

Chapter 3

Point Estimation

A major theme in statistics is what we call **data reduction**. We may have acquired data in a highly multi-dimensional sample space, but are interested in answering questions of low dimensionality. For example, we may be interested in whether CP violation in the decay $B \rightarrow J\psi K_S^0$ is consistent with the constraints of the Cabibbo-Kobayashi-Maskawa matrix description in the standard model. To answer this questions may require looking at the results of 100 million particle collision events or more, where each event involves the recording of order 10^4 datums from the apparatus. That is a lot of random variables to answer a one-dimensional question.

In this example, we take the measured data for each event (largely pulse heights and times from different elements of the detector) and turn it into more intuitive information, such as momenta, directions, energy deposits, and particle types. This reduces the information in a single event from 10^4 to order 10^2 dimensions. This information may then be used to decide whether the event is of interest to the question or not. If not, it is dropped from the remaining analysis. The data in the events that pass this selection are then used to address the question at hand. This is done by forming some function of the random variables for each event that contains the relevant information, and then combining the information from all of the events. In this way, our many-dimensional sample space is reduced to one or a few dimensions. This is the process of data reduction.

Experiments are generally expensive, both in money and time. So, we want this data reduction process to make effective use of our resources. For example, we would prefer not to waste relevant information as we reduce the data. We'll introduce a variety of criteria for "effectiveness" in our discussion of "point estimation" in this chapter.

3.1 Parametric Statistics

Making a measurement corresponds to taking a sample from some probability distribution, or sampling distribution. The goal of the measurement process is to learn something about this distribution.

Often we know (or assume) something about the form of the sampling distribution. For example, we may know that the mean of the distribution is a quantity of physical interest. In this case, the sampling gives a direct estimate of the quantity we are interested in. We say “estimate” because of the fluctuations – any given measurement will typically have some deviation, or **error**, from the correct value. The quantity of interest is a **parameter of the distribution**. The branch of statistics that deals with estimation of parameters of distributions is called **parametric statistics**.

It may also happen that we do not know, or do not wish to assume, anything about the form of the sampling distribution. It may be the distribution itself that we are trying to measure, for example a spectrum. In such a case, we are dealing with **non-parameteric statistics**. We will concentrate on parametric statistics for now, returning to the non-parametric case in later chapters.

We will develop the subject of hypothesis testing in a later chapter. However, it is a ubiquitous concept, and it is useful to introduce the concept in the present context. We may regard each guess for the true sampling distribution as a **hypothesis**. In the context of parameteric statistics, the possible points in parameter space correspond to hypotheses. That is, the parameter space is the space of possible hypotheses.

We suppose that θ is a physically-interesting parameter, or possibly a vector of parameters. In parametric statistics, we assume that the sampling distribution (PDF) is of the form:

$$f(x) = f(x; \theta), \quad (3.1)$$

where f is a function of x and θ known to whatever level we require in our discussion. Our goal is to estimate θ . This is the problem of **Point Estimation**. We'll introduce the notation of a “hat” accent mark to indicate an estimator. For example, $\hat{\theta}$ is an estimator for the unknown parameter θ . An estimator is a function of our sampled data $\hat{\theta} = \hat{\theta}(x)$, thus an estimator is a random variable with its own sampling distribution, derived from the sampling distribution for x .

A particularly simple form of parameterization is a **location parameter**:

Definition 3.1 *If the probability density function is of the form:*

$$f(x; \theta) = f(x - \theta),$$

*then θ is called a **location parameter** for x .*

If $f(x)$ is plotted as a function of x , then changes in θ will shift $f(x)$ by an amount equal to the change in θ , but the shape of the function is not changed. For example, in the normal distribution:

$$f(x) = \frac{1}{2\pi} \exp \left[-\frac{(x - \theta)^2}{2} \right],$$

θ is a location parameter for random variable X .

Of course, we would like to get the best estimate possible with the data we have. Defining what “best” means, however, is not entirely straightforward. There are several properties we might wish to have in defining what we mean by best:

- An unbiased estimator is one that “gets it right on the average”:

Definition 3.2 The **bias**, b , of an estimator $\hat{\theta}$ for θ is

$$b(\theta) \equiv \langle \hat{\theta} \rangle - \theta. \quad (3.2)$$

Note that the bias is in general a function of θ . An **unbiased** estimator is one for which $b = 0$, independent of θ .

Given a biased estimator, it is sometimes possible to modify it to “correct” for the bias and obtain an unbiased estimator. Note that bias is often thought of as a “systematic error”, particularly when we know a bias may exist but don’t attempt to correct for it.

Example: Given an IID sample of size n from some distribution with finite variance s , we wish to use our data to estimate s . Consider the **sample variance**¹:

$$\hat{s} = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2, \quad (3.3)$$

where m is the sample mean, $m = \frac{1}{n} \sum_{i=1}^n x_i$.

We compute the bias of the sample variance as an estimator for s . We’ll show this in detail, as an instructive sample of such computations. Let θ be the mean of the distribution.

$$\begin{aligned} b(s) &= \langle \hat{s} \rangle - s \\ &= \left\langle \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 \right\rangle - s \\ &= \frac{1}{n} \sum_{i=1}^n \langle (x_i - m)^2 \rangle - s \\ &= \langle (x_1 - m)^2 \rangle - s, \quad \text{since IID} \\ &= \langle (x_1 - \theta + \theta - m)^2 \rangle - s \\ &= \langle (x_1 - \theta)^2 \rangle + \langle (\theta - m)^2 \rangle + 2\langle (x_1 - \theta)(\theta - m) \rangle - s \\ &= s + \left\langle \left[\frac{1}{n} \sum_{i=1}^n (\theta - x_i) \right]^2 \right\rangle + 2\langle (x_1 - \theta)(\theta - \frac{x_1}{n} - \frac{1}{n} \sum_{i=2}^n x_i) \rangle - s \end{aligned}$$

¹There are differing definitions of sample variance, including the quantity \hat{s}' in Eq. 3.6. We’ll adopt the definition given here, but the possible confusion can be avoided by the description “sample second central moment”.

$$\begin{aligned}
&= \frac{1}{n^2} \sum_{i=1}^n \langle (\theta - x_i)^2 \rangle + \frac{2}{n^2} \sum_{i < j} \langle (\theta - x_i)(\theta - x_j) \rangle \\
&\quad + \frac{2}{n} \langle (x_1 - \theta)(\theta - x_1) \rangle + 2 \langle (x_1 - \theta) \frac{1}{n} \sum_{i=2}^n (\theta - x_i) \rangle \\
&= \frac{s}{n} - \frac{2s}{n} + 2 \langle (x_1 - \theta) \rangle \langle \frac{1}{n} \sum_{i=2}^n x_i \rangle, \quad \text{since independent} \\
&= \frac{s}{n} - \frac{2s}{n}, \tag{3.4}
\end{aligned}$$

Thus, the bias of this estimator is $-\frac{s}{n}$.

In this case, we can make a simple modification to obtain an unbiased estimator for s . We note that

$$\langle \widehat{s} \rangle = s \left(1 - \frac{1}{n} \right). \tag{3.5}$$

This, we will have an unbiased estimator for s if we simply multiply \widehat{s} by $n/(n-1)$. That is,

$$\widehat{s}' = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2 \tag{3.6}$$

is an unbiased estimator for the variance.

