

Chapter 5

Parameter Estimation

5. Introduction

Many geophysical problems involve the need to estimate some set of unknown quantities from a collection of observational data. We gave a simple example at the end of the previous chapter, where we assumed that measurements of distance x for a falling body at different times could be modeled by a random variable

$$X_i \sim \phi(x - \frac{1}{2}gt_i^2) \quad (1)$$

with ϕ being some known pdf, assumed the same for all the measurements. The **point estimation problem** would then be, given actual data x_1, x_2, \dots, x_n , at times t_1, t_2, \dots, t_n , to find the “best” value of the parameter g —and, though this is a more subtle problem, within what range we think it might lie. *Note* that we use a lower-case variable x to denote actual data—these are not random variables, but some set of definite values. Also note that we make the number of data n , partly to distinguish this from the dimension of a multivariate rv that we discussed in the previous chapter—though we will see soon enough that understanding estimation will require multivariate pdf’s.

5.1. A Simple Example: Three Sets of Estimates

We will start with a simpler case, which is the first dataset shown in Chapter 1. We assume that the data, x_1, x_2, \dots, x_n , can be modeled by a random variable

$$X_i \sim \phi\left(\frac{x-l}{c}\right)$$

where the two parameters are the location l and the spread c . We begin by discussing some ways in which we might estimate these, partly to introduce some conventional and useful methods.

5.1.1. The Method of Moments

As an estimate of l we might use the arithmetic average, or mean, often called the **sample mean**:¹

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{analogous to} \quad \mu = \int_{-\infty}^{\infty} x \phi(x) dx \quad (2)$$

where the “analogous to” is the same kind of operation performed on a pdf to get (in this case) the first moment, or expected value. If we make the same analogy for the second moment, we get, as an estimate of the variance, the **sample variance**:

¹Th term “sample” comes from the idea that there is a large (potentially infinite) collection of random variables, from which we have chosen a sample of n values. However appealing this image may be for the sciences in which such a population exists (which are quite a few, ranging from astronomy to economics), this concept does not seem appropriate for many parts of geophysics (though it does work for statistical seismology). We shall continue to say, instead, that we model data by random variables, and use the idea of a population as little as possible.

$$s^2 = \frac{1}{n} \left[\sum_{i=1}^n (x_i - \bar{x})^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i - \bar{x} \right)^2 \right] \quad \text{analogous to} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \phi(x) dx \quad (3)$$

The expression for s^2 has two summations; in an exact computation the second one would be identically equal to zero, but in an actual, finite-precision, computation it improves the accuracy for large n .²

Equations (2) and (3) illustrate the **method of moments** as a procedure for estimating parameters, particularly ones that directly describe the pdf. In principle, if we know all the moments of a pdf we know everything there is to know about it, and so can deduce all the parameters it includes. So using the method of moments means estimating the sample moments of the data (something that is computationally easy to do) and taking these to be the moments of the pdf. Of course, in practice we cannot find all the moments, so unless we are quite sure of the form of the pdf, this procedure is somewhat limited, and may be misleading; so we do not recommend it. We will see it in use below, when we come to the Monte Carlo and bootstrap methods of evaluating estimation methods—though in these it is used, not for the data, but for results derived from them.

In the case of our GPS data, the results are $\bar{x} = -0.0234$ and $s^2 = 0.407$ ($s = 0.638$). If we compare the distribution of the data (as shown by the histogram) with the moments we may feel that something has gone badly wrong, at least with the second moment: it seems much too large. The reason is fairly simple: while almost all the data are between -1 and 1 (see Figure 5.1), there is one value at -10.32 , one at -1.56 , and one at 6.34 . Including these in the arithmetic average (equation 3) might bias the sample mean based on only a tiny fraction of the data; but these three outlying values unquestionably have a large effect on inflating the estimated second moment.

5.1.2. Order Statistics

Given these results, we might seek estimation procedures that are less affected by what values are taken on by a very small fraction of the data; the technical term for such procedures is **robust**. One method of getting robust estimates is to sort the data into increasing order

$$x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)}$$

where the use of parentheses on the subscripts is the standard notation to indicate a sorted set of values. Procedures that make use of sorted data are called **order statistics**. One such, which provides an estimate of the location parameter l , is the **sample median**

$$x_{med} = \frac{1}{2} [x_{(n/2)} + x_{(n/2+1)}] \quad \text{analogous to} \quad \int_{-\infty}^{x_m} \phi(x) dx = \int_{x_m}^{\infty} \phi(x) dx$$

where the definition given for x_{med} is for n even (true in the case of these GPS data, with $n = 386$); for n odd, $x_{med} = x_{(n/2)}$. For the spread, we can use the **interquartile range**, which is just $x_{(0.75n)} - x_{(0.25n)}$. If we were to apply this to a normal

² Chan, T. F., G. H. Golub, and R. J. LeVeque (1983), Algorithms for computing the sample variance: analysis and recommendations, *Amer. Statistician*, **37**, 242-247.

distribution, we would find that the IQR is 1.349σ ; for the GPS data the IQR is 0.14, giving 0.10 for the estimate of the equivalent of the standard deviation. This is much more satisfying as an approximation to the spread seen in the histogram.

5.1.3. Trimmed Estimates

However, we might be concerned that these two procedures do not make enough use of the data. For the median, the actual values of almost all the data are irrelevant, except to set the midpoint; only one or two values actually enter into the calculation. As an intermediate approach, we could consider removing data from the two ends of the sorted distribution, and computing what would then be called a **trimmed mean**. Note that this is still an order statistic, dependent on having the data sorted. For example, the 10% trimmed mean is

$$\bar{x}_{10\%} = \frac{1}{0.9n} \sum_{i=0.05n}^{0.95n} x_{(i)}$$

and the 10% **trimmed variance** is

$$s_{10\%}^2 = \frac{1.64}{0.9n} \sum_{i=0.05n}^{0.95n} (x_{(i)} - \bar{x}_{10\%})^2$$

where the constant in the expression applies for large n (and 10% trimming), and assumes that the pdf is in fact normal. The resulting values for the mean and variance are $\bar{x}_{10\%} = -0.0125$ and $s_{10\%}^2 = 0.112$, which are much closer to the values from the median and IQR.

But, all this is more than a bit *ad hoc*. How can we decide which of these methods is better than another, either globally, or for a particular case? Answering that question is what we turn to next.

5.2. Three Ways to Test an Estimator

We begin with some terminology. An **estimator** is some algorithm for finding the value of a parameter; we have described three such procedures above. What such a procedure produces when applied to data is an **estimate**. So the question we are discussing is the relative performance of different estimators. We describe three ways to assess this: the classical method using analytic procedures; Monte Carlo methods; and the bootstrap. (As usual, more cryptic names—but their meaning will become clearer). Each has its advantages and problems; we examine them using the example just given, of finding parameters for the location and spread of the pdf used to model the data.

5.2.1. The Classical Method of Evaluating a Statistic

The title of this subsection may have caused you to stop in confusion, because we were just promising to evaluate estimators, and now are talking about “a statistic”, which at first sight would appear to be the singular of the field, statistics, we are studying. But in fact a statistic³ is yet another kind of random variable, and one

³ Blame where blame is due: this unhappy terminology is due to R. A. Fisher, whose policy seems to have been never to coin a name from Greek or Latin if a common English word could be used instead—the difficulty being that such words often had a common meaning at odds with the

we use to evaluate an estimator.

