CHAPTER 5

# PARAMETER ESTIMATION

## 5.1 Introduction

Perhaps the most common data analysis in geophysics is estimating some quantities from a collection of data. We gave an example in Section 1.1.2, where we sought a good value for the distance between two points. Section 4.9 gave another example: we assumed that measurements of distance $x$ for a falling body at different times could be modeled by a random variable

$$Y_i \sim \phi(x - \tfrac{1}{2}gt_i^2) \tag{5.1}$$

with $\phi$ being some known pdf, assumed the same for all the measurements. The **point estimation** problem would then be, given actual data $x_1, x_2, \ldots, x_n$, at times $t_1, t_2, \ldots, t_n$, to find the "best" value of the parameter $g$ – and, though this is a more subtle problem, what range we think it might lie in. *Note* that because data have some definite value, we treat them as conventional variables and represent them by lowercase letters such as $x$. Also note that we use lowercase $n$ for the number of data, in part to distinguish this from $N$, the dimension of a multivariate rv.

## 5.2 A Simple Example: Three Sets of Estimates

We start with the example from Section 1.1.2, We assume that the $n$ data, $x_1, x_2, \ldots, x_n$, can be modeled by a random variable $X$ with a single pdf; it is just because $X$ is random (varying) that we can use one $X$ to model many $x_i$. We take the pdf of $X$ to be one that has only two parameters: the location $l$ and the spread $c$:

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

so our goal is to estimate these two parameters. We start by describing some conventional (and often useful) ways to do so.

### 5.2.1 The Method of Moments

One possible estimate of $l$ is the arithmetic average, or mean, often called the **sample mean**:[1]

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{analogous to} \quad \mu = \int_{-\infty}^{\infty} x\phi(x)\,dx \tag{5.2}$$

where the equation after "analogous to" describes the similar operation, but performed on a pdf – in this case, to get the first moment (expected value). If we use the same analogy for the second moment, we get, as an estimate of the variance, the **sample variance**:

$$s^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i - \bar{x}\right)^2\right] \quad \text{analogous to} \quad \sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2\phi(x)dx \tag{5.3}$$

The expression for $s^2$ contains two summations; while the second one would be identically equal to zero in an exact computation, using it in a computation subject to roundoff error improves the accuracy for large $n$ [*Chan et al.*, 1983].

Equations (5.2) and (5.3) are examples of an estimation procedure called the **method of moments**. We showed in Section 2.12.1 that a pdf (including its parameters) is completely specified if we know all its moments. So, in principle, we can estimate the pdf by estimating the sample moments from the data, which is computationally easy. But in practice we cannot find the moments exactly; unless we are quite sure of the form of the pdf, this procedure can be very misleading. We do not recommend it as a way of finding parameters directly from data, though it can be valuable, as we will see below when we come to the Monte Carlo and bootstrap methods.

The GPS data illustrate the danger of applying this procedure blindly, since equations (5.2) and (5.3) give $\bar{x} = -0.0234$ and $s^2 = 0.407$ ($s = 0.638$).

[1] The 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 many cases in which such a population exists, the population concept is not appropriate for many parts of geophysics. We shall continue to say, instead, that we model data by random variables, and use the idea of a population as little as possible.

If we compare the distribution of the data (Figure 1.3) with the moments we may feel that something has gone badly wrong, at least with the second moment, which seems much too large. There is a simple reason for this: almost all of the data, as shown in Figure 1.1, are between $-1$ and 1; but there is one value at $-10.32$, one at $-1.56$, and one at 6.34. Including these in equation (5.2) does not much affect the sample mean; but these three outlying values increase $s^2$ substantially.

### 5.2.2 Order Statistics

Are there estimation procedures that are less affected by the values of a very small fraction of the data? The answer is yes; we have already alluded to such procedures, the technical term for which is **robust**. A good many robust procedures start by sorting the data into increasing order:

$$x_{(1)} \le x_{(2)} \le x_{(3)} \ldots x_{(n-1)} \le x_{(n)}$$

where the parentheses on the subscripts is the standard notation for a sorted set of values. Procedures that make use of sorted data are called **order statistics**.[2]

One order statistic, which provides an estimate of the location parameter $l$, is the **sample median**

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

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 sample **interquartile range**, which is just $x_{(0.75n)} - x_{(0.25n)}$; again, this is analogous to the definition (Section 2.7) for the pdf. For a normal distribution find that the IQR is $1.349\sigma$. For the GPS data the IQR is 0.14, giving 0.10 for the estimate of the equivalent of $\sigma$. This is a much more satisfying representation of the spread in Figure 1.1.

### 5.2.3 Trimmed Estimates

While the median and IQR are certainly unaffected by a few outlying values, it is fair to ask if these procedures make full use of the data. For

[2] As we discuss in more detail in Section 5.3.1, the word "statistic" has multiple meanings, largely for historical reasons; here it is equivalent to "estimate".

example, in finding the median, the actual values of all but one or two data are irrelevant, except to set the midpoint. So an intermediate approach is sort the data, remove data from the two ends of the sorted set, and compute the mean and variance using what is left: these estimates are called the **trimmed mean** and **trimmed variance**. The 10% trimmed mean and trimmed variance are

$$\bar{x}_{10\%} = \frac{1}{0.9n} \sum_{i=0.05n}^{0.95n} x_{(i)} \quad \text{and} \quad s^2_{10\%} = \frac{1.64}{0.9n} \sum_{i=0.05n}^{0.95n} (x_{(i)} - \bar{x}_{10\%})^2$$

where the constants (0.9 and 1.64) apply for large $n$ and a Normal pdf. For the GPS data $\bar{x}_{10\%} = -0.0125$ and $s^2_{10\%} = 0.112$, which are close to the estimates from the median and IQR.

But, all this is *ad hoc*, and leaves us unable to decide if one method is better than another, either in general, or for a particular case. So we now discuss how to compare different estimators.