- A **consistent** estimator gets it right for large statistics:

Definition 3.3 An estimator, $\widehat{\theta}$, is **consistent** if

$$\lim_{n \rightarrow \infty} \widehat{\theta}(x_1, x_2, \dots, x_n) = \theta. \tag{3.7}$$

That is, an estimator is **consistent** if it converges to the parameter in the limit of large statistics.

Note the distinction between consistency and bias. For example, our sample variance,

$$\widehat{s} = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2, \tag{3.8}$$

is a consistent estimator for the variance, as is \widehat{s}' .

- A **sufficient** estimator uses all of the relevant information:

Definition 3.4 A statistic $S = S(X)$ is **sufficient** for parameter θ if the conditional probability for X , given a value of S , is independent of θ :

$$\frac{\partial p(X|S)}{\partial \theta} = 0.$$

Intuition: A sufficient statistic contains all of the information in the data concerning the parameter of interest. Once S is specified, there is no additional information in X concerning θ .

For example, consider sampling n times from the normal distribution $N(\theta, 1)$, with result $x = (x_1, x_2, \dots, x_n)$. Our sampling distribution is:

$$f(x; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2}. \quad (3.9)$$

Let m be the sample mean, and consider:

$$\begin{aligned} \sum_{i=1}^n (x_i - \theta)^2 &= \sum_{i=1}^n (x_i - m + m - \theta)^2 \\ &= \sum_{i=1}^n [(x_i - m)^2 + (m - \theta)^2 + 2(x_i - m)(m - \theta)] \\ &= \sum_{i=1}^n (x_i - m)^2 + n(m - \theta)^2. \end{aligned} \quad (3.10)$$

Thus, we can rewrite our PDF in the form:

$$f(x; \theta) = \left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - m)^2} e^{-\frac{n}{2} (m - \theta)^2}. \quad (3.11)$$

Then, given the sample mean, we have

$$f(x|m; \theta) = \left(\frac{1}{\sqrt{2\pi}} \right)^{n-1} e^{-\frac{1}{2} \sum_{i=1}^n (x_i - m)^2}, \quad (3.12)$$

which is independent of θ . Hence, the sample mean is a sufficient statistic for the mean of this normal distribution.

- A **robust** estimator is insensitive to large fluctuations.

In general, we don't know exactly the probability distribution from which we are sampling when we do an experiment. In particular, there are often extended "tails" above our approximate forms (e.g., non-Gaussian tails on an approximately Gaussian distribution). A robust statistic is one which is relatively insensitive to the existence of these tails.

It may be remarked that there is another sense in which the word robust makes sense. Besides controlling sensitivity to errors in the model, it may be desirable to control for fluctuations within the model itself. While "robust" is usually used in the context of model errors, we'll adopt the broader meaning here and also not attempt a formal definition.

The median of a distribution is typically a more robust estimator for a location parameter than the mean. For example, the Cauchy distribution

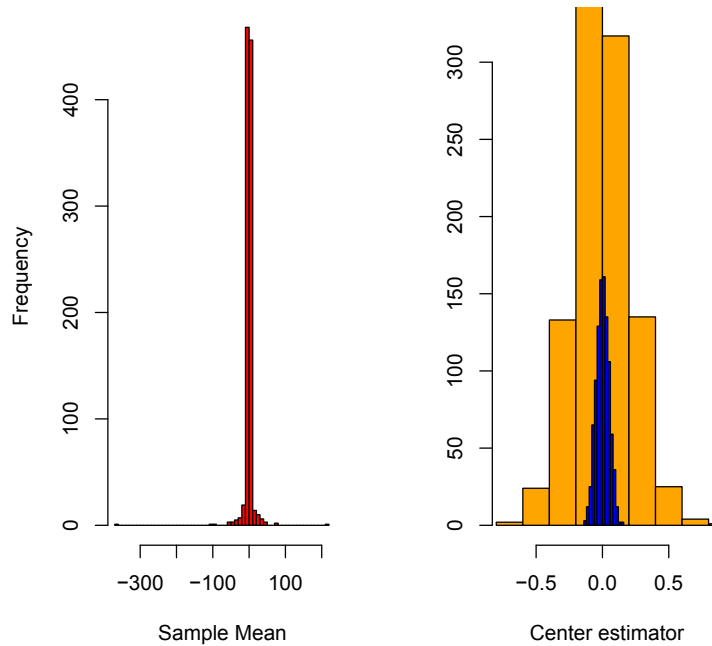


Figure 3.1: Distributions of estimators for the center of a Cauchy distribution. Each estimation (“experiment”) is based on 1000 draws from a Cauchy distribution with center at zero and FWHM equal to two. There are 1000 such experiments simulated. Left: Distribution of the sample mean. Right: The broad histogram (in orange) is the distribution of a “trimmed mean”, in which the upper and lower 1% of samplings are discarded in each experiment. The narrow histogram (blue) is the distribution of the sample median.

is particularly problematic with its long tails. Figure 3.1 shows the performance of three different estimators for the symmetry point (“center”) of a Cauchy distribution. The sample mean may be very far off, due to the high probability of large fluctuations. The **trimmed mean**, in which some of the lowest and highest samplings are discarded before forming a sample mean with the remaining values does much better. However, the sample median does still better.

- An **efficient** estimator has a small variance. Estimator $\hat{\theta}_a$, is said to be more efficient than another estimator $\hat{\theta}_b$, if its variance is smaller.

Note that the goal of good efficiency, by itself, is readily achieved with useless estimators. For example, we spend millions of dollars to measure CP -violation parameter $\sin 2\beta$. The estimator we use has a non-zero variance, improving as the data size increases. However, we could avoid all

this if we only want an efficient estimator. Forget the experiment, and use $\sin 2\hat{\beta} = 0.5$. You can't get more efficient than zero variance!

Later we'll give a useful theorem for how well we can do as a function of bias. As we'll be able to prove, the sample mean is an optimally efficient unbiased estimator for the mean of a normal distribution.

- It might be deemed important to have a “physical” estimator. Here, I simply mean that it may be desirable to have an estimator which is guaranteed to be in some restricted range, corresponding to theoretically allowed values for the parameter of interest. However, if all you are trying to do is summarize the information content of a measurement, as opposed to making some statement about the true value of a parameter, this is not an interesting property to require. This gets back to the discussion in Chapter 3.
- Often an important consideration is “tractableness”, that is an estimator that is practical to obtain given available resources and other constraints. We may be willing to sacrifice other goals to get an answer at all, as long as we can get something “good enough”.

3.2 Information

When we make measurements relevant to some question of interest, such as the value of a physical parameter, we are acquiring relevant information. We may formulate a statistical measure for information.

Definition 3.5 *If $L(\theta; x)$ is a likelihood function depending on parameter θ , the Fisher Information Number, corresponding to θ , is:*

$$I(\theta) = \left\langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right\rangle.$$

An intuitive view is that if L varies rapidly with θ , the experimental sampling distribution will be very sensitive to θ . Hence, a measurement will contain a lot of “information” relevant to θ . It will be useful to note that (exercise for the reader):

$$\left\langle \left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right\rangle = - \left\langle \frac{\partial^2 \ln L}{\partial \theta^2} \right\rangle.$$

The quantity $\partial_\theta \ln L$ is known as the **Score Function**.