To state more generally the problem we are trying to solve, we suppose that we have n observations x_1, x_2, \dots, x_n that we model by n random variables X_1, X_2, \dots, X_n . These random variables have a joint probability distribution of

$$\phi(x_1, x_2, \dots, x_n, \theta_1, \theta_2, \dots, \theta_p)$$

where the values of the parameters $\theta_1, \theta_2, \dots, \theta_p$ are unknown, and are to be determined from the data, x_1, x_2, \dots, x_n . Note that we use lower-case letters (for example, x) to denote both data and the arguments to the pdf for the X 's. Since the observations are not used as arguments in the pdf, this should not cause confusion.

You need to appreciate that the above problem, of estimating the θ 's given the x 's, is quite different from what we have studied in probability theory. In those problems, we are given the parameters $\theta_1, \theta_2, \dots, \theta_p$ and use the function $\phi(x_1, x_2, \dots, x_n, \theta_1, \theta_2, \dots, \theta_p)$ to tell us the probability of obtaining values of the random variables X_1, X_2, \dots, X_n that fall within specified ranges.

In estimation we are given the measured values of x_1, x_2, \dots, x_n , and try to make statements, perhaps using the pdf $\phi(x_1, x_2, \dots, x_n, \theta_1, \theta_2, \dots, \theta_p)$ about the “best values”, and range of reasonable values, for the different parameters $\theta_1, \theta_2, \dots, \theta_p$. Estimation is thus a kind of “inverse” to probability theory—*except* that in this framework neither the data nor the parameters are random variables, so we cannot make probability statements about them.⁴

How do we express what the data tell us about the parameters, when neither is a random variable. What we do is to introduce a random variable that “looks like” a parameter. To keep matters simple, we assume we have only one random variable X , and one parameter θ , so the pdf of X is $\phi(\theta)$. We assume that X is a model for the data x_1, x_2, \dots, x_n ; note that there can be n data even if there is only one X .

Next, suppose we have some estimator (such as the average, the median, or the trimmed mean) that supplies us, given data, with a number that we think is an estimate of the parameter θ . We can imagine applying this estimator (which is just an algorithm), not to the data, but to the random variable we are using as a model. Performing this algorithm on n random variables X will give us a new random variable, which we symbolize by $\hat{\theta}$. This variable $\hat{\theta}$ will have some pdf, which is called the **sampling distribution** of $\hat{\theta}$,⁵ which we write as $\hat{\phi}(X_1, X_2, \dots, X_n)$. This pdf depends on

1. The pdf for X .
2. The way the estimator is computed; that is, its functional form.
3. The number of data n .

technical one. However, considering such coinages from the Greek as “homoscedasticity”, perhaps his choice was understandable.

⁴ We say “in this framework” because there is another way to look at the problem, namely the **Bayesian** approach, in which we take the parameters to indeed be modeled by random variables, with the meaning of the associated probabilities then being our degree of belief in a particular range of values. For reasons of space we have not introduced this methodology, which certainly deserves more than a footnote!

⁵ This term comes from the idea of sampling from a population.

The random variable $\hat{\theta}(X_1, X_2, \dots, X_n)$ is called the **statistic** corresponding to θ . The **estimate** is the result of applying the estimation procedure to actual data, $\hat{\theta}(x_1, x_2, \dots, x_n)$.⁶ Note that the statistic is a random variable; the estimate is a conventional variable. Because the statistic is a random variable, but is tied to something we do with data, it provides the theoretical basis for evaluating different estimation procedures.

To see how this works, suppose we have n data that we assume can be described by a normal distribution with mean μ and variance σ^2 . Note that we do not, for the following discussion, need to know what the values of these parameters actually are; we do of course need to know the functional form of the pdf. The procedure (2), applied to these random variables, will produce another random variable, the statistic,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

The sampling distribution of this statistic is given by the convolution of n normal pdf's, divided by n : this pdf is again normal, with mean μ and variance σ^2/n , a result we derived in Chapter 4. So, if we are justified in our assumption that we can model the data by random variables with the normal distribution, we know that: (1) the expected value of the statistic is equal to the actual value of the parameter; and (2) the variance of the statistic decreases as the number of data, n , increases.

What about the median? If we apply this procedure to the same distribution of random variables (normal) we get, again, a new random variable. It turns out that the expected value of this is again μ ; the variance is approximately

$$\mathcal{V}[\hat{\mu}_{med}] = \frac{\pi \sigma^2}{2(n+2)} \left[1 + \frac{\pi}{2(n+4)} \right] \quad (4)$$

which for large n is 1.57 times the variance of the sample mean, $\hat{\mu}$.

5.2.2. Monte Carlo Methods

The method we have just described can be most tersely summarized as, find the pdf of some combination of n random variables with known pdf's (assumed to model the data). If we can do all this analytically, taking this approach will give the most rigorous result; but sometimes there is no analytical result readily available. If not, we can replace analysis with computation, using what is called a **Monte Carlo** evaluation⁷

The basic Monte Carlo procedure is simply described:

1. Generate n random variables with the assumed distribution.
2. Compute an estimate (call it $\hat{\theta}_1$) from these imitation data.

⁶ [Notation alert:] we may (as here) use the same function to describe an estimate and a statistic; what sets them apart is their arguments.

⁷ The term of course comes from the famous (less so now than when statistics was founded) gambling center. There are lots of Monte Carlo methods, actually; we refer only to those in statistics. We should also note that it is common to call any method that uses computed random numbers a Monte Carlo method, which makes the bootstrap (discussed below) such a method. For clarity, we have taken a more restricted meaning here.

3. Repeat steps 1 and 2 a total of k times, to get estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$.
4. Find the empirical first and second moments of the $\hat{\theta}$'s:

$$\bar{\theta} = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_i \quad \sigma_{\hat{\theta}}^2 = \frac{1}{k} \sum_{i=1}^k (\hat{\theta}_i - \bar{\theta})^2$$

and take these to be the first and second moments of the pdf of the statistic $\hat{\theta}$. Since this is all done numerically, we can be assured that our imitation X 's have the distribution we want. We may be interested in other properties of the pdf (e.g., the bounds within which 95% of the mass lies); these can also be found from the $\hat{\theta}$'s.

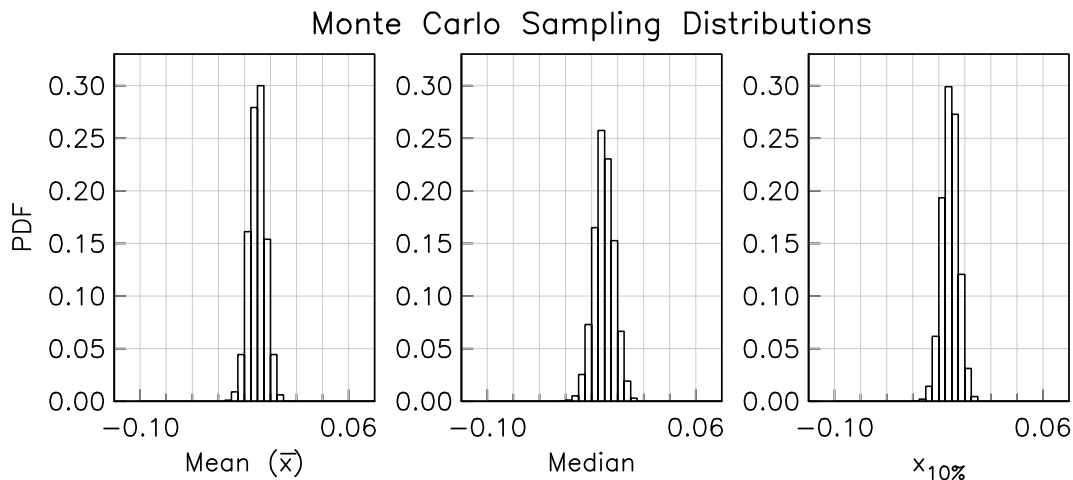
The parallel with the analytic procedure should be clear:⁸ we assume we can represent the data by random variables, from which we form a statistic. The difference is that, rather than trying to find the distribution of this statistic analytically, we do it through lots and lots of computation, generating a large number of (simulated) statistics and then looking at their distribution.