## 5.3 Three Ways to Test an Estimator

We have to begin with some terminology. An **estimator** is a procedure for finding the value of some parameter; we have described six such procedures for two parameters. When applied to data, an estimator produces an **estimate**. Using these terms, we would say that our aim is to find methods of evaluating the relative performance of different estimators. We describe three such methods: the classical one, using analytic procedures; Monte Carlo methods; and the bootstrap. (The reasons for these cryptic names will become clearer). Each method has advantages and disadvantages. To make the discussion more concrete, we continue to use the GPS data, and consider estimators for the location and spread of the pdf that we use to model these data.

### 5.3.1 The Classical Method of Evaluating a Statistic

The title of this subsection may seem confusing: 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 statistic[3] is yet a particular kind of random variable:

[3] Blame where blame is due: this unfortunate 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

one that we use in evaluating an estimator.

Suppose that we have $n$ data $x_1, x_2, \dots, x_n$ that we model by $n$ random variables $X_1, X_2, \dots, X_n$, which are described by their joint probability distribution:

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

where the values of the parameters $\theta_1, \theta_2, \dots, \theta_p$ are to be estimated from the data. Note that we use lower-case letters (for example, $x_1$) 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 $\theta$'s given the $x$'s, is quite different from the problems we have studied in probability theory. In those problems, we assume we know the parameters $\theta_1, \theta_2, \dots, \theta_p$ and can 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 instead know the data 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 the parameters are not random variables, so we cannot make probability statements about them.[4]

How do we express what the data tell us about the parameters, when the latter are not random variables? We introduce a random variable that "looks like" a parameter. To keep the discussion as simple as possible, assume that we have only one random variable $X$, and one parameter $\theta$, so the pdf of $X$ is $\phi(\theta)$. We assume that $X$ is a model for all the data $x_1, x_2, \dots,, x_n$; again, there can be $n$ data even if there is only one $X$.

Next, consider some estimator (for example the average, the median, or the trimmed mean); this will supply us, given data, with a value that is our estimate of the parameter $\theta$. Now, imagine applying this estimator (which is just an algorithm), not to the data, but to the random variable we are

English word could be used instead – the difficulty being that such words already had a meaning somewhat at odds with the new one. However, considering such coinages 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 usually being our degree of belief in a particular range of values. This methodology certainly deserves more than a footnote!

using as a model. Performing this algorithm on $n$ random variables $X$ will give us a new random variable, symbolized by $\hat{\theta}$, and called the **statistic** corresponding to $\theta$. The pdf of this is a function of $\hat{\theta}(X_1,X_2,\ldots,X_n)$; this is called the **sampling distribution**[5] of $\hat{\theta}$,[6] The pdf for $\hat{\theta}(X_1,X_2,\ldots,X_n)$ depends on all three of

1. The way the estimate is computed: that is, the estimator.

2. The pdf of $X$, the rv we assume models the data.

3. And the number of data $n$.

Because the statistic is a random variable, but is tied to something we do with data, it is the basis for evaluating different estimation procedures.

And finally, we symbolize the **estimate**, which is the result of applying the estimation procedure to actual data, by $\theta(x_1,x_2,\ldots,x_n)$; this is a conventional variable.

To see how the statistic is used to evaluate an estimate, suppose we have $n$ data, and assume that they can be modeled by an rv $X$ whose pdf is a Normal distribution with mean $\mu$ and variance $\sigma^2$. Note that we do need to know the functional form of the pdf, but do not need to know what the values of $\mu$ and $\sigma$ actually are. If our estimator is given by the average (equation (5.2), we apply it to the $n$ rv's $X$ to produce another random variable, the statistic $\hat{\mu}$:

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

All the $X_i$'s, by assumption, have the same pdf and are independent. Then, by the results of Section 2.12, the pdf of the rv[7] $\hat{\mu}$ is the convolution of $n$ normal pdf's, divided by $n$: which is another Normal pdf with expected value $\mu$ and variance $\sigma^2/n$, So, if we are justified in our assumption that we can model the data by random variables with the normal distribution, we have found 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.

Suppose instead we used the median as our estimator. If we apply the procedure for finding the median to the $n$ rv's $X_i$ (again with a Normal

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

[6] *Notation alert* We use $\hat{\theta}$ to denote both a random variable and the function that produces its pdf; what sets these apart is their arguments.

[7] That is, the sampling distribution of this statistic.

distribution) we get a different random variable, $\hat{\mu}_{med}$. It can be shown the expected value of $\hat{\mu}_{med}$ is also $\mu$, while the variance is approximately

$$\mathcal{V}[\hat{\mu}_{med}] = \frac{\pi\sigma^2}{2(n+2)}\left[1+\frac{\pi}{2(n+4)}\right] \tag{5.4}$$

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

You might think that both estimators look satisfactory, since for both the expected value of the statistic is the actual value of the parameter, while both variances go to zero as $n$ goes to infinity. This view is correct, but before exploring these properties further in Section 5.5 we describe some other ways to determine the behavior of an estimator by finding the behavior of the associated statistic.

### 5.3.2 Monte Carlo Methods

A terse description of the method of the previous section would be, for $n$ random variables with known pdf's (assumed to model the data), find the pdf of some combination of them corresponding to an estimator. Doing this analytically, as we did for the average, gives the most rigorous result; but if we cannot, we can replace analysis with computation, using what is called a **Monte Carlo** evaluation. The term comes from the famous gambling center. There are actually lots of Monte Carlo methods; 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. We have taken a more restricted meaning here.

The basic Monte Carlo procedure is simple:

1. Generate $n$ random variables with the assumed pdf.

2. Compute an estimate (call it $\hat{\theta}_1$) from these imitation data.

3. Repeat steps 1 and 2 a total of $k$ times, to get estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$.

4. Find the first and second moments of the $\hat{\theta}$'s:

$$\bar{\theta} = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i \qquad \hat{\sigma}_{\theta}^2 = \frac{1}{k}\sum_{i=1}^{k}(\hat{\theta}_i - \bar{\theta})^2$$

and take these to be the first and second moments of the pdf of the statistic $\hat{\theta}$. 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. Note that we can be certain that our imitation data $X$ have the distribution we want – after all, we computed them.