There is a great deal of controversy and confusion over the properties of the likelihood function. We will gradually address the issues, starting here with the “likelihood theorem”:

Theorem 3.1 *Let \mathcal{H} be the space of all possible hypotheses, including the truth. Note that we needn't restrict to parametric statistics. Denote a possible hypothesis by $H_i \in \mathcal{H}$. For any given hypothesis H_i , let $P(x|H_i)$ be the probability of event x , and let $P(x'|H_i)$ be the probability of some other event x' . If*

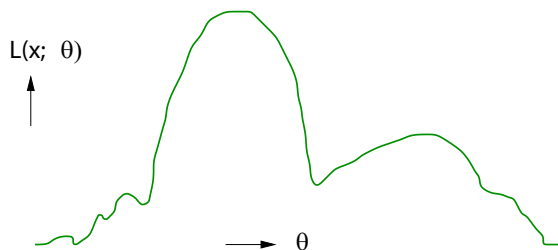


Figure 3.2: A possible likelihood function, used in computing information.

$P(x|H_i) = cP(x'|H_i)$ for all $H_i \in \mathcal{H}$, where $c > 0$ is a constant, then

$$\mathcal{L}(H_i|x) = \mathcal{L}(H_i|x') \quad \forall H_i \in \mathcal{H}. \quad (3.13)$$

The proof of this theorem relies on Bayes theorem, as used in the first and last lines below. Suppose $H_k \in \mathcal{H}$. Then:

$$\mathcal{L}(H_k|x) = \frac{P(x|H_k)P(H_k)}{P(x)} \quad \forall H_k \in \mathcal{H} \quad (3.14)$$

$$= \frac{P(x|H_k)P(H_k)}{\sum_i P(x|H_i)P(H_i)} \quad (3.15)$$

$$= \frac{cP(x'|H_k)P(H_k)}{\sum_i cP(x'|H_i)P(H_i)} \quad (3.16)$$

$$= \mathcal{L}(H_k|x') \quad \forall H_k \in \mathcal{H}. \quad (3.17)$$

Note in this proof that we use the concept of the “probability of a hypothesis”, which may be the “degree of belief” interpretation of Bayesian statistics. However, as long as the probability measure on \mathcal{H} is properly defined, the theorem does not rely on any particular interpretation.

The likelihood theorem tells us that if the probability of two outcomes, x and x' , is in a fixed ratio, independent of model, then both outcomes provide the same likelihood function.

3.2.1 Rao-Cramer-Frechet Inequality

We are ready for an important result which gives us a bound on the best possible efficiency for a given bias. We suppose that we have an estimator $\hat{\theta} = \hat{\theta}(x)$ for a parameter θ , where $x = (x_1, x_2, \dots, x_n)$, with a bias function $b(\theta)$. Let the likelihood function be $L(\theta, \text{other parameters}; x)$.

With this understanding, we have the theorem:

Theorem 3.2 Rao-Cramer-Frechet (RCF) Assume:

1. The range of x is independent of θ .

2. The variance of $\hat{\theta}$ is finite, for any θ .
3. $\partial_{\theta} \int_{-\infty}^{\infty} f(x)L(\theta; x)dx = \int_{-\infty}^{\infty} f(x)\partial_{\theta}L(\theta; x)dx$, where $f(x)$ is any statistic of finite variance.

Then:

$$\sigma_{\hat{\theta}}^2 \geq \frac{[1 + \partial_{\theta}b(\theta)]^2}{I(\theta)}.$$

The proof is left as an exercise, but here is a sketch: First, show that

$$I(\theta) = \text{Var}(\partial_{\theta} \ln L).$$

Next, find the linear correlation parameter, ρ , between the score function and $\hat{\theta}$. Finally, note that $\rho^2 \leq 1$.

3.2.2 Efficient Estimators

This leads to an interesting question: Under what (if any) circumstances can the minimum variance bound be achieved? If an unbiased estimator achieves the minimum variance bound, it is called “efficient”. We have the following:

Theorem 3.3 *An efficient (perhaps biased) estimator for θ exists iff:*

$$\frac{\partial \ln L(\theta; x)}{\partial \theta} = [f(x) - h(\theta)]g(\theta).$$

An unbiased efficient estimator exists iff we further have:

$$h(\theta) = \theta.$$

That is, an efficient estimator exists for members of the exponential family.

The proof is again left to an exercise, but here is a hint: The RCF bound made use of the linear correlation coefficient, in which equality holds iff there is a linear relation:

$$\partial_{\theta} \ln L(\theta; x) = a(\theta)\hat{\theta} + b(\theta).$$

3.3 Maximum Likelihood Method

A popular method for parameter estimation with many desirable properties is the **Maximum Likelihood Method**:

Definition 3.6 *Given measurements x , the **Maximum Likelihood Estimator (MLE)**, $\hat{\theta}$, for a parameter θ , is the value of the parameter for which the likelihood function, $L(\theta; x)$, is maximized:*

$$L(\hat{\theta}; x) = \max_{\theta} L(\theta; x).$$

The intuition behind this is that the MLE is that value of the parameter which would make the actual observed data values the most likely observation (compared with other possible parameter values). This isn't the same as saying that it is somehow the "most likely" value of θ . This would be a statement outside of classical (frequentist) statistics, but is in fact the statement we would make in Bayesian statistics. A complete Bayesian analysis would also multiply by a prior distribution to obtain a posterior distribution. The value of θ for which the posterior is maximal is then the Bayesian estimator for θ .

It is interesting, and not a little confusing, that the likelihood function may be used in both frequentist and Bayesian methodologies. Its suitability for Bayesian analysis is clear, given its role in the use of Bayes theorem for such an analysis. In fact, the maximum likelihood estimator has a number of useful properties making it attractive for a frequentist analysis as well. We shall examine some of these properties, after looking at an example of the MLE method.

3.3.1 MLE – Poisson example

Let us imagine that we are trying to measure the rate for some signal process. We count signal-like events for a period of time. Unfortunately, there is also a background process that is indistinguishable from signal. However, the background rate is known, so we should be able to subtract it off of the total rate. The distribution of background counts in our time interval is Poisson:

$$f_b(n_b; b) = \frac{b^{n_b} e^{-b}}{n_b!}, \quad (3.18)$$

where b is the expected number of background events according to the known background rate. Likewise, the distribution of signal events in our time interval is Poisson:

$$f_s(n_s; \theta) = \frac{\theta^{n_s} e^{-\theta}}{n_s!}, \quad (3.19)$$

where θ is the unknown expected number of signal events, the parameter we wish to estimate using our data. Unfortunately, we cannot distinguish signal and background, hence we cannot separately measure n_b and n_s . We can only measure the sum, $n \equiv n_s + n_b$.