How does this method compare with the classical approach?

- A. So far as the assumptions involved go, it is the same: we assume we know the distribution the data come from (including all the parameters), which may or may not be valid. (We will see in the next section how we can escape this assumption).
- B. It makes many fewer demands on our analytical skills. So long as we have an algorithm for generating appropriate random variables, and another for finding the estimates from data, we can proceed—of course we always have the latter. Confronted with some novel estimation procedure, and given ever-faster computers, Monte Carlo methods may often be the quickest way to find out the sampling distribution of the statistic.
- C. However, the replacement of the analytical determination of the pdf by step (4), finding it empirically, can be a disadvantage, because it may take a very large value of k to make this empirical distribution well-determined. If all we want to do is find the expected value and variance (the first two moments) a relatively small number of simulations will do; but if we want to determine, reliably, such features as the limits within which 95% of the sampling distribution lies, we are trying to make a statement about the tails of the sampling distribution—and this inevitably requires a great many simulations.
- D. Monte Carlo methods can only answer the question “How good is this estimator?” They cannot provide information on whether an estimator is, in some sense, the best possible—that is to say, **optimal**. For this reason we will spend a fair amount of time, in the rest of this chapter, on analytical methods, precisely to lay out some general procedures for finding and evaluating estimators.

Monte Carlo methods are seeing increasing use, and we certainly suggest you consider them—but they should, in general, be the second choice, after at least some

⁸ And this is a situation in which the idea of sampling from a population makes perfect sense.



attempt to find an analytical result for the sampling distribution of the statistic. And, if you have derived your own analytical result, we strongly suggest doing a Monte Carlo computation as a check!

Figure 5.2 shows an example of a Monte Carlo calculation, for the sample mean, median, and 10% trimmed mean, using simulated normal random variables with the mean and standard deviation taken from the trimmed estimates for the GPS data. As expected, the distribution of the sample median is somewhat broader than that of the sample mean; in fact, the ratio of the variances is 1.55, very close to the 1.57 expected for this large a value of n . The Monte Carlo method easily handles the 10% trimmed mean, which has a variance essentially the same as that of the sample mean.

5.2.3. Lifting Ourselves by Our Bootstraps

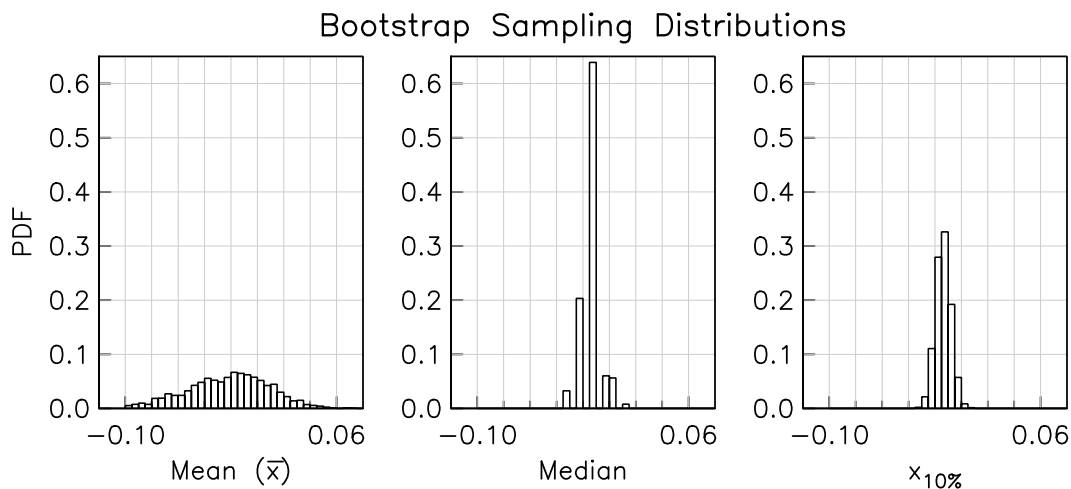
Our final procedure for evaluating an estimator is also computationally intensive, and very like what we have called Monte Carlo, with one big exception: instead of assuming we know the distribution of the random variables that we use to model the data, we assume that the data themselves can be used as the model random variables—we do not need to compute such variables at all. This seems so much like getting something for nothing—or at least, more than we ought to get—that it provoked the inventors of the method to refer to the image used as the title of this section. For this reason such an approach is known as a **bootstrap method**, a bit of jargon now firmly embedded in the statistical lexicon.

A bootstrap approach to finding the sampling distribution of a statistic works as follows:

1. Generate n random variables distributed as integers j_1, j_2, \dots, j_n , with a uniform distribution from 1 to n . Note that this means that some of these integers will, almost certainly, be identical. Use, as imitation random variables, the data indexed by these integers (that is, x_{j_i}); again, some of these values will be the same data values, taken more than once—for which reason this procedure is called sampling with replacement. (Imagine taking

data values out of a container at random, and putting each one back after you record it).

2. Compute an estimate (call it $\hat{\theta}_1$) from these imitation random variables.
3. Repeat steps 1 and 2 a total of k times, to get estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$.
4. Find the empirical first and second moments of the $\hat{\theta}$'s, or whatever other property we are interested in.



The parallel with the Monte Carlo method is clear, not least in that this is something that can be done only with access to significant amounts of computer power—for which reason these methods were developed relatively recently. A major strength is the paradoxical feature that we do not need to select a pdf—this is done for us, as it were, by the data; and the method is certainly straightforward. At the same time, there is the same disadvantage as Monte Carlo, that a large number of iterations (large k) may be needed to determine some aspects of the sampling distribution. If the number of data n is not large, it is possible that k may be large enough that we exhaust all possible arrangements of the data—in which case the randomness becomes less clear.

Figure 5.3 shows the bootstrap applied to the GPS data, with $k = 4000$, for the same three estimators as were used in the Monte Carlo procedure in the previous section. The presence of a few outliers makes a huge difference, with the mean being much worse than either the median or the trimmed mean. The lowest variance of the three is for the 10% trimmed mean, with 0.72 of the variance of the median (the histogram for the median looks odd because the original data, one of which will form the median value, are only given to multiples of 0.01; so every other bin is empty). In this case the bootstrap has made clear that the trimmed mean is the preferred approach; and the Monte Carlo simulation of the previous section has shown that even for normal data, it has a variance that is nearly the same as that of the mean.

The biggest danger in using the bootstrap is that the data will not obey one of the restrictions that are required for the method to work. For example, the theory

requires that the values be independent—something easy enough to assure if we are creating random variables on a computer, but less easy to be sure about in a real data set. (Indeed, this is probably not true for the GPS data, though not to an extent that would significantly vitiate the results given here).