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. But rather than trying to find the distribution of this statistic analytically, we do it through 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?

1. The assumptions involved are the same: we assume the pdf of the rv's used to model the data, an assumption that may or may not be valid. (In the next section we see how to eliminate this assumption).

2. Monte Carlo evaluation demands many fewer analytical skills. Given an algorithm for generating appropriate random variables, and another for finding the estimates from data, we can proceed – and we must have the latter, else how could we compute an estimate? Given a novel estimation procedure Monte Carlo methods may often be the quickest way to find out the sampling distribution (the pdf) of the statistic.

3. However, replacing the analytical determination of the pdf by step (4), finding it empirically, can be a disadvantage: we may need to have $k$ very large to determine this pdf well. If we only want the expected value and variance (the first two moments) a relatively small value of $k$ will usually suffice. But sometimes we want to reliably determine, for example, the limits within which 95% of the sampling distribution lies; since this is making a statement about the tails of the pdf, we need a large value of $k$ to do so reliably.

But the biggest problem with Monte Carlo methods is that while they can answer the question, "How good is this estimator?", they cannot answer the question, "Is this estimator the best possible (**optimal**)?" We still need analytical methods to answer this second question.

We certainly suggest that you consider using Monte Carlo methods – but they should, in general, be your second choice, after some attempt to find an analytical result for the sampling distribution of the statistic. If

[8] And here the idea of sampling from a population makes perfect sense.

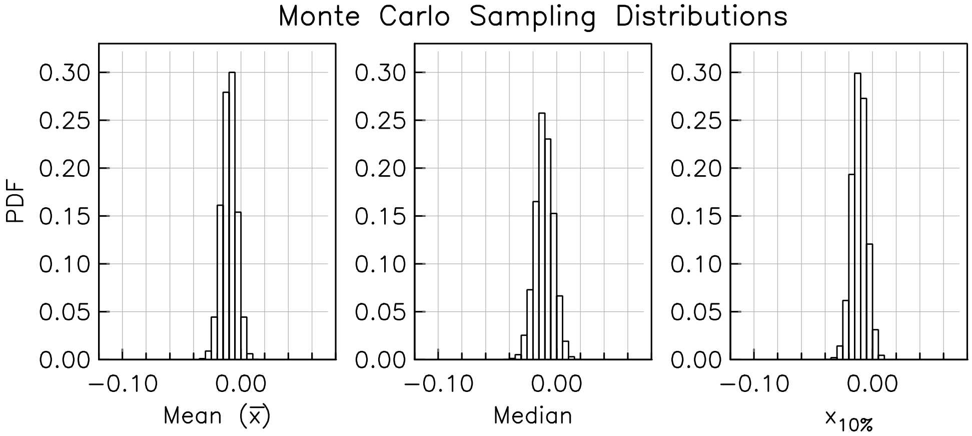

Figure 5.1: Histograms of the statistic $\hat{\mu}$, the estimate of the mean, for Monte Carlo simulations of (left to right) the average, the median, and the 10% trimmed mean. The simulated rv's have a mean of −0.01 and a variance of 0.0144; remember, that what is being shown here is not that (the distribution of the data), but of the statistic, $\hat{\mu}$, for $n = 386$.

you have derived your own analytical result, we strongly suggest doing a Monte Carlo computation as a check!

Figure 5.1 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, the variance of which is nearly the same as the $\sigma^2/n$ of the sample mean.

### 5.3.3 Lifting Ourselves by Our Bootstraps

Our final procedure for evaluating an estimator is also computationally intensive, and very like the Monte Carlo procedure, 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 – so 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.

To find the sampling distribution of a statistic using the bootstrap we do the following:

1. Generate $n$ random variables, whose distribution is uniform over the integers from 1 to $n$; denote these variables by $j_1, j_2, \ldots, j_n$. Note that some of these $n$ variables will, almost certainly, be identical – since otherwise we would have managed to select $n$ values from a uniform distribution that happen to include each possible value once and once only.

2. 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).

3. Compute an estimate (call it $\hat{\theta}_1$) from these imitation random variables.

4. Repeat steps 1 through 3 a total of $k$ times, to get estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$.

5. Find the first and second moments of the $\hat{\theta}$'s, or whatever other property we are interested in.

The parallel with the Monte Carlo method should be clear; like that, the bootstrap can be done only if we can perform significant amounts of computation – for which reason these methods were developed relatively recently. The method is certainly straightforward, and its 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; But the bootstrap has the same disadvantage as Monte Carlo: 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 the $k$ we need might may be large enough to exhaust all possible arrangements of the data – in which case the randomness of the selection is less clear.

Figure 5.2 shows the bootstrap applied to the GPS data, with $k = 4000$, for the same three estimators as were used in the Monte Carlo procedure

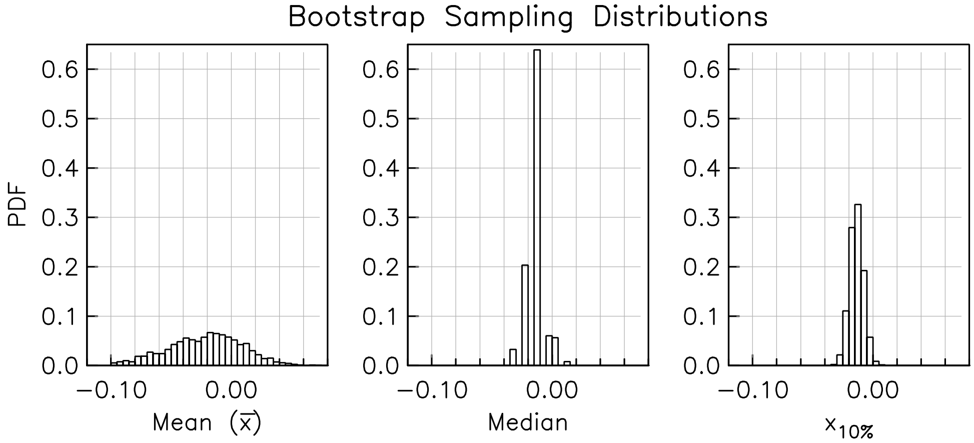

Figure 5.2: The sampling distribution of the statistics corresponding to three estimators, found by applying the bootstrap to the actual GPS data. The histogram for the median looks odd because the original data, one of which will form the median value, are only given to multiples of 0.01; so every other bin is empty.

in Section 5.3.2. The presence of a few outlying values makes a huge difference in the behavior of the different statistics: the mean is much more variable than either the median or the trimmed mean. The lowest variance is for the 10% trimmed mean, which has 0.72 of the variance of the median. The bootstrap method shows clearly that the trimmed mean is the preferred estimator given the actual data distribution. The Monte Carlo simulation in Section 5.3.2 shows that even if the data were Normal, the trimmed mean has a variance nearly the same as that of the mean – so there is no disadvantage to the trimmed mean for data without outliers, and a clear advantage if these are present.