Let us determine the distribution of the sum. Make a transformation from (n_s, n_b) to (n, n_b) :

$$f(n, n_b; \theta, b) = \frac{b^{n_b} e^{-b} \theta^{n-n_b} e^{-\theta}}{n_b! (n-n_b)!} \quad (3.20)$$

Now sum over all possible n_b consistent with a given value of n :

$$f(n; \theta, b) = \sum_{n_b=0}^n \frac{b^{n_b} e^{-b} \theta^{n-n_b} e^{-\theta}}{n_b! (n-n_b)!} \quad (3.21)$$

$$= \sum_{n_b=0}^n \frac{e^{-b-\theta}}{n_b! (n-n_b)!} \theta^{n-n_b} b^{n_b} \quad (3.22)$$

$$= \frac{e^{-b-\theta}}{n!} \sum_{n_b=0}^n \frac{n!}{n_b!(n-n_b)!} \theta^{n-n_b} b^{n_b} \quad (3.23)$$

$$= \frac{(\theta + b)^n e^{-\theta-b}}{n!}. \quad (3.24)$$

We have just demonstrated that the Poisson distribution possesses the **reproductive property** – the sum of two Poisson-distributed random variables is also Poisson-distributed.

Suppose that we do the experiment and observe n events. The likelihood function is:

$$L(\theta; n) = \frac{e^{-\theta-b}(\theta + b)^n}{n!},$$

The MLE for θ is conveniently found by taking:

$$\begin{aligned} \partial_\theta \log L &= \partial_\theta [-\theta - b + n \log(\theta + b) - \log n!] \\ &= -1 + n/(\theta + b). \end{aligned} \quad (3.25)$$

Setting this to zero gives the MLE:

$$\hat{\theta} = n - b,$$

which is intuitive! (You consider $n = 0$ case...)

Note that:

$$\langle \hat{\theta} \rangle = \langle n - b \rangle = (\theta + b) - b = \theta,$$

so this estimator is unbiased. Furthermore,

$$-\langle \frac{\partial^2 \log L}{\partial \theta^2} \rangle = \langle \frac{n}{(\theta + b)^2} \rangle \quad (3.26)$$

$$= 1/(\theta + b). \quad (3.27)$$

Hence, the minimum variance bound is $\theta + b$.

What is the variance of our MLE? It is:

$$\sigma_{\hat{\theta}}^2 = \langle (n - b)^2 \rangle - \langle n - b \rangle^2.$$

Noting that

$$\langle n(n-1) \cdots (n-k) \rangle = (\theta + b)^{k+1},$$

we obtain:

$$\sigma_{\hat{\theta}}^2 = \theta + b,$$

which is the minimum bound. Thus, this MLE estimator is unbiased and efficient, even for small Poisson samples.

Let us examine some of the properties of the oft-misunderstood maximum likelihood estimator.

Theorem 3.4 *The MLE will be unbiased and efficient, if an unbiased efficient estimator exists.*

Proof: The maximum likelihood prescription (assuming no “endpoint” troubles) corresponds to:

$$\left. \frac{\partial \ln L(\mathbf{x}; \theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0.$$

Suppose an efficient, unbiased estimator exists. Then

$$\frac{\partial \ln L}{\partial \theta} = [f(\mathbf{x}) - \theta]g(\theta),$$

and, hence:

$$\left. \frac{\partial \ln L}{\partial \theta} \right|_{\hat{\theta}} = [f(\mathbf{x}) - \hat{\theta}]g(\hat{\theta}) = 0.$$

Thus $\hat{\theta} = f(\mathbf{x})$ is the maximum likelihood estimator. The remainder, to show that it is unbiased and efficient, is left as an exercise. But note that the MLE is otherwise not unbiased and efficient. However, the MLE has very nice asymptotic properties:

Theorem 3.5 *The MLE is asymptotically (i.e., as the sample size $n \rightarrow \infty$) efficient, unbiased, consistent, and normal (assuming the sampling space does not depend on the parameter value).*

To prove this, make a Taylor series expansion of the maximum likelihood condition in terms of $\log L$ about the true parameter value. Use the Central Limit Theorem.

We conclude with a few other remarks about the maximum likelihood estimator:

1. The MLE is parameterization-independent. Given function $\alpha(\theta)$,

$$\hat{\alpha}_{\text{ML}} = \alpha(\hat{\theta}_{\text{ML}}).$$

Note that, since $\langle f(x) \rangle \neq f(\langle x \rangle)$ in general, this means that MLE are “typically” biased.

2. The MLE is sufficient, if a sufficient statistic exists.
3. The MLE may not be robust.
4. Watch out for multiple maxima!
5. This method typically requires a numerical search to find the maximum.

3.3.2 Case Study – Analysis of Bias in m_τ

The BES experiment made a precision measurement of m_τ , by measuring the $e^+e^- \rightarrow \tau^+\tau^-$ cross section near threshold [J. Z. Bai et al. [BES Collaboration], “Measurement of the Mass of the τ Lepton” *Phys. Rev. Lett.* **69** (1992) 3021; J. Z. Bai et al. [BES Collaboration], “Measurement of the Mass of the Tau

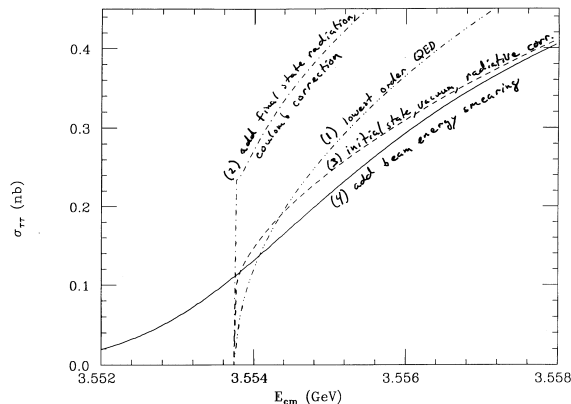


Figure 3.3: The theoretical cross section for $e^+e^- \rightarrow \tau^+\tau^-$ near threshold.

Lepton *Phys. Rev. D* **53** (1996) 20]. To optimize running time, they used a “data-driven” algorithm to update the energy setting of the storage ring in real time, to try to run at the energy where the cross section is most sensitive to the mass, roughly where the derivative is greatest in Fig. 3.3.

The final mass value is obtained by a maximum likelihood fit to the observed cross section as a function of energy. The likelihood function used is:

$$L(m; n) = \prod_{i=1}^k \frac{e^{-\theta_i(m)} \theta_i(m)^{n_i}}{n_i!},$$

where k is the number of energy points, n_i is the number of events observed at energy point i , and $\theta_i(m)$ is the expected number of events at scan point i if the τ mass is m . Thus, this likelihood is maximized as a function of m to obtain the MLE for the tau mass.

There is an important question in this analysis: Is this method of measurement biased? Since it is a precision measurement, even a small bias may be significant. Indeed, the method is in general biased, and it is important to estimate this bias. In order to estimate the bias, we simulate the experiment with Monte Carlo. In the Monte Carlo, we know what mass we put in, so we can compute the bias once we determine the maximum likelihood mass. The bias is determined to desired precision by avergaing together the results of many simulated experiments.

The choice of energy is based on the cleanest channel, $e^+e^- \rightarrow \tau^+\tau^- \rightarrow e^\pm\mu^\mp + \text{neutrinos}$, referred to as the “driving channel”. Other tau decay modes are included later to improve precision. The algorithm to determine the energy setting is, roughly:

1. Start at the best previous measurement of the tau mass.

The $e\mu$ channel is the one which drives the scan. Hence, this is the channel where we might expect to see the greatest sensitivity of the bias to the scan algorithm. As a quick check on the simulation, Figure 1 shows the distribution of the number of events in an experiment. It may be noted that the number we observed in the BES scan is in the region of the peak.