Finally, we note that what we have described here as “the bootstrap” is actually one of a number of bootstrap methods. Suppose, for example, we used the Monte Carlo methods of the previous section, but, instead of assuming *a priori* the parameters for the pdf of our computed random variables, took them from estimates from the data (of course, we sort of did this in our example). This would be called a **parametric bootstrap**: we are, again, using the data, though this time not directly but through a simulation of data with a known pdf but parameters determined from the data. This would be appropriate, for example, when we have so few data that direct use of the data values themselves would not be possible for more than a few simulations.

5.3. Confidence Limits for Statistics

Up to this point we have talked only about the first two moments of the sampling distribution of a statistic: the expected value $\mathcal{E}[\hat{\theta}] = \hat{\theta}$ and the variance $\mathcal{V}[\hat{\theta}]$. If we define $\sigma_{\hat{\theta}} =_{\text{def}} (\mathcal{V}[\hat{\theta}])^{\frac{1}{2}}$, then it is conventional to say that our estimate $\hat{\theta}(\vec{x})$ has a standard error of $\sigma_{\hat{\theta}}$, and we write the result, again conventionally, as being $\theta \pm \sigma_{\hat{\theta}}$.

This is often interpreted as being a statement about the range within which θ , the true value, will fall most of the time—but if we look at this statement carefully, we realize that this is nonsense. Any statement that includes something like “most of the time” is a probabilistic one; but we cannot make probabilistic statements about the true value of something, for this is a conventional variable and not a random one. And furthermore, just saying “ $x \pm \sigma$ ” is not even a statement of probability, for it does not have any information about a pdf (or rather, has only the first two moments).

How can we make this more sensible and more precise?

The answer is what are called **confidence limits**. These arise by broadening our view from the question of finding a single value for a statistic (called a **point estimate** of a parameter), to finding the probable range for the statistic (called an **interval estimate**). Note that since we said “a statistic” we can talk about probability, since a statistic is a random variable.

To make the discussion more concrete, suppose that the sampling distribution of $\hat{\theta}$ is normal. Then, 0.95 (95%) of the area under the pdf will lie between $\pm 2\sigma_{\hat{\theta}}$, relative to the center of the pdf. But, the center of the pdf is at $\hat{\theta} = \mathcal{E}[\hat{\theta}(\vec{X})]$. Let us further assume that $\mathcal{E}[\hat{\theta}] = \theta$ (that is, the statistic is unbiased—we discuss this in more detail in the next section). Then the probability statement we can make is

$$\mathcal{P}[\theta - 2\sigma_{\hat{\theta}} \leq \hat{\theta} \leq \theta + 2\sigma_{\hat{\theta}}] = 0.95 \quad (5)$$

Now this is not a statement about what interval the true value θ will fall in, but rather a statement about how often the random variable $\hat{\theta}$ will fall into an interval around the true value θ . This is not very satisfying, since we do not know θ .

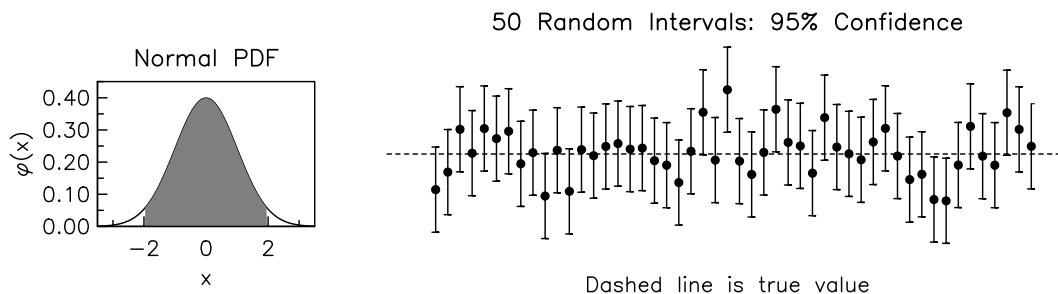


Figure 5.4

But, we can make this more useful, while retaining the probabilistic statement, if we look at the **random interval** $[\hat{\theta} - 2\sigma_{\hat{\theta}}, \hat{\theta} + 2\sigma_{\hat{\theta}}]$, which would be a **95% confidence interval**. Then the implication of the distribution of $\hat{\theta}$, equation (5), is that this interval will cover (that is include) the true value, 95% of the time; we write this as

$$\mathcal{P}[\hat{\theta} - 2\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + 2\sigma_{\hat{\theta}}] = 0.95 \quad (6)$$

You might be forgiven for thinking that in this equation we have just done what we said we shouldn't: making a probabilistic statement about θ , by giving a probability that it lies between two limits. However, there is in fact no problem (though some chance of confusion) because the limiting values are now random variables, not conventional ones. Expressions of the type of (6) are standard in the statistical literature; they should be taken to be a probability statement about an interval around $\hat{\theta}$, not a statement about the probability of θ —for the latter is, once again, not a random variable.⁹

Figure 5.4 shows how the concept of a random interval would work; on the left we show a pdf for the random variable with the 95% range shaded in; this range gives the size of the random interval. On the right we show 50 examples of an estimate, with the error bars indicating the range of the random interval at the same level of confidence (0.95, or 95%), along with the true value. We see that 3 of the intervals do not cover this value; we would expect 2 to 3. What this means is that, given a single estimate and confidence interval based on a particular dataset, we have high confidence that the true value lies within the interval, since it is relatively improbable that the interval would not include it.

We should note that the idea of a confidence interval can be made more general. What we have done above is to find two values for the random variable, $\hat{\theta}$, say $\hat{\theta}_1 < \hat{\theta}_2$, such that the pdf of $\hat{\theta}$ means that

$$\mathcal{P}[\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2] = 1 - \alpha$$

so that in the case above $\alpha = 0.05$ (a smaller α means a longer interval, in general). Again, this is a statement about the random interval $[\hat{\theta}_1, \hat{\theta}_2]$ compared to the true value θ . This can actually be broken down into the two statements of the reverse

⁹ We should note one slight ambiguity in the idea of a confidence interval: it is not uniquely determined by the requirement that it contain a certain fraction of the probability, since different limits can produce this. The usual approach is to make whatever choice minimizes the length of the interval.

case that $\hat{\theta}$ lies outside an interval:

$$\mathcal{P}[\theta \leq \hat{\theta}_1] = \alpha_1 \quad \mathcal{P}[\theta \geq \hat{\theta}_2] = \alpha_2 \quad \text{with} \quad \alpha_1 + \alpha_2 = \alpha \quad (7)$$

Usually, we take $\alpha_1 = \alpha_2$, and take the shortest interval consistent with the probability statement (7). If we have a symmetric pdf about the expected value $\mathcal{E}[\hat{\theta}]$, say $\phi_{\hat{\theta}}$, with cumulative distribution function $\Phi_{\hat{\theta}}$, the confidence limits will be given by the inverse of the cdf, and the interval will be

$$[\Phi^{-1}(0.5) + \Phi^{-1}(\alpha/2), \Phi^{-1}(0.5) + \Phi^{-1}(1 - \alpha/2)]$$

where we have made use of the fact that for a symmetric pdf, the expected value has $\Phi(\mathcal{E}(\hat{\theta})) = 0.50$. (We can add the values since $\Phi^{-1}(\alpha/2) < 0$ for a normal distribution).

5.3.1. Confidence Limits for the Mean, Variance Unknown

To give a more complete example, we look at the confidence limits for the mean when we do not know the variance, but have to estimate it. We assume we can model our data with normally distributed random variables with true variance σ^2 and true mean μ . We estimate the mean using \bar{x} and the variance as s^2 , as in equations (2) and (3). The statistic, $\hat{\mu}$, for \bar{x} , will be distributed, as we described in Section 5.2.1, as a normal random variable with variance σ^2/n .