The biggest danger in using the bootstrap is that the data might not in fact 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 for actual data.[9]

What we have described here as "the bootstrap" is actually just one kind of bootstrap method; there are others. For example, we might use the Monte Carlo method of Section 5.3.2; but, instead of assuming *a priori* the

[9] The GPS data are not independent, though not so much so that the results given here are vitiated.

parameters of the pdf for our computed random variates, estimate them from the data[10] This is called a **parametric bootstrap** since we are using the data, though not directly but through a simulation using a known pdf but parameters determined from the data. This approach might be preferable to the regular bootstrap if we had so few data that the number of simulations would be so large as to exhaust the possible combinations of the data.

## 5.4 Confidence Limits for Statistics

So far we have looked only at the first two moments of the sampling distribution of a statistic: that is, at the expected value $\mathscr{E}[\hat{\theta}] = \hat{\theta}$ and the variance $\mathscr{V}[\hat{\theta}]$. If we define $\hat{\sigma}_\theta \overset{\text{def}}{=} \sqrt{\mathscr{V}[\hat{\theta}]}$, then it is conventional to say that our estimate $\theta(\vec{x})$ has a standard error of $\hat{\sigma}_\theta$, and we write the result, again conventionally, as being $\theta \pm \hat{\sigma}_\theta$.

This is often interpreted as being a statement about the range within which $\theta$, the true value, is likely to fall – but if we look at this statement carefully, we can see that this is nonsense. Any statement that includes something like "is likely" is a probabilistic one; but we cannot make probabilistic statements about something ($\theta$) that is represented by a conventional variable and not a random one. Furthermore, a statement like "$x \pm \sigma$" is not even a statement of probability, because it contains no information about a pdf besides 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 (a **point estimate** of a parameter), to finding the probable range for the statistic (an **interval estimate**). Note that we said "a statistic" so 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 $\pm 2\hat{\sigma}_\theta$, relative to the center of the pdf. But, the center of the pdf is at $\hat{\theta} = \mathscr{E}[\hat{\theta}(\vec{X})]$. Let us further assume that $\mathscr{E}[\hat{\theta}] = \theta$ (that is, the statistic is unbiased – we discuss this in more detail in the next section). Then the probability statement we can make is

$$\Pr[\theta - 2\hat{\sigma}_\theta \le \hat{\theta} \le \theta + 2\hat{\sigma}_\theta] = 0.95 \tag{5.5}$$

[10] Actually, we did this in our Monte Carlo example.

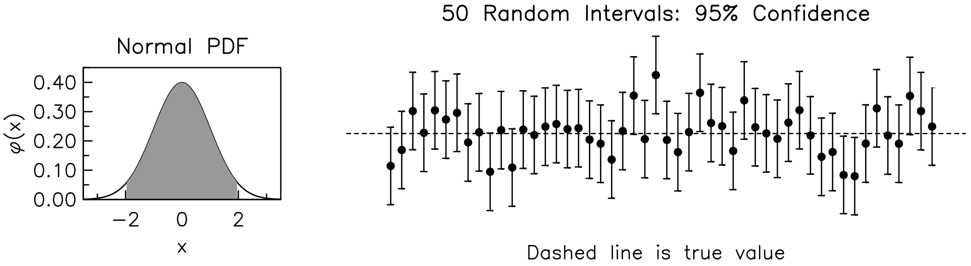

Figure 5.3: On the left, the shaded area under the pdf is 0.95 of the probability, and so defines the length of the 95% confidence interval. The right-hand plot shows the intervals we might get from 50 different data sets: 95% of them include the true value.

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

But, we can make a probabilistic statement, by instead considering the **random interval** $[\hat{\theta}-2\hat{\sigma}_\theta, \hat{\theta}+2\hat{\sigma}_\theta]$, which is called a 95% **confidence interval**. Then equation 5.5 implies, through the distribution of $\hat{\theta}$, that this interval will cover (that is include) the true value 95% of the time; we write this as

$$\Pr[\hat{\theta}-2\hat{\sigma}_\theta \le \theta \le \hat{\theta}+2\hat{\sigma}_\theta] = 0.95 \tag{5.6}$$

You might think that in this equation we have just done what we said we shouldn't: making a probabilistic statement about $\theta$, 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 like equation (5.6) 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 $\theta$.[11]

Figure 5.3 illustrates the concept of a random interval. 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

[11] We should note one slight ambiguity in the idea of a confidence interval: the interval is not uniquely determined by the requirement that it contain a certain fraction of the probability, since this is possible for different limits – so usually the interval is taken to be the shortest one that satisfies the probability requirement.

interval at the same level of confidence (0.95, or 95%), along with the true value. We see that three of the intervals do not cover this value; we would expect two to three. Given a single dataset from which we find a single estimate and confidence interval, it is relatively improbable that the interval would not include the true value: so we have high confidence that the true value lies within the interval,