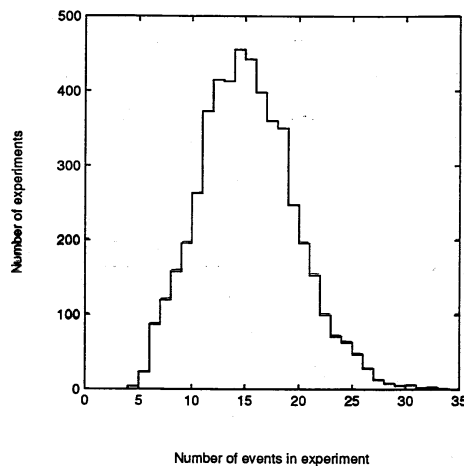


Figure 1. Distribution of the number of $e\mu$ events in an experiment, for a starting error of $\text{start}=7$ MeV, and a maximum step size of 100 MeV.

Figure 3.4: Distribution of the number of $e^+e^- \rightarrow \tau^+\tau^- \rightarrow e^\pm\mu^\mp + \text{neutrinos}$ events in a set of simulated experiments.

2. Run for a fixed amount of integrated luminosity and measure the cross section for the driving channel.
3. Use the measured information to revise the estimate of the tau mass and adjust the energy accordingly.
4. Repeat (2) and (3) until the end of the data-taking run period.

In fact, there are some additional features to ensure some robustness against large fluctuations. For example, the change in energy is limited for any single step.

The distribution of the total number of events in the experiment, according to the simulation, is shown in Fig. 3.4. It may be noted that it is not a priori known how close the starting point is to the true mass, so the simulation has to be performed for different starting energies to determine the effect of this uncertainty.

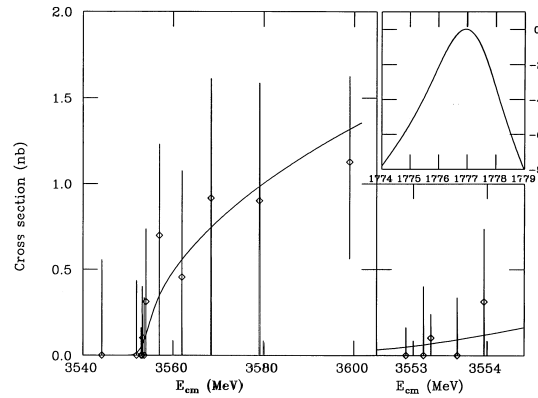


Figure 3.5: The result of the actual experiment. The insert at the upper right shows the likelihood function; the insert at the lower right shows an expanded view of the threshold region.

The result from the real experiment is shown in Fig. 3.5. The small upper inset shows the likelihood as a function of tau mass; the peak position provides the MLE for the tau mass. The other parts show the measured cross section as a function of center-of-mass energy, with curves overlaid for the cross section according to the MLE mass. We will discuss the displayed error bars in the chapter on interval estimation; they are only for display purposes, and are not used in the MLE evaluation.

Figure 3.6 shows results from 5000 simulated experiments, in which the first energy step was 7 MeV above the true mass (as happened in the actual experiment). The vertical axis shows an estimate for the size of the positive error bar. This estimate is based on the likelihood function and will be discussed in the chapter on interval estimation. Here it will suffice to note that this is an estimate for the size of the fluctuations to be expected in the measurement. The horizontal axis shows the error of the measurement, that is, the MLE mass minus the true mass. Most of the 5000 simulated experiments are in the tight clump of points around zero on the horizontal axis. However, there are long tails in the plot, and these are worrisome as they indicate the possibility of large errors in a measurement that is supposed to be precise at the level of tenths of an MeV. We'll refer to experiments falling in these tails as outliers. Of most concern are those with a large deviation from the true value, but a small estimated error, since in these cases the error estimate does not reflect the true error made.

Figure 3.7 shows the likelihood function for one of the outlier experiments with a small estimated error. These outliers have this characteristic likelihood

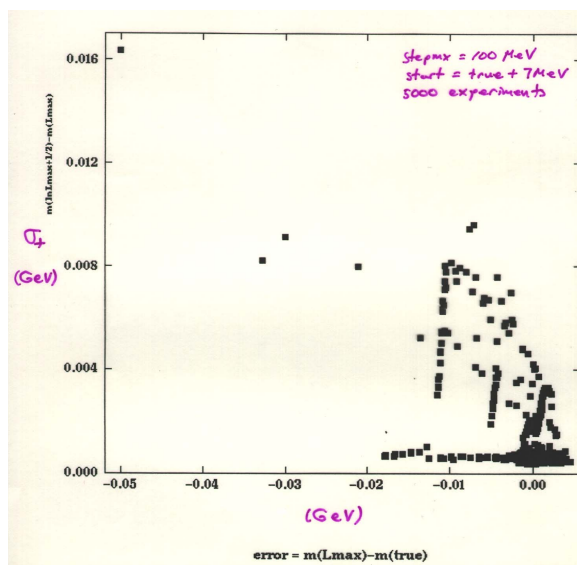


Figure 3.6: Graph of the estimated upper uncertainty on mass vs. the error in the measured mass, from simulations of the experiment. The maximum stepsize is here 100 MeV.

function, with a sharp peak at a low mass, and a broad peak at higher mass. The source of these outliers is traceable to an initial positive fluctuation causing a large step to low mass, then a background event occurring late in the scan. Fortunately, this situation is easily recognized by the form of the likelihood function, and did not occur in the actual experiment. In fact, the maximum stepsize allowed in this simulation is 100 MeV, practically no constraint on the step size. A more realistic maximum step size (the actual maximum is somewhat unclear) is 10 MeV. Figure 3.8 is the same as Fig. 3.6, except for a 10 MeV maximum energy step (note that there is a multiplier of 10^{-4} on the vertical axis. The outlier problem is now essentially gone.

Finally, we can look at the expected bias in the measurement. Figure 3.9 shows the bias as a function of the maximum stepsize. Two curves are shown, one for the driving channel alone, and the other for the measurement combining all channels. For reasonable step sizes, the bias is of the order of a tenth of an MeV, which is small compared with the other uncertainties in the measurement.

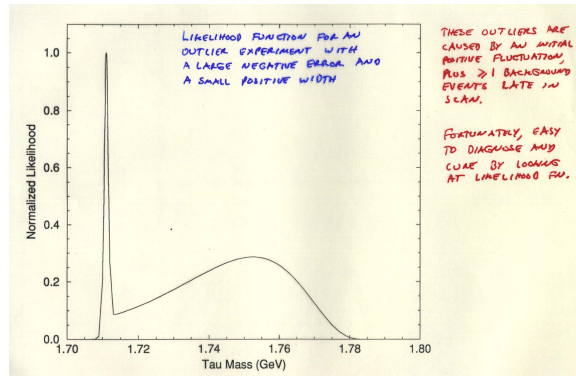


Figure 3.7: The likelihood function for an outlier experiment (from simulation).