The statistic for s^2 , being related to the sum of the squares of a set of normally-distributed random variables, turns out to be distributed as χ_{n-1}^2 : a chi-squared variable with $n-1$ degrees of freedom. To show this we note that the variable

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n [(X_i - \hat{\mu}) + (\hat{\mu} - \mu)]^2 \quad (8)$$

is exactly a sum of squares of normal random variables, and so distributed as χ_n^2 . But if we compute the square in (8), the cross-product term

$$\sum_{i=1}^n 2(X_i - \hat{\mu})(\hat{\mu} - \mu)$$

vanishes by the definition of $\hat{\mu}$. Hence (8) is

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \hat{\mu})^2 + \frac{n(\hat{\mu} - \mu)^2}{\sigma^2}$$

Now, it can be shown that $\hat{\mu}$ is independent of any of the X_i 's. The second term on the right-hand side is the square of a normally distributed rv, and so is distributed as χ_1^2 . The left-hand side is distributed as χ_n^2 . The sum of a random variable distributed as χ_{n-1}^2 , and an independent variable distributed as χ_1^2 , is distributed as χ_n^2 ; this is somewhat obvious from the construction of χ^2 from sums of rv's. Therefore, the distribution of

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \hat{\mu})^2$$

is, as stated before, χ_{n-1}^2 .

Having established the distribution of the sample mean and variance, we construct a scaled ("standardized") variable by subtracting the true mean (we don't know this, but don't worry), and scaling by the square root of the estimated variance divided by n .¹⁰ The result, $\frac{\bar{x} - \mu}{s/\sqrt{n}}$, has the sample distribution of the random variables

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{(\hat{\mu} - \mu)}{\sigma/\sqrt{n}} \left[\frac{\hat{\sigma}^2}{\sigma^2} \right]^{\frac{1}{2}}$$

The first part of this is distributed as a normal random variable with zero mean and unit variance. The second part (inside brackets) is distributed as n^{-1} times the square root of χ_{n-1}^2 . Hence the product is distributed as (Chapter 3, Section 10), Student's t distribution with n degrees of freedom, so we may

¹⁰ This is sometimes called Studentizing.

use the cdf of this distribution to come up with exact confidence limits for the sample mean, using only the sample mean and the sample variance. (This is, of course, why the distribution was invented). The t distribution is symmetric, so the confidence limits are also. We can write

$$\mathcal{P}[\Phi_{-1}(\alpha/2) \leq \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \leq \Phi^{-1}(1 - \alpha/2)] = 1 - \alpha$$

whence we get the confidence interval expression as being

$$\mathcal{P}[\hat{\mu} + (\hat{\sigma}/\sqrt{n})\Phi_{-1}(\alpha/2) \leq \mu \leq \hat{\mu} + (\hat{\sigma}/\sqrt{n})\Phi^{-1}(1 - \alpha/2)] = 1 - \alpha$$

and which we may interpret, for the actual estimates, as a statement that the confidence limits are $\pm s\Phi^{-1}(1 - \alpha/2)/\sqrt{n}$. The inverse cdf values are what are usually given in a “Table of the t Distribution”.

Since finding confidence intervals requires a determination of the pdf of the sampling distribution, they are more difficult to determine than the standard error; of course, if we can demonstrate (or choose to assume) normality for the sampling distribution (as in the example above) finding these is equivalent.

We finish with two important points about how to present limits and errors. It is conventional to always assume that the \pm values in tables refer to standard errors; if you mean them to be confidence limits, it is important to say so. There is no set convention for what to show as error bars in a plot; we suggest that this should be 95% confidence, since the purpose of error bars is to give a sense of where the true value might reasonably fall. Whatever you choose, you should always state what these bars mean.

5.4. Desirable Properties for Estimators

We turn now to general considerations of the properties of different estimators, and discuss some that are desirable.

5.4.1. Unbiasedness

We say that an estimator that produces $\hat{\theta}$ is an **unbiased estimator**¹¹ for θ if the expectation of $\hat{\theta}$ is the true value θ : that is,

$$\mathcal{E}(\hat{\theta}) = \theta \quad \text{for all } \theta$$

If this is not so, then the estimator is said to be **biased**; the bias of $\hat{\theta}$ is

$$b(\hat{\theta}) = \mathcal{E}(\hat{\theta}) - \theta$$

In general, the pdf of $\hat{\theta}$, and hence its expected value, and the bias $b(\hat{\theta})$ will depend on n , the number of observations in the sample. If $b(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ then $\hat{\theta}$ is **asymptotically unbiased**, the “asymptotically” referring to what happens as we move towards there being an infinite number of observations. A large part of statistics is devoted to showing asymptotic properties of estimators (and other things, as we shall see). This focus on very large amounts of data is not because we always have this, but because asymptotic properties are sort of a minimum expectation: we would at least like whatever we do to get better as we get more data. If this is not true, as for example an estimator being asymptotically unbiased, it is usually a serious defect. But biased estimators may still be useful provided they are

¹¹ **Terminology alert:** strictly speaking we should say “unbiased statistic,” but we blur (as is common in the statistical literature) the distinction between an estimator and the statistic it produces.

asymptotically unbiased.

Considering some of the estimators described in Section 5.1, \bar{x} is an unbiased estimator for μ , since

$$\mathcal{E}[\bar{x}] = \mathcal{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathcal{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

but the same is not true for s^2 as we defined it earlier:

$$\begin{aligned} \mathcal{E}[s^2] &= \frac{1}{n} \mathcal{E}\left[\sum_{i=1}^n (X_i - \bar{x})^2\right] \\ &= \frac{1}{n} \mathcal{E}\left[\sum_{i=1}^n (X_i^2 - 2X_i\bar{x} + \bar{x}^2)\right] \\ &= \frac{1}{n} \mathcal{E}\left[\sum_{i=1}^n (X_i^2 - 2\mu X_i + \mu^2 + \bar{x}^2 + 2\mu X_i - 2X_i\bar{x} - \mu^2)\right] \\ &= \frac{1}{n} \mathcal{E}\left[\sum_i (X_i - \mu)^2 + n\bar{x}^2 + 2\mu \sum_i X_i - 2 \sum_i X_i\bar{x} - n\mu^2\right] \\ &= \frac{1}{n} \mathcal{E}\left[\sum_i (X_i - \mu)^2\right] - \mathcal{E}[\bar{x}^2 - 2\mu\bar{x} + \mu^2] \end{aligned}$$

Thus

$$\begin{aligned} \mathcal{E}[s^2] &= \mathcal{E}\left[\frac{1}{n} \sum_i (X_i - \mu)^2\right] - \mathcal{E}[(\bar{x} - \mu)^2] = \frac{1}{n} \mathcal{V}[\sum_i (X_i - \mu)] - \mathcal{V}[\bar{x} - \mu] \\ &= \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \left(1 - \frac{1}{n}\right) \end{aligned}$$

So s^2 will always be biased (low) relative to σ^2 , although it is asymptotically unbiased. The unbiased estimator is then easily seen to be

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

Obviously, the bias is small unless n is small (in which case this is not going to be a well-determined result anyway).