Having explained what a confidence interval is, we now explain how to find it. A random interval is defined by two values for the random variable, $\hat{\theta}$, $\hat{\theta}_1 < \hat{\theta}_2$, such that, for the given pdf of $\hat{\theta}$,

$$\Pr[\hat{\theta}_1 \le \theta \le \hat{\theta}_2] = 1 - \alpha$$

For a 95% interval, $\alpha = 0.05$: in general a smaller $\alpha$ means a longer interval. Again, this is a statement about the random interval $[\hat{\theta}_1, \hat{\theta}_2]$ compared to the true value $\theta$, and one that can be broken down into two statements of the reverse case, namely that $\hat{\theta}$ lies outside an interval:

$$\Pr[\theta \le \hat{\theta}_1] = \alpha_1 \qquad \Pr[\theta \ge \hat{\theta}_2] = \alpha_2 \quad \text{with} \quad \alpha_1 + \alpha_2 = \alpha \tag{5.7}$$

Usually we take $\alpha_1 = \alpha_2$, and choose the shortest interval consistent with the probability statement 5.7. Let the pdf for $\hat{\theta}$, be $\phi_{\hat{\theta}}$, with cumulative distribution function $\Phi_{\hat{\theta}}$; if the pdf is symmetric about the expected value $\mathscr{E}[\hat{\theta}]$, 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 used the result that for a symmetric pdf, the expected value has $\Phi(\mathscr{E}(\hat{\theta})) = 0.50$. (We can add the values since $\Phi^{-1}(\alpha/2) < 0$ for a normal distribution).

### 5.4.1 Confidence Limits for the Mean, Variance Unknown

We now use these results to find the confidence limits for the mean when the variance is not known, but has to be estimated. We have to assume a pdf for our data, and choose the Normal distribution – which, you should by now appreciate, is both a common assumption and a potentially risky one. We assume the pdf of the data has mean $\mu$ and variance $\sigma^2$; we do not know these, but the whole point of our derivation is to make this irrelevant. We use equations (5.2) and (5.3) to estimate the mean $\bar{x}$ and the variance $s^2$. As we described in Section 5.2.1, our assumptions lead to the result that the statistic $\hat{\mu}$ for $\bar{x}$ is distributed as a normal random variable with variance $\sigma^2/n$.

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

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\hat{\mu})+(\hat{\mu}-\mu)]^2 \tag{5.8}$$

is exactly a sum of squares of normal random variables, and so will be distributed as $\chi^2_n$. But if we compute the square in equation (5.8), the cross-product term

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

vanishes by the definition of $\hat{\mu}$, and equation (5.8) becomes

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

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 $\chi^2_1$. The left-hand side is distributed as $\chi^2_n$. The sum of a random variable distributed as $\chi^2_{n-1}$, and an independent variable distributed as $\chi^2_1$, is distributed as $\chi^2_n$; this is somewhat obvious from the construction of $\chi^2$ 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, $\chi^2_{n-1}$.

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 $n$.[12] The result, $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, has the sample distribution of the random variables

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

[12] This is sometimes called Studentizing.

The first part of this is distributed as a normal random variable with zero mean and unit variance. The second part (inside the brackets) is distributed as $n^{-1}$ times the square root of $\chi^2_{n-1}$. By the results of Section 3.5.3, the product has the pdf of Student's $t$ distribution with $n$ degrees of freedom. We may use the cdf of this distribution to produce confidence limits for the sample mean, using only the sample mean and the sample variance.[13] The $t$ distribution is symmetric, so the confidence limits are also. We can write

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

giving the confidence interval expression

$$\Pr\left[\hat{\mu}+(\hat{\sigma}/\sqrt{n})\Phi^{-1}(\alpha/2) \le \mu \le \hat{\mu}+(\hat{\sigma}/\sqrt{n})\Phi^{-1}(1-\alpha/2)\right] = 1-\alpha$$

We may interpret this as a statement that the confidence limits are $\pm\hat{\sigma}\Phi^{-1}(1-\alpha/2)/\sqrt{n}$. The inverse cdf values are what are usually given in a "Table of the $t$ Distribution".

Since finding confidence intervals requires that we know the complete pdf of the sampling distribution, not just the first and second moments, they are more difficult to determine than the variance – though if we can demonstrate (or choose to assume) that the sampling distribution is Normal, the variance and mean determine the pdf, and so give the confidence limits directly.

We conclude with two points about how to present confidence limits in tables and plots. It is conventional to always assume that the ± values in tables refer to "standard errors", that is to say $\sqrt{\sigma^2}$. There is a further implication that the pdf that goes with these numbers is Normal. If your ± values are confidence limits, you should say so. There is no universal convention for what error bars in a plot refer to; so you should always state what your error bars mean. We suggest that if possible you should show 95% confidence limits, since the purpose of the error bars is to indicate where the true value almost certainly is.

## 5.5 Desirable Properties for Estimators

Next we consider how to evaluate different estimators: what properties are especially desirable, and which ones are less important.

[13] This what the Student's $t$ distribution was invented for.

### 5.5.1 Unbiasedness

We say that an estimator is an **unbiased estimator**[14] for $\theta$ if the expectation of the statistic $\hat{\theta}$ is $\theta$: that is, if $\mathscr{E}(\hat{\theta}) = \theta$ for all $\theta$. If this is not so, then the estimator is **biased**, and the bias is $b(\hat{\theta}) = \mathscr{E}(\hat{\theta}) - \theta$. Usually, $n$, the number of data, affects the pdf of $\hat{\theta}$, including its expected value, and hence also affects the bias $b(\hat{\theta})$. If $b(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ then $\hat{\theta}$ would be called **asymptotically unbiased**, where "asymptotically" refers to what happens as the number of data increases to infinity. Many theorems about estimators are about their asymptotic properties; this is not because we necessarily have large amounts of data, but because asymptotic properties are a minimum expectation: we would like our estimate to get better as we get more data. An estimator that was (for example) biased asymptotically would not be a good one, though a biased estimator may be useful so long as it is asymptotically unbiased.

