})$ as a function of θ. The EM algorithm works well for problems in which some kind of data augmentation greatly simplifies the problem, so that $Q(\theta, \theta^{(j)})$ is easy to compute (as in Step 2 of the mixture of two Gaussians illustration above). In addition to models that incorporate latent variables, the EM algorithm is often applied to problems with missing data, where the missing data are treated as augmenting the observed data. (See also the related discussion of Gibbs sampling in Section 16.2.2.)

One way to see that this iterative scheme should work is to apply the formula[11]

$$\frac{d}{d\theta}Q(\theta, \theta^*)|_{\theta=\theta*} = \ell'_Y(\theta^*) \tag{8.51}$$

(see the details below). If $\theta^{(1)}, \theta^{(2)}, \ldots$ is a sequence of EM iterates that converge to a value θ^* then, because each iterate maximizes $Q(\theta, \theta^{(j)})$ its derivative is 0, i.e.,

$$\frac{d}{d\theta}Q(\theta, \theta^*)|_{\theta=\theta*} = 0.$$

From (8.51) we then have

$$\ell'_Y(\theta^*) = 0.$$

Thus, for sufficiently good initial values, when the EM algorithm converges to θ^* we get $\theta^* = \hat{\theta}$, i.e., the EM algorithm converges to the MLE $\hat{\theta}$.

Details: We derive Eq. (8.51). From (8.50) we have

$$Q(\theta, \theta^*) = \int \frac{f(y, z|\theta^*)}{f(y|\theta^*)} \log f(y, z|\theta) dz.$$

[11] This formula was used by Fisher, in his discussion of sufficiency, to substantiate the argument mentioned in Section 8.2.2 (see p. 200 and Kass and Vos 1997, Section 2.5.1)

We differentiate under the integral:

$$\begin{aligned}\frac{d}{d\theta}Q(\theta,\theta^*)|_{\theta=\theta*} &= \int \frac{f(y,z|\theta^*)}{f(y|\theta^*)}\frac{\frac{d}{d\theta}f(y,z|\theta)|_{\theta=\theta*}}{f(y,z|\theta^*)}dz\\ &= \int \frac{\frac{d}{d\theta}f(y,z|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz.\end{aligned}$$

We continue using $f(y,z|\theta) = f(z|y,\theta)f(y|\theta)$, differentiate the product, and rewrite:

$$\begin{aligned}&\frac{d}{d\theta}Q(\theta,\theta^*)|_{\theta=\theta*}\\ &= \int \frac{\frac{d}{d\theta}f(z|y,\theta)f(y|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz\\ &= \int \frac{f(y|\theta^*)\frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*} + f(z|y,\theta^*)\frac{d}{d\theta}f(y|\theta)|_{\theta=\theta*}}{f(y|\theta^*)}dz\\ &= \int \frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*} + f(z|y,\theta^*)\frac{d}{d\theta}\log f(y|\theta)|_{\theta=\theta*}dz.\end{aligned} \tag{8.52}$$

In this last expression the integral of the first term vanishes because

$$\int f(z|y,\theta)dz = 1$$

so that

$$\frac{d}{d\theta}\int f(z|y,\theta)dz = 0$$

and taking the derivative under the integral gives

$$\int \frac{d}{d\theta}f(z|y,\theta)|_{\theta=\theta*}dz = 0.$$

Therefore, expression (8.52) reduces to (8.51). □

8.4.6 Maximum likelihood may produce bad estimates.

The method of ML is not universally applicable, nor does it guarantee good statistical results. The most serious concern with ML is that it is predicated on the description of the data according to a particular statistical model. If that model is seriously deficient, the MLE will be misleading. This underscores the essential role of model assessment, and the iterative nature of model building, emphasized in Chapter 1.

The provably good performance of ML estimation also applies only for large samples. What constitutes "large" is difficult to specify precisely, though attempts have been made occasionally. A key observation is that sample size must be judged relative to the number of parameters being estimated. In problems having large numbers of parameters and only modest sample sizes, we should expect neither ML estimates, nor their associated SEs, to be accurate. One standard approach to making progress in such situations is to build models that effectively reduce the number of parameters by restricting them in some way (often by introducing additional probability distributions). In some cases, however, ML must be abandoned. There is a large body of methods that are *nonparametric*, in the sense that they do not posit a statistical model with a finite number of parameters. There are many situations where nonparametric methods perform well, and save the difficulty and worry associated with careful model building.