5.4.2. Relative Efficiency

An **efficient estimator** is one that produces an **efficient statistic** $\hat{\theta}_A$, relative to some other statistic $\hat{\theta}_B$. The measure of efficiency is the ratio of variances of the sampling distributions of the two statistics. Assuming that $\hat{\theta}_A$ and another one $\hat{\theta}_B$ are both unbiased, the relative efficiency is

$$\text{Relative efficiency} = \left[\frac{\mathcal{V}(\hat{\theta}_A)}{\mathcal{V}(\hat{\theta}_B)} \times 100 \right] \%$$

Clearly we should prefer the more efficient estimator $\hat{\theta}_A$, since the probability that $\hat{\theta}_A$ lies in some interval around the true value θ (say, in $[\theta - \varepsilon, \theta + \varepsilon]$) will be higher than the probability of finding $\hat{\theta}_B$ (which is more spread out) in the same interval. It turns out that under general conditions there is a lower bound to the variance of the sampling distribution of any unbiased estimate of a parameter θ . This lower bound

is given by the Cramer-Rao inequality, which we discuss below. If the variance of $\hat{\theta}$ achieves the Cramer-Rao bound then the estimator that produces it is called a **fully efficient estimator** for θ , or alternatively a **Minimum Variance Unbiased Estimator (MVUE)**.

If we can model our data by a normal distribution, the sample mean is a more efficient estimate of the mean than the sample median is, as implied by equation (4), and confirmed by the Monte Carlo simulations in Section 5.2.2.

However, this result is very much dependent on the pdf we assume for the random variables X that we use to model the data. To take an extreme case, if their pdf was a Cauchy distribution rather than a normal, the variance of \bar{x} would be infinite, but the variance of the median is not; for a Cauchy pdf

$$\mathcal{V}[x_{med}] = \frac{4\pi^2}{n}$$

making this estimator infinitely more efficient.

In the particular case of the slightly heavy-tailed distribution of the GPS data, the bootstrap evaluations show that, for the particular dataset under consideration, the median is much more efficient than the mean—and the 10% trimmed mean is even better. This suggests, again, that what can be of most practical importance is not full efficiency under restricted assumptions, but something close to full efficiency under less restrictive ones.

5.4.3. Mean Square Error Criterion

If two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are biased, then we cannot use variance alone to measure efficiency. A more general criterion is the **mean square error**, which includes both variance and bias:

$$M^2(\hat{\theta}) = \mathcal{E}[(\hat{\theta} - \theta)^2]$$

This may be expanded as

$$\begin{aligned} M^2(\hat{\theta}) &= \mathcal{E}[(\hat{\theta} - \mathcal{E}(\hat{\theta})) + (\mathcal{E}(\hat{\theta}) - \theta)]^2 \\ &= \mathcal{E}[(\hat{\theta} - \mathcal{E}(\hat{\theta}))^2] + [\mathcal{E}(\hat{\theta}) - \theta]^2 + 2(\mathcal{E}(\hat{\theta}) - \theta)\mathcal{E}[\hat{\theta} - \mathcal{E}(\hat{\theta})] \end{aligned}$$

Since $\mathcal{E}[\hat{\theta} - \mathcal{E}(\hat{\theta})] = 0$ we get

$$M^2(\hat{\theta}) = \mathcal{V}[\hat{\theta}] + b^2(\hat{\theta})$$

M^2 can be used to choose among competing biased estimators; naturally, we choose the one with smaller M^2 . Note that for unbiased estimators M^2 reduces to the variance, so that this then becomes just the relative efficiency.

5.4.4. Consistency

Consistency is yet another “good adjective” we can apply (or not) to an estimator: this is the requirement that an estimator gets closer and closer to the true value as the sample size increases. Such behavior covers (asymptotically) both bias and variance. The formal definition of consistency requires a probabilistic statement:

$\hat{\theta}$ is a consistent estimator for θ if $\hat{\theta} \rightarrow \theta$ in probability as the sample size $n \rightarrow \infty$.

To explain this, we have to explain the idea of **convergence in probability**. Suppose we have a random variable X whose pdf $\phi(x)$ depends on some parameter θ . Then we say that X converges in probability to x as θ approaches some value θ_1 if, for arbitrarily small positive values of the variables ε and η , there exists a value of θ , θ_0 such that

$$\mathcal{P}[|X - x| < \varepsilon] > 1 - \eta \text{ for all } \theta \text{ in } [\theta_0, \theta_1] \quad (9)$$

In this case, the parameter we have is n , the sample size, and the limiting value is ∞ ; so n_0 is a value such that, for all larger values of n , the condition in (9) holds: we can find an n_0 such that, for all larger values of n , the pdf of X is arbitrarily closely concentrated around x . If $\mathcal{V}(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ and $b^2(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$, then $M^2(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ and $\hat{\theta}$ will be consistent. For a normal distribution the sample mean is a consistent estimator: $\bar{x} \rightarrow \mu$ as n increases.

5.5. The Method of Maximum Likelihood

Now that we have established some desirable properties for estimators we need a recipe for constructing them. We will discuss two very important general estimation procedures: maximum likelihood and (in less detail) least squares. We will restrict the discussion almost entirely to the univariate case, so we will discuss estimating one parameter: the location parameter for a univariate distribution. This allows us to focus on the principles. The very important case of estimating many parameters we leave until later.

The method of maximum likelihood was developed by R.A. Fisher in the 1920's and has dominated the field of statistical inference ever since, because it can (in principle) be applied to any type of estimation problem, provided that one can write the joint probability distribution of the random variables which we are assuming model the observations.

Suppose we have n observations x_1, x_2, \dots, x_n which we believe can be modeled as random variables with a univariate pdf $\phi(x, \theta)$; we wish to find the single parameter θ . The joint pdf for n such random variables, $\vec{X} = (X_1, X_2, \dots, X_n)$ is, because of independence,

$$\phi(\vec{X}, \theta) = \phi(X_1, \theta)\phi(X_2, \theta)\dots\phi(X_n, \theta)$$

In probability theory, we think of $\phi(\vec{X}, \theta)$ as a function which describes the X 's for a given θ . But for inference, what we have are the x 's—that is to say, the actual data: so we think of $\phi(\vec{x}, \theta)$ as a function of θ for the given values of \vec{x} . When we do this, ϕ is called the **likelihood function** of θ :

$$\mathcal{L}(\theta) =_{\text{def}} \phi(\vec{x}, \theta)$$

it has values that will be like those of the pdf, but we cannot integrate over them to get probabilities of θ , for the simple (but subtle) reason that θ is, we remind you again, *not* a random variable.

The use of the likelihood function for inference goes as follows: if $\phi(\vec{x}, \theta_1) > \phi(\vec{x}, \theta_2)$ we would say that θ_1 is a more plausible value for θ than θ_2 ; because ϕ is larger, θ_1 makes the observed \vec{x} more probable than θ_2 does. The **maximum likelihood** method is simply to choose the value of θ that maximizes the value of the pdf function for the actual values of \vec{x} ; that is, we find the θ which maximizes $\mathcal{L}(\theta) = \phi(\vec{x}, \theta)$, because this would maximize the probability of getting the data we actually have.