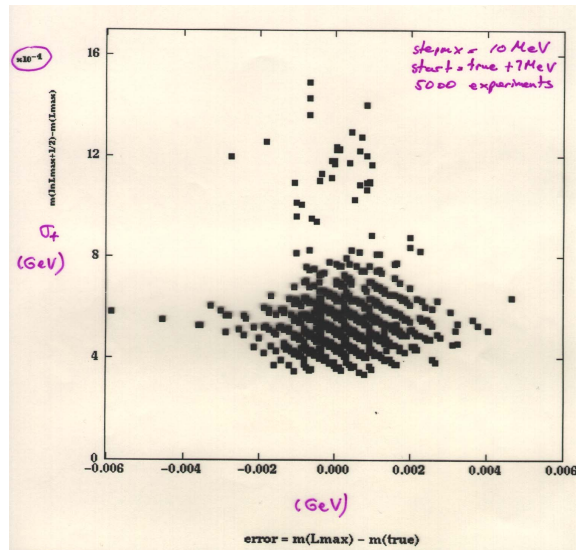


Figure 3.8: Same as 3.6, except for a maximum stepsize of 10 MeV.

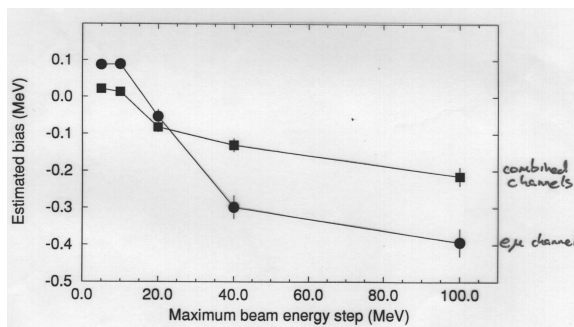


Figure 3.9: Measurement bias as a function of maximum step size.

3.4 Substitution (Moment) Method

Suppose we have PDF $f(x; \theta)$ for random variable X , depending on unknown parameter θ . Any statistic $U(X)$ computed from X has expectation value:

$$\langle U \rangle = \int u(x) f(x; \theta) dx = \phi_U(\theta).$$

If ϕ_U is invertible, we have:

$$\theta = \phi_U^{-1}[\langle U \rangle].$$

Thus, we can define a plausible estimator for θ if we substitute the sample average $m_u = \frac{1}{n} \sum_i^n u(x_i)$ for the expectation value of U :

$$\hat{\theta} = \phi_u^{-1}[m_u].$$

This method is often very easy to apply (and is sometimes applied without even thinking about it), but there is no reason to expect it to be efficient.

A common application of the moment method is in the estimation of parent angular distributions as in a scattering experiment. For example, suppose we want to estimate a in an assumed angular distribution of the form:

$$\frac{d\sigma}{d\Omega} = A(1 + a \cos \vartheta), \quad (3.28)$$

where our measurement consists of the n samplings, $\{x_1, \dots, x_n\}$ of $x = \cos \vartheta$.

Taking $U(X) = X$, we have

$$\langle X \rangle = \int x(1 + ax) dx / 2 = a/3 = \phi_x(a).$$

This is readily inverted, and we obtain the estimator for a :

$$\hat{a} = \frac{3}{n} \sum_{i=1}^n x_i. \quad (3.29)$$

Case study: CP violation via mixing

For a more involved example, consider the measurement of CP violation via mixing at the $\Upsilon(4S)$. This measurement involves measuring the time difference between two B meson decays. The PDF for this time random variable may be written:

$$f(t; A) = \frac{1}{2}e^{-|t|}(1 + A \sin xt),$$

where $t \in (-\infty, \infty)$, $x = \Delta m/\Gamma$ is known, and A is the CP asymmetry parameter of interest.

In the early days, when the experiment was being designed, there was some small dispute concerning the importance of the “dilution factor” in what amounts to a moment method. We have the tools to analyze this now.

The simplified analysis under discussion was to simply count the number of times $t < 0$, n_- , and the number of times $t > 0$, n_+ . The expectation value of the difference between these, for a total sample size $n = n_- + n_+$, is:

$$\langle n_+ - n_- \rangle = n \frac{xA}{1 + x^2}.$$

This is readily inverted, leading to the estimator:

$$\hat{A} = d^{-1} \frac{n_+ - n_-}{n},$$

where $d = x/(1 + x^2)$ is known as the “dilution factor”. We note that \hat{A} is by definition an unbiased estimator for A . The question is, how efficient is it? In particular, we are throwing away detailed time information – does that matter very much, assuming our time resolution isn’t too bad?

First, what is the variance of \hat{A} ? For a given n , we may treat the sampling of n_{\pm} as a binomial process, giving:

$$\delta \hat{A} = d^{-1} \sqrt{(1 - d^2 A^2)/n}.$$

Second, how well can we do, at least in principle, if we do our best? Let’s use the RCF bound to estimate this (and argue that, at least asymptotically, we can achieve this bound, e.g., with the maximum likelihood estimator):

For n independent time samplings, the RCF bound on the variance of any unbiased estimator for A is:

$$\delta^2 \hat{A} \geq 1 / \left\langle \left[\frac{\partial}{\partial A} \sum_1^n \log f(t_i; A) \right]^2 \right\rangle \quad (3.30)$$

$$\geq 1/n \left\langle \left(\frac{\sin xt}{1 + A \sin xt} \right)^2 \right\rangle. \quad (3.31)$$

Performing the integral gives:

$$\delta^2 \hat{A} = \frac{1}{n} \sum_{k=1}^{\infty} A^{2(k-1)} \frac{x^{2k} (2k)!}{[1 + (2x)^2][1 + (4x)^2] \cdots [1 + (2kx)^2]}.$$

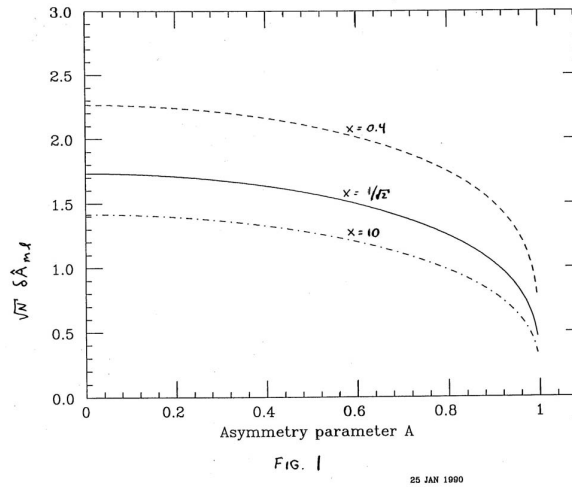


Figure 3.10: RCF bound on error in asymmetry parameter estimators.

This function is graphed as a function of the asymmetry, for selected values of x , in Fig. 3.10. The value of closest to the true value is $1/\sqrt{2}$. Figure 3.11 provides a comparison of this bound with the variance from the moment method. We may conclude that, especially for large asymmetries, significant gains may be obtained by using the detailed time information. The actual measured value for A is about 0.7.

3.5 Least Squares Method

A third popular method is the method of Least Squares Estimation [A nice discussion of this subject appears in: F. T. Solnitz, *Ann. Rev. Nucl. Sci.*, vol 14, 375-402 (1964)]:

Definition 3.7 Given a set of observations $\{x_1, \dots, x_n\}$, with expectation values $\{g_1(\theta) = \langle x_1 \rangle, \dots, g_n(\theta) = \langle x_n \rangle\}$ and covariance (moment) matrix M , then the set of parameter values $\hat{\theta}$ that minimizes the quantity:

$$S = (x - g)^T M^{-1} (x - g) \quad (3.32)$$

is called the **Least Squares Estimate (LSE)** for θ .