Of the estimators described in Section 5.2, $\bar{x}$ is an unbiased estimator for $\mu$, since

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

[14] *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.

but this is not true for $s^2$:

$$
\begin{aligned}
\mathscr{E}[s^2] &= \frac{1}{n}\mathscr{E}\left[\sum_{i=1}^{n}(X_i-\bar{x})^2\right] \\
&= \frac{1}{n}\mathscr{E}\left[\sum_{i=1}^{n}(X_i^2-2X_i\bar{x}+\bar{x}^2)\right] \\
&= \frac{1}{n}\mathscr{E}\left[\sum_{i=1}^{n}(X_i^2-2\mu X_i+\mu^2+\bar{x}^2+2\mu X_i-2X_i\bar{x}-\mu^2)\right] \\
&= \frac{1}{n}\mathscr{E}\left[\sum_{i}(X_i-\mu)^2+n\bar{x}^2+2\mu\sum_{i}X_i-2\sum_{i}X_i\bar{x}-n\mu^2\right] \\
&= \frac{1}{n}\mathscr{E}\left[\sum_{i}(X_i-\mu)^2\right]-\mathscr{E}[\bar{x}^2-2\mu\bar{x}+\mu^2] \\
&= \left[\frac{1}{n}\sum_{i}(X_i-\mu)^2\right]-\mathscr{E}[(\bar{x}-\mu)^2] \\
&= \frac{1}{n}\mathscr{V}[\sum_{i}(X_i-\mu)]-\mathscr{V}[\bar{x}-\mu] \\
&= \sigma^2-\frac{\sigma^2}{n}=\sigma^2\left(1-\frac{1}{n}\right)
\end{aligned}
\tag{5.9}
$$

So $s^2$ is always biased to be smaller than $\sigma^2$, although it is asymptotically unbiased. It is then obvious that the unbiased estimator for $\sigma^2$ is

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

But equally obviously, the bias in equation (5.9) is small unless $n$ is small – in which case $\hat{s}^2$ will have such a large variance that the bias will not matter much.

### 5.5.2 Relative Efficiency

An **efficient estimator** is one that produces an **efficient statistic** $\hat{\theta}_A$, relative to some other statistic $\hat{\theta}_B$. What "efficiency" means in statistics has nothing to do with its meaning in engineering or physics; rather, the relative efficiency of estimators is given by the ratio of the variances of the sampling distributions of the two statistics. Assuming that the statistics

$\hat{\theta}_A$ and $\hat{\theta}_B$ are both unbiased, the relative efficiency is

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

where the factor of 100 is conventional so that the result can be given as a percentage.

If the relative efficiency is less than 100%, $\mathcal{V}(\hat{\theta}_A) < \mathcal{V}(\hat{\theta}_B)$; and 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 $\theta$ (say, in $[\theta-\epsilon, \theta+\epsilon]$) will be higher than the probability of finding $\hat{\theta}_B$ (which is more spread out) in the same interval.

Can we find estimators that make $\mathcal{V}(\hat{\theta})$ arbitrarily small? No, because there is a lower bound to the variance of the sampling distribution of any unbiased estimate of a parameter $\theta$. We discuss the value of this lower bound (which is given by the Cramer-Rao inequality) in Section 5.6.1. This lower bound allows us to evaluate the absolute efficiency of an estimator; if this is 100%, meaning that the variance of $\hat{\theta}$ achieves this lower bound, the estimator is called a **fully efficient estimator** for $\theta$, or alternatively a **M**inimum **V**ariance **U**nbiased **E**stimator (**MVUE**).

For data modeled by a Normal distribution, equation 5.4 shows that the sample median is a less efficient estimate of the mean than the sample mean is – a result confirmed by the Monte Carlo simulations in Section 5.3.2. However, this result depends very much on the pdf we assume for the random variables that we use to model the data. To take an extreme case, if the appropriate pdf was a Cauchy distribution, the variance of $\bar{x}$ would be infinite, but the variance of the sample median is finite, namely $\mathcal{V}[x_{med}] = (4\pi^2)/n$: so the median is infinitely more efficient.

For the slightly heavy-tailed distribution of the GPS data, the bootstrap evaluations of Section 5.3.3 show that the median is much more efficient than the sample mean – and the 10% trimmed mean is even better. This shows, again, that in practice the most important behavior of an estimator is not full efficiency under restricted assumptions, but nearly full efficiency under less restrictive ones.

### 5.5.3 Mean Square Error Criterion

If two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are biased, variance alone does not provide a useful measure. A more general criterion is **mean square error**, which includes both variance and bias:

$$\begin{aligned}M^2(\hat{\theta}) =&\mathscr{E}[(\hat{\theta}-\theta)^2]\\ =&\mathscr{E}[[(\hat{\theta}-\mathscr{E}(\hat{\theta}))+(\mathscr{E}(\hat{\theta})-\theta)]^2]\\ =&\mathscr{E}[(\hat{\theta}-\mathscr{E}(\hat{\theta}))^2]+[\mathscr{E}(\hat{\theta})-\theta]^2+2(\mathscr{E}(\hat{\theta})-\theta)\mathscr{E}[\hat{\theta}-\mathscr{E}(\hat{\theta})]\end{aligned} \tag{5.10}$$

and since $\mathscr{E}[\hat{\theta}-\mathscr{E}(\hat{\theta})]=0$ this becomes $M^2(\hat{\theta})=\mathscr{V}[\hat{\theta}]+b^2(\hat{\theta})$ Among competing biased estimators; we would choose the one with the smallest $M^2$. For unbiased estimators $M^2(\hat{\theta})=\mathscr{V}(\hat{\theta})$: the mse reduces to the variance, and we are back to the relative efficiency.

### 5.5.4 Consistency

**Consistency** is yet another "good attribute" that we can apply to an estimator. A consistent estimator is one for which, as the sample size increases, the estimator converges to the true value: this covers (asymptotically) both bias and variance. Defining consistency requires a probabilistic statement: $\hat{\theta}$ is a consistent estimator for $\theta$ if $\hat{\theta}\to\theta$ in probability as the sample size $n\to\infty$.