Since ϕ is a pdf, it, and hence $\mathcal{L}(\theta)$, is everywhere positive; so we can take the logarithm of the pdf function. Because the log function (we use natural log, \ln) is single-valued, the maximum of the log of the likelihood function will occur for the same value of θ as for the function itself:

$$\max_{\theta} \mathcal{L}(\theta) = \max_{\theta} \phi(\vec{x}, \theta) = \max_{\theta} \left[\ln[\phi(\vec{x}, \theta)] \right] =_{\text{def}} \max_{\theta} \ell(\theta)$$

We take logs because the log-likelihood function, which we write as $\ell(\theta)$, is often more convenient, since we can replace the product of the ϕ 's by a sum, which is a lot easier to differentiate:

$$\ell(\theta) = \sum_{i=1}^n \ln \left[\phi(x_i, \theta) \right]$$

For future use, note the following relationships between derivatives:

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \theta} \quad (10)$$

As an initial example, consider finding the rate of a Poisson process from the times between events. As described in Chapter 3, the interevent times have a pdf $\phi(x) = \lambda e^{-\lambda x}$. Then the log-likelihood is

$$\ell(\lambda) = \sum_{i=1}^n \ln(\lambda) - \lambda x_i = n \ln(\lambda) - \lambda \sum_{i=1}^n x_i = n \ln(\lambda) - \lambda x_s$$

where x_s is the sum of all the interevent times—which is to say, the time between the first and the last event. Taking the derivative with respect to λ and setting it to zero gives the maximum likelihood estimate (MLE) as

$$\hat{\lambda} = \frac{n}{x_s}$$

Note that this implies that all we need to get our estimate is the total span x_s and number of events n ; the ratio of these is called a **sufficient statistic** because it is sufficient to completely specify the pdf of the data: no additional information or combination of data can tell us more. Obviously, establishing sufficiency is valuable, since it tells us that we have enough information.

For another example of an MLE, with equally banal results, consider estimating the mean of a normally-distributed random variable, X . The pdf is

$$\phi(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2}[(x-\mu)/\sigma]^2}$$

which makes the log-likelihood

$$l(\mu) = n \ln \left[\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \right] - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}$$

Taking the derivative

$$\frac{\partial l}{\partial \theta} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \mu$$

which is zero for $\mu = \bar{x}$ (see equation (2)).

These are (intentionally) simple examples; in general the MLE may not have a closed-form solution. But, given a pdf, and a set of data, we can always find it, since numerical methods exist to find the maximum of any function we can compute. But does the MLE have properties that make it a “good” estimator in the sense we described in the previous section? The answer is yes, but to show this we need to take a slight detour, though one that involves the likelihood function.

5.5.1. Cramer-Rao Inequality

In Section 5.4.2 above, we described a desirable characteristic of an estimator as having a high efficiency, meaning that the sample distribution of the associated statistic had a small variance. We now show that there is a lower bound to this variance; one that certain maximum likelihood estimators reach, making them **fully efficient**. This minimum variance is set, as we noted in that section, by the **Cramer-Rao bound**.

Consider a function $\tau(\theta)$, and let $\hat{\tau}$ be an unbiased statistic for τ . Then, by the definition of bias,

$$E[\hat{\tau}] = \tau(\theta) = \int \cdots \int \hat{\tau} \mathcal{L}(\vec{X}, \theta) d^n \vec{X}$$

Note that while we are writing the pdf as a likelihood function, we are integrating this function over the random variables. Taking the derivative (and taking it inside the integral, which we shall always assume we can do), we get

$$\frac{\partial \tau}{\partial \theta} = \int \cdots \int \hat{\tau} \frac{\partial \mathcal{L}}{\partial \theta} d^n \vec{X} \quad (11)$$

where we have made use of (10). Now, since \mathcal{L} is a pdf, the integral

$$\int \cdots \int \mathcal{L} d^n \vec{X} = 1$$

and hence the derivative of this is zero:

$$\int \cdots \int \frac{\partial \mathcal{L}}{\partial \theta} d^n \vec{X} = 0$$

which, when we multiply it by $\tau(\theta)$ and subtract it from (11), gives us

$$\frac{\partial \tau}{\partial \theta} = \int \cdots \int (\hat{\tau} - \tau(\theta)) \frac{\partial \mathcal{L}}{\partial \theta} d^n \vec{X}$$

We now make use of the Cauchy-Schwarz inequality, which states that for any two functions a and b

$$\left[\int ab \right]^2 \leq \int a^2 \int b^2 \quad (12)$$

with equality occurring only if a and b are proportional to each other. Setting

$$a = \hat{\tau} - \tau(\theta) \quad b = \frac{\partial \ell}{\partial \theta}$$

then gives us

$$\left[\frac{\partial \tau}{\partial \theta} \right]^2 =_{\text{def}} [\tau'(\theta)]^2 \leq \left(\int \cdots \int [\hat{\tau} - \tau(\theta)]^2 \mathcal{L} d^n \vec{X} \right) \left(\int \cdots \int \left[\frac{\partial \ell}{\partial \theta} \right]^2 \mathcal{L} d^n \vec{X} \right)$$

But since the integrals are of something times the pdf \mathcal{L} over all X , they just give the expected values, so this inequality becomes

$$[\tau'(\theta)]^2 \leq \mathcal{E}[(\hat{\tau} - \tau(\theta))^2] \cdot \mathcal{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] =_{\text{def}} \mathcal{E}[(\hat{\tau} - \tau(\theta))^2] \cdot I(\theta)$$

where $I(\theta)$, which depends only on the pdf ϕ , is called the **Fisher information**. Noticing that the term multiplying it is the variance of $\hat{\tau}$, we obtain, finally, the Cramer-Rao inequality:

$$\mathcal{V}[\hat{\tau}] \geq \frac{[\tau'(\theta)]^2}{I(\theta)}$$

If we look at the MLE for the mean of a normal pdf, above, we see that

$$\frac{\partial \ell}{\partial \mu} = - \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$

but also, $\tau(\theta) = \mu$, so $\hat{\tau} - \tau(\theta) = \hat{\mu} - \mu$. But then it is just the case that the two elements of the Cauchy-Schwartz inequality ((12)) are proportional to each other, and the inequality becomes an equality. The sample mean thus reaches the Cramer-Rao bound, so that the statistic \bar{x} is an unbiased and fully efficient statistic for μ . Also

$$\mathcal{V}[\bar{x}] = \frac{1}{I(\theta)} = \frac{\sigma^2}{n}$$

5.5.2. Some Properties of MLE's

Besides the philosophical appeal of choosing $\hat{\theta}$ corresponding to the highest probability, maximum likelihood estimators are asymptotically fully efficient. That is, for large sample sizes they yield the estimate for θ that has the minimum possible variance. They can also be shown to be asymptotically unbiased, and consistent—so they have most of the desirable properties we want in an estimator. Note that these asymptotic results do *not* necessarily apply for n finite.