The intuition behind this method is that the best “fit” to the data is that set of parameter values that minimizes a measure of the deviations between the “model” ($g(\theta)$) and the data. In this case, the measure of a deviation is the squared difference, weighted according to the moment matrix, so that imprecise data with large variances carries less weight than precise data with small variances.

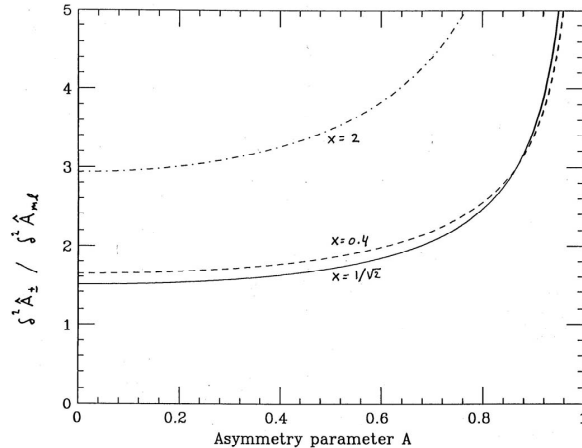


FIG. 2

25 JAN 1990

Figure 3.11: The variance according to the moment method divided by the RCF bound on the variance in the asymmetry parameter estimates.

If the x_i are sampled from a multi-variate normal distribution with known moment matrix, then the LSE is the same as the MLE. This is clear since S is what appears in the exponential of the normal PDF, with a factor of $-1/2$. We also have that S is distributed according to a χ^2 distribution with $n - r$ degrees of freedom, where r is the number of independent parameters being estimated. We will later see that this provides us with a test for “Goodness of fit”, although this is already intuitive – smaller values of S mean that the agreement between the data and the model is better than for large values of S .

Even if the observations are not normally distributed, the LSE may be useful, for example, if the distribution is approximately normal.

3.5.1 LSE - Sample Application

Suppose our data consists of a histogram which we wish to fit to some model, including the estimation of some parameters.

In general, the histogram bin contents are described by Poisson distributions, rather than normal distributions. However, if the contents are large, the normal approximation may suffice. In this case, the bins are independent, so the moment matrix is diagonal. The moment matrix is not actually known, so it must be estimated in order to apply this method. There are two common approaches to estimating these variances to be used in the fit:

1. Use the value of g_i as the estimated variance for the i th bin. In this case, the variance estimate changes as the parameters are varied. In principle this approaches correct estimates as the fit approaches correct parameter

values, but allowing the variances to change as the minimum is searched for may result in an unstable fit.

2. Use the value of x_i (observed bin contents) as the estimated variance for the i th bin. This approach is likely to be more stable, but has the danger that downward fluctuations in bin contents will carry more weight than upward fluctuations, introducing a downward bias on the estimated model.

One rule-of-thumb is that the normal approximation is typically reasonable (and the χ^2 goodness of fit valid) if each bin has at least 7 counts, although higher values are also used for this minimum. Note that it is quite permissible to combine bins until this is satisfied. The bin width need not be constant.

3.5.2 Linear Least Squares Methodology

Suppose the expectation values g_i for x_i are n linear functions of the r parameters θ :

$$\langle x \rangle = g = g_0 + F\theta,$$

where F is a matrix with n rows and r columns. It is convenient to translate the measurement vector by the constant vector g_0 :

$$y = x - g_0. \quad (3.33)$$

Then

$$S = (y - F\theta)^T M^{-1} (y - F\theta). \quad (3.34)$$

It is readily demonstrated that $\langle y \rangle = F\theta$, and that $\text{Var}(y) = M$.

We obtain $\hat{\theta}$, the values that minimize S by:

$$\left. \frac{\partial S}{\partial \theta_i} \right|_{\hat{\theta}} = 0. \quad (3.35)$$

Or, with

$$\nabla_{\theta} \equiv \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \dots \\ \frac{\partial}{\partial \theta_r} \end{pmatrix}, \quad (3.36)$$

we have (noting that the taking the transpose of a scalar does nothing)

$$\begin{aligned} 0 &= \left\{ [\nabla_{\theta}(y - F\theta)^T] M^{-1} (y - F\theta) + [(y - F\theta)^T M^{-1} \nabla_{\theta}^T (y - F\theta)]^T \right\} \Big|_{\hat{\theta}} \\ &= 2 [\nabla_{\theta}(y - F\theta)^T] M^{-1} (y - F\theta) \Big|_{\hat{\theta}} \end{aligned} \quad (3.37)$$

$$= -2F^T M^{-1} (y - F\hat{\theta}). \quad (3.38)$$

Let $H \equiv F^T M^{-1} F$; this is an $r \times r$ matrix. Then we may write

$$F^T M^{-1} F \hat{\theta} = F^T M^{-1} y, \quad (3.39)$$

or $H\hat{\theta} = F^T M^{-1}y$. Assuming H is non-singular, we solve for estimator $\hat{\theta}$:

$$\hat{\theta} = H^{-1}F^T M^{-1}y. \quad (3.40)$$

Let us check the expectation value of this estimator:

$$\langle \hat{\theta} \rangle = \langle H^{-1}F^T M^{-1}y \rangle \quad (3.41)$$

$$= H^{-1}F^T M^{-1}\langle y \rangle \quad (3.42)$$

$$= H^{-1}F^T M^{-1}F\theta \quad (3.43)$$

$$= H^{-1}H\theta \quad (3.44)$$

$$= \theta. \quad (3.45)$$

Our estimator is unbiased. We leave it as an exercise to demonstrate that

$$\text{Var}(\hat{\theta}) = H^{-1}, \quad (3.46)$$

and that we may write:

$$S = (y - F\hat{\theta})^T M^{-1}(y - F\hat{\theta}) + (\hat{\theta} - \theta)^T H(\hat{\theta} - \theta) \quad (3.47)$$

Notice the similarity between Eq. 3.47 and Eq. 3.10. Suppose that our sampling distribution is in fact multivariate normal:

$$f(y; \theta) = A \exp \left[-\frac{1}{2}(y - F\theta)^T M^{-1}(y - F\theta) \right], \quad (3.48)$$

where we leave the determination of the normalization A as an exercise. The conditions giving $\hat{\theta}$ are r linear functions of the observations x . Imagine that we “complete” this linear transformation with a transformation that takes the n variables y to the r variables $\hat{\theta}$ and $n - r$ variables z , constructed to be independent of the $\hat{\theta}$. We thus conclude that the likelihood function must be of the form:

$$L(\theta; \hat{\theta}, z) = \exp \left[-\frac{1}{2}(y - F\hat{\theta})^T M^{-1}(y - F\hat{\theta}) \right] \exp \left[-\frac{1}{2}(\hat{\theta} - \theta)^T H(\hat{\theta} - \theta) \right]. \quad (3.49)$$

This is just the original PDF, with θ replaced by the estimators $\hat{\theta}$, times a “correction term”, taking into account that $\hat{\theta}$ may differ from θ . We have split the likelihood into two independent probabilities, the probability that we will observe $\hat{\theta}$, given θ , times the probability that we will observe y given a PDF with parameters $\hat{\theta}$. The second exponential is the PDF for $\hat{\theta}$. The first exponential compares y with the predictions based on $\hat{\theta}$. But the $\hat{\theta}$ are r linear functions of the y 's, so there are really only $n - r$ variables left. Thus, the first exponential expresses the probability distribution in the remaining $n - r$ variables z , and the quadratic form:

$$\chi^2(\hat{\theta}) = (y - F\hat{\theta})^T M^{-1}(y - F\hat{\theta}) \quad (3.50)$$

is distributed according to the χ^2 distribution with $n - r$ degrees of freedom [we'll demonstrate this connection in class]. We will find this useful in testing whether the data are consistent with the “model” expressed by 3.48.