What do we mean by $\hat{\theta}\to\theta$ "in probability"? This is called **convergence in probability**, and is defined as follows. Suppose we have a random variable $X$ whose pdf $\phi(x)$ depends on some parameter $p$. Then we say that "$X$ converges in probability to $x$ as $p$ approaches some value $p_l$" if, for arbitrarily small positive values of the variables $\epsilon$ and $\eta$, there exists a value of $p$ (denoted by $p_0$) such that

$$\Pr[|X-x|<\epsilon]>1-\eta \qquad \text{for all } p \text{ in} \qquad [p_0,p_l] \tag{5.11}$$

In this case, the parameter $p$ is $n$, the sample size, and the limiting value is infinity; so $n_0$ is a value such that, for all larger values of $n$, the pdf of $X$ is arbitrarily closely concentrated around $x$. If $\mathscr{V}(\hat{\theta})\to 0$ as $n\to\infty$ and $b^2(\hat{\theta})\to 0$ as $n\to\infty$, then $M^2(\hat{\theta})\to 0$ as $n\to\infty$ and $\hat{\theta}$ will be consistent. For a normal distribution the sample mean is a consistent estimator: $\bar{x}\to\mu$ as $n$ increases.

Consistency is a stronger property than asymptotic lack of bias or high efficiency, since it involves the entire pdf of the statistic, not just its first two moments. A consistent estimator is guaranteed to give better and better results with more data; an inconsistent one does not, and so should be avoided if possible. Note that all the estimators we have discussed so far have variances that decrease as $n^{-1}$, and so are consistent.

## 5.6 The Method of Maximum Likelihood

We now have desirable properties for estimators, what we need is a recipe for constructing good ones. We discuss two very important estimation procedures that can be applied to many specific problems: maximum likelihood and (in less detail) least squares. We restrict the discussion almost entirely to the univariate case and discuss estimating one parameter, namely the location parameter for a univariate distribution. This simplification allows us to focus on the principles of what we are doing. We leave until later the very important case of estimating many parameters.

The method of maximum likelihood was developed by R. A. Fisher in the 1920's and has dominated the construction of estimators ever since, because (in principle) it can be applied to any type of estimation problem, so long as we can write the joint probability distribution of the random variables which model the observations.

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

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

In probability theory, we think of $\phi(\vec{X},\theta)$ as a function which describes the $X$'s for a given $\theta$. But for estimation, 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 $\theta$ for the given values of $\vec{x}$. When we do this, $\phi$ is called the **likelihood function** of $\theta$:

$$\mathscr{L}(\theta) \overset{\text{def}}{=} \phi(\vec{x},\theta)$$

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

We use the likelihood function to find an estimate in the following way: if $\phi(\vec{x},\theta_1) > \phi(\vec{x},\theta_2)$ we would say that $\theta_1$ is a more plausible value for $\theta$ than $\theta_2$; because $\phi$ is larger: $\theta_1$ makes the observed $\vec{x}$ more likely than $\theta_2$ does. The **maximum likelihood** method is simply to choose the value of $\theta$ that maximizes the value of the $\phi$ for the actual values of $\vec{x}$; that is, we find the $\theta$ which maximizes $\mathscr{L}(\theta) = \phi(\vec{x},\theta)$, because this would maximize the probability of getting the data we actually have.

Since $\phi$ is a pdf, it, and hence $\mathscr{L}(\theta)$, 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 $\theta$ as for the function itself:

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

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

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

This sum is much easier to differentiate than the product would be. For later use, note the following relationship between derivatives:

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\mathscr{L}} \frac{\partial \mathscr{L}}{\partial \theta} \tag{5.12}$$

As a concrete example, consider finding the rate of a Poisson process from the times between events. As described in Section 3.4.1, the interevent times for a Poisson process 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 $\lambda$ and setting it to zero gives the maximum likelihood estimate (MLE):

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

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 do not need more 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

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

which makes the log-likelihood

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

Taking the derivative gives

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\sigma_2} \sum_{i=1}^{n} (x_i - \mu) \tag{5.13}$$

which is zero for $\mu = \bar{x}$ (equation 5.2).

These are simple examples for which the algebra is easy; but in general the MLE may not have a closed-form solution. However, given a pdf, and a set of data, we can always find the MLE using numerical techniques to find the maximum of the function.

But does the MLE have properties that make it a “good” estimator in the senses we described in Section 5.5.1? Yes, but to show this we need to take a slight detour.

### 5.6.1 Cramer-Rao Inequality

In Section 5.5.2 we described one desirable characteristic of an estimator: 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, called the **Cramer-Rao bound**. Very importantly, this bound is reached by certain maximum likelihood estimators, making them **fully efficient**.

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

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

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

$$\frac{\partial \tau}{\partial \theta} = \int \cdots \int \hat{\tau} \frac{\partial \ell}{\partial \theta} \mathscr{L}(\vec{X}, \theta)\, d^n \vec{X} \tag{5.14}$$

where we have made use of 5.12. Now, since $\mathscr{L}$ is a pdf, the integral

$$\int \cdots \int \mathscr{L}(\vec{X}, \theta)\, d^n \vec{X} = 1$$

and hence the derivative of this is zero:

$$\int \cdots \int \frac{\partial \ell}{\partial \theta} \mathscr{L}(\vec{X}, \theta)\, d^n \vec{X} = 0$$

which, when we multiply it by $\tau(\theta)$ and subtract it from 5.14, gives us

$$\frac{\partial \tau}{\partial \theta} = \int \cdots \int (\hat{\tau} - \tau(\theta)) \frac{\partial \ell}{\partial \theta} \mathscr{L}(\vec{X}, \theta)\, d^n \vec{X} = 0$$

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

$$\left[ \int ab \right]^2 \leq \int a^2 \int b^2 \tag{5.15}$$

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

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

then gives us

$$\begin{aligned} \left[ \frac{\partial \tau}{\partial \theta} \right]^2 &\overset{\text{def}}{=} [\tau \prime(\theta)]^2 \\ &\leq \left( \int \cdots \int [\hat{\tau} - \tau(\theta)]^2 \mathscr{L} d^n \vec{X} \right) \left( \int \cdots \int \left[ \frac{\partial \ell}{\partial \theta} \right]^2 \mathscr{L} d^n \vec{X} \right) \end{aligned}$$