It can also be shown (though we do not give the proof) that for large n , the MLE becomes normally distributed; more precisely, under certain smoothness conditions on ϕ , the variable

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$$

is distributed with a normal distribution with zero mean and unit variance; θ_0 is the

true value of θ , which of course we do not know. Since the MLE is unbiased for large n , it is safe (and standard) to use $\hat{\theta}$ as the argument for I . The confidence limits then become those for a normal distribution; for example, the 95% limits become

$$\pm \frac{1.96}{\sqrt{nI(\hat{\theta})}}$$

5.5.3. Multiparameter Maximum Likelihood

We touch briefly on the extension of the maximum likelihood method to estimating more than one parameter. We have the likelihood function

$$\phi(\vec{x}, \theta_1, \theta_2, \dots, \theta_p)$$

which is now a function of p variables; the log-likelihood is

$$l(\theta_1, \theta_2, \dots, \theta_p) = \ln \left[\phi(\vec{x}, \theta_1, \theta_2, \dots, \theta_p) \right]$$

The maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$ are obtained by maximizing l with respect to each of the variables, i.e., by solution of

$$\frac{\partial l}{\partial \theta_1} = 0, \quad \frac{\partial l}{\partial \theta_2} = 0, \quad \dots \quad \frac{\partial l}{\partial \theta_p} = 0$$

For example, suppose we want to estimate both μ and σ^2 from x_1, x_2, \dots, x_n , assuming that these can be modeled by random variables with a normal distribution. As before,

$$\phi(\vec{x}, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

yielding

$$l(\vec{x}, \mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (13)$$

If we now find the maximum for the mean, we have

$$\frac{\partial l}{\partial \mu} = 0$$

which implies

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu}) = 0$$

giving the maximum likelihood estimate for μ , which is again the sample mean:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

For the variance, the derivative is

$$\frac{\partial l}{\partial \sigma^2} = 0$$

which implies, from (13), that

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu})^2 = 0$$

giving the result for the MLE of

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

So the maximum likelihood estimates of μ and σ^2 are the sample mean and sample variance, equations (2) and (3). Note though that the sample variance $\hat{\sigma}^2$ is the biased version, i.e., it uses n instead of $n-1$ in the divisor.

5.5.4. Least Squares and Maximum Likelihood

Estimation of parameters using least squares is perhaps the most widely used technique in geophysical data analysis, having been developed by Gauss and others in the early nineteenth century. How does it fit in with what we have discussed? We will deal with the full least-squares method in a later chapter; for now we simply show how it relates to maximum likelihood. Basically, these are equivalent when we are trying to estimate a mean value of a random variable with a normal distribution.

As a specific problem, consider the estimation problem given by equation (1). We can write the pdf that we use to model the data x_1, x_2, \dots, x_n as

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x - \frac{1}{2}gt_i^2)^2/2\sigma^2} \quad (14)$$

from which the likelihood function for g is

$$\mathcal{L}(g | \vec{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \frac{1}{2}gt_i^2)^2/2\sigma^2}$$

which makes the log-likelihood

$$l(g) = -n \ln[\sqrt{2\pi}\sigma] - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \frac{1}{2}gt_i^2)^2$$

which has a maximum if we minimize the sum of squares of residuals:

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

Taking the derivative of this,

$$\frac{\partial l(g)}{\partial g} = \sum_{i=1}^n t_i^2 (x_i - gt_i^2)$$

and setting it to zero gives the solution

$$\hat{g} = \frac{\sum_{i=1}^n x_i t_i^2}{\sum_{i=1}^n t_i^4}$$

which of course would reduce to the sample mean (2) if all the t_i 's were one.

One further generalization of this is worth noting, namely the case of different errors for each observation, so the pdf of the i -th observation is, instead of (14)

$$\phi_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x - \frac{1}{2}gt_i^2)^2/2\sigma_i^2}$$

from which the likelihood function for g is

$$\mathcal{L}(g | \vec{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x_i - \frac{1}{2}gt_i^2)^2/2\sigma_i^2}$$

which makes the log-likelihood

$$\ell(g) = - \sum_{i=1}^n \ln[\sqrt{2\pi}\sigma_i] - \sum_{i=1}^n \frac{(x_i - \frac{1}{2}gt_i^2)^2}{2\sigma_i^2}$$

Again taking the derivative, we get

$$\frac{\partial \ell(g)}{\partial g} = 2 \sum_{i=1}^n \frac{t_i^2}{2\sigma_i^2} (x_i - \frac{1}{2}gt_i^2)$$

which makes the maximum likelihood estimate equal to

$$\hat{g} = \frac{\sum_{i=1}^n x_i t_i^2 \sigma_i^{-2}}{\sum_{i=1}^n t_i^4 \sigma_i^{-2}}$$

which if all the t_i 's were one becomes the **weighted mean**

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i \sigma_i^{-2}}{\sum_{i=1}^n \sigma_i^{-2}}$$

which is a particular case of a weighted least squares solution, meaning that different observations are weighted differently. As these equations show, this differential weighting may arise from different errors being assigned to different data, the structure of the problem (as the t^2 dependence between g and x) or both. We will explore all this much more completely in a later chapter.

5.5.5. L_1 -norm Estimation

Least squares estimation operates by minimizing the sum of the squares of the difference between a model and some data; this is often called minimization of the **L_2 norm**. We close this section by showing how a different model for errors would correspond to minimizing a different norm, and what estimator this would correspond to in a simple case.

Suppose our model for the data is that the pdf is a double exponential: the X_i 's are each distributed with a pdf

$$\phi(x) = \frac{1}{2} e^{-|x - \mu|}$$

where we have assumed a scale factor of 1. This distribution is, obviously, more

heavy-tailed than the normal. It is immediately clear that the log-likelihood for μ is

$$l(\mu) = -\sum_{i=1}^n |x_i - \mu| \quad (15)$$

so that the MLE for μ will be that value of μ that minimizes the sum of the absolute values of the differences between μ and all the observations x_i ; this is known as minimizing the **L₁ norm**.

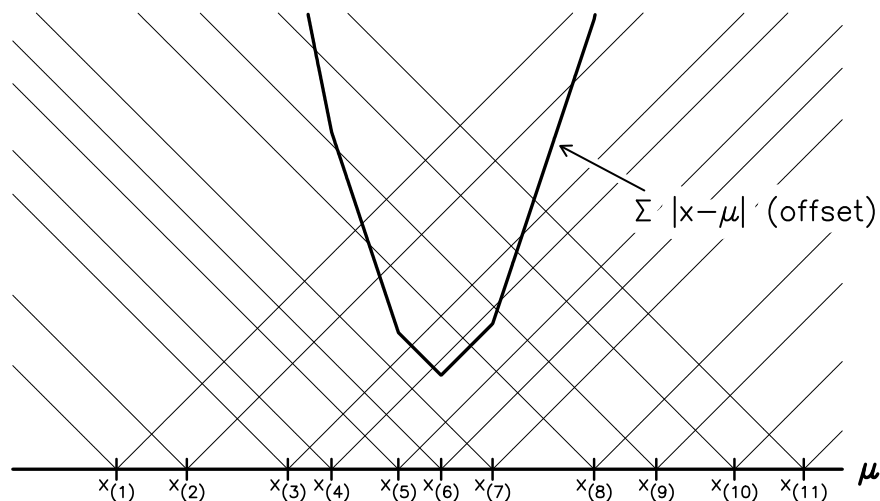


Figure 5.5

If we consider the individual terms in (15), we see that each one, gives rise to a V-shaped function of μ , with the tip of the V being at $\mu = x_i$, with value 0; the slope is -1 below this and 1 above it, as shown in Figure 5.5. The slope of the sum (heavy line in the figures) will thus be the number of x 's below μ , minus the number above μ , so this sum will be a minimum when the slope is zero, or changing from -1 to 1 . This will happen when the number of x 's on each side of μ is the same—so, the L_1 norm is minimized by taking μ to be the median, which is thus the maximum likelihood estimator for the location parameter of a double-exponential distribution. Figure 5.5 shows this for the case of eleven data points, indexed as sorted into increasing order.

Viewing the median as an estimate that minimizes the L_1 norm allows us to generalize it to settings, such as data on the circle and the sphere, in which the idea of sorting does not make sense. The **circular median** minimizes the sum of angles from the median point to all the data; the **spherical median** minimizes the sum of angular distances on the sphere from the median point to all the data. Unlike the median on a line, neither the circular or spherical medians will necessarily coincide with one of the data values.