3.5.3 Non-linear Least Squares

In general, we are not lucky enough to have a linear problem. In this case:

1. First, see whether it is equivalent to a linear problem.
2. Second, if you don't need to do it often, plug it into a general-purpose minimizer. This is usually very compute intensive compared with other methods, so should only be done if you won't need to do it very many times.
3. Or, third, especially if you need to do it many times (e.g., track fitting or kinematic fitting) it may be a good approximation to linearize the problem via a Taylor series expansion about some starting value for the parameters. The process is iterated until convergence is (hopefully) attained.

The procedure in the third option is as follows: Make a Taylor series expansion of the function giving the expectation values about some initial guess for the parameter values:

$$g_i(\theta) = g_i(\theta^0) + \sum_{j=1}^r (\theta_j - \theta_j^0) \left. \frac{\partial g_i}{\partial \theta_j} \right|_{\theta^0} + \dots \quad (3.51)$$

It is desirable to pick a starting θ^0 that is near the value that minimizes S , in order for the fit to converge well. Neglecting the higher order terms, we have a problem of the form:

$$g(\theta) = g_0 + F\theta, \quad (3.52)$$

where

$$g_0 = g(\theta^0) - F\theta^0, \quad (3.53)$$

$$F_{ij} = \left. \frac{\partial g_i}{\partial \theta_j} \right|_{\theta^0}. \quad (3.54)$$

We then solve this linear problem as discussed already. Often the first solution will not be close enough to the desired minimum. In this case, we re-expand about the new estimate and iterate for a new solution. We may continue to iterate until convergence is achieved, as may be determined by small differences between iterations.

3.5.4 Constraints

When we find the minimum of

$$S = (x - g)^T M^{-1} (x - g) \quad (3.55)$$

we are attempting to find those functions g which give a "best fit". The g are n functions of r parameters. Thus, there are $n - r$ equations relating the g_i 's, that is, we have constrained the possible values of g by using these equations.

We may approach the problem differently: Let us take g themselves as n “independent” parameters, and use the method of Lagrange multipliers to introduce the constraints on the allowed values for g .

Thus, we may write:

$$S = (x - g)^T M^{-1}(x - g) + 2\lambda^T c(g, u), \quad (3.56)$$

where the factor of two is introduced for convenience, and c are k equations of constraint. That is, they are equations of the form $c = 0$. These constraint equations could, perhaps, depend not only on g , but also on some m additional unknowns u . The λ is a vector of k Lagrange multipliers. The desired “best fit” is obtained by minimizing S with respect to g , u , and the Lagrange multipliers.

If we are lucky, c is linear in g and u , otherwise we may perform a linear approximation and iterate. Thus, assume:

$$c(g, u) = c_0 + G^T g + U^T u, \quad (3.57)$$

where G is a $k \times n$ matrix:

$$G_{ij} = \left. \frac{\partial c_j}{\partial g_i} \right|_{g^0, u^0}, \quad (3.58)$$

and U is a $k \times m$ matrix:

$$U_{ij} = \left. \frac{\partial c_j}{\partial u_i} \right|_{g^0, u^0}. \quad (3.59)$$

Then

$$S = (x - g)^T M^{-1}(x - g) + 2\lambda^T c_0 + 2\lambda^T G^T g + 2\lambda^T U^T u. \quad (3.60)$$

Setting the derivatives equal to zero with respect to g , u , and λ yields the equations:

$$0 = -M^{-1}(x - \hat{g}) + G\hat{\lambda} \quad (3.61)$$

$$0 = U\hat{\lambda} \quad (3.62)$$

$$0 = c_0 + G^T \hat{g} + U^T \hat{u}. \quad (3.63)$$

To solve these equations, we may first eliminate \hat{g} and then $\hat{\lambda}$ to obtain

$$\hat{u} = -K^{-1}UH^{-1}(c_0 + G^T x), \quad (3.64)$$

where H is the $k \times k$ matrix $H \equiv G^T M G$, and K is the $m \times m$ matrix $K \equiv UH^{-1}U^T$. Then we back-substitute to find the estimators for g and λ :

$$\hat{\lambda} = H^{-1}(c_0 + G^T x + U^T \hat{u}) \quad (3.65)$$

$$\hat{g} = x - M G \hat{\lambda}. \quad (3.66)$$

Letting $E \equiv M G$ and $J \equiv UH^{-1}$, we may express our estimators as:

$$\hat{u} = -K^{-1}J(c_0 + G^T x) \quad (3.67)$$

$$\hat{\lambda} = H^{-1}(c_0 + G^T x + U^T \hat{u}) \quad (3.68)$$

$$\hat{g} = x - E\hat{\lambda}. \quad (3.69)$$

3.5.5 Explicit case of two dimensions

Consider the case of sampling from a bivariate normal distribution with common mean and known moment matrix.

$$\chi^2 = (x - \vartheta)^T M^{-1} (x - \vartheta), \quad (3.70)$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \vartheta = \begin{pmatrix} \theta \\ \theta \end{pmatrix}, \quad (3.71)$$

and

$$M = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \quad (3.72)$$

We form the least-squares estimator, $\hat{\theta}$, for θ according to

$$\left. \frac{\partial \chi^2}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0. \quad (3.73)$$

The result is

$$\hat{\theta} = \frac{\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} - (x_1 + x_2) \frac{\rho}{\sigma_1\sigma_2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}}. \quad (3.74)$$

3.6 Gauss-Markov

We introduce some additional terms at this point:

1. The term **error** is used to describe the difference between a datum and its expectation value, or between an estimator for a parameter and the true value of the parameter. We have already used this notion in our statement of the Gauss-Markov theorem.
2. The term **residual** is used to describe the difference between a datum and the “fitted” value.

It is readily demonstrated that the LSE is efficient and unbiased if the observations are normal and the parameter functions are linear. However, the LSE has some favorable properties beyond this, as implied in the following theorem:

Theorem 3.6 Gauss-Markov *Consider the linear model for our observations:*

$$y_i = \sum_{j=1}^r \theta_j s_{ji} + \epsilon_i, \quad (3.75)$$

where s_{ji} is given, and the “error” ϵ_i is sampled from some distribution, not necessarily normal. If $\langle \epsilon_i \rangle = 0$ and $\text{Var}(\epsilon_i) < \infty$, then the LSE estimator for θ is unbiased and of minimum variances among all linear unbiased estimators.

This property of the LSE is sometimes denoted “BLUE”, for “Best Linear Unbiased Estimator”.

We’ll save the proof of this as an exercise.

The “pulls” (or normalized residuals), are a handy way to tell whether the fit assumptions (