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

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

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

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

Looking at the MLE for the mean of a normal pdf (equation 5.13), 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 (5.15) are proportional to each other, and the inequality becomes an equality. The variance of the sample mean thus reaches the Cramer-Rao bound, so that the statistic $\bar{x}$ is an unbiased and fully efficient statistic for $\mu$. Also

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

### 5.6.2 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 $\theta$ 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 that for large $n$, the MLE becomes normally distributed; more precisely, under certain smoothness conditions on $\phi$, the variable

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

is distributed with a normal distribution with zero mean and unit variance; $\theta_0$ is the true value of $\theta$, 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

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

### 5.6.3 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

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

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

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

The maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$ are obtained by maximizing $\ell$ with respect to each of the variables – that is, by solving

$$\frac{\partial \ell}{\partial \theta_1} = 0 \quad , \quad \frac{\partial \ell}{\partial \theta_2} = 0 \quad , \quad \ldots \frac{\partial \ell}{\partial \theta_p} = 0$$

For example, suppose we want to estimate both $\mu$ and $\sigma^2$ from $x_1, x_2, \ldots, x_n$, 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

$$\ell(\vec{x}, \mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{5.16}$$

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

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

which implies

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

giving the sample mean as the maximum likelihood estimate for $\mu$:

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

For the variance, the derivative is

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

which implies, from 5.16, 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 that the MLE is

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

So the maximum likelihood estimates of $\mu$ and $\sigma^2$ are the sample mean and sample variance, equations 5.2 and 5.3. But note that the sample variance $\hat{\sigma}^2$ is the biased version, since it uses $n$ instead of $n-1$ in the divisor. This is not untypical of MLE estimators: their properties of unbiasedness and full efficiency are guaranteed asymptotically, not for finite $n$.

### 5.6.4 Least Squares and Maximum Likelihood

Estimating parameters using least squares, a procedure developed by Gauss and others in the early nineteenth century, is perhaps the most widely used technique in geophysical data analysis, How does it fit in with what we have discussed? We will describe 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 5.1. We can write the pdf that we use to model the data $x_1, x_2, \ldots, x_n$ as

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\frac{1}{2}gt_i^2)^2/2\sigma^2} \tag{5.17}$$

from which the likelihood function for $g$ is

$$\mathscr{L}(g|\vec{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x_i - \frac{1}{2}gt_i^2)^2}{2\sigma^2}\right)$$

so the log-likelihood is

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

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

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

The derivative of the log-likelihood is

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

and setting this 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 (equation 5.2) if all the $t_i$'s were one.

One further generalization of this is worth noting at this point namely the case of different errors for each observation. Instead of equation (5.17) the pdf of the $i$-th observation is then

$$\phi_i(x) = \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left( \frac{-(x - \frac{1}{2} g t_i^2)^2}{2\sigma_i^2} \right)$$

from which the likelihood function for $g$ is

$$\mathscr{L}(g|\vec{x}) = \prod_{i=1}^{n} = \frac{1}{\sqrt{2\pi\sigma}} \exp\left( \frac{-(x_i - \frac{1}{2} g t_i^2)^2}{2\sigma^2} \right)$$

which makes the log-likelihood

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

Again taking the derivative, we get

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

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.

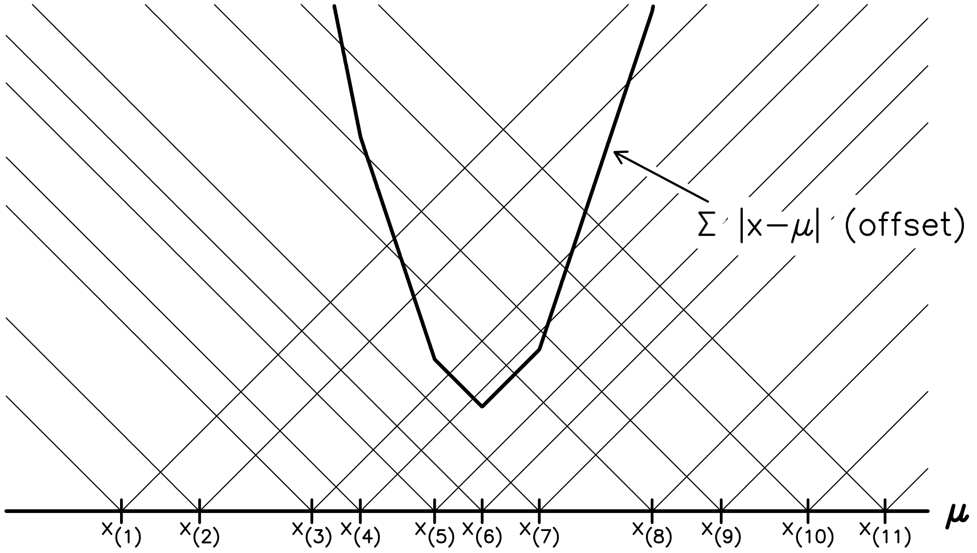

Figure 5.4: Graphical display of how the $L_1$ norm is minimized for finding the location parameter (the median)

### 5.6.5 $L_1$-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 $\mathbf{L_2}$ **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

$$\phi(x) = {}^1\!/\!_2 e^{-|x-\mu|}$$

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 $\mu$ is

$$\ell(\mu) = -\sum_{i=1}^{n} |x_i - \mu| \qquad (5.18)$$

so that the MLE for $\mu$ will be that value of $\mu$ that minimizes the sum of the absolute values of the differences between $\mu$ and all the observations $x_i$; this is known as minimizing the $\mathbf{L_1}$ **norm**.

If we consider the individual terms in 5.18, we see that each one, gives rise to a V-shaped function of $\mu$, 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 5.4. The slope of the sum (heavy line in the figure) will thus be the number of $x$'s below $\mu$, minus the number above $\mu$, 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 $\mu$ is the same – so, the $L_1$ norm is minimized by taking $\mu$ to be the median, which is thus the maximum likelihood estimator for the location parameter of a double-exponential distribution. Figure 5.4 shows this for 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. But both are robust measures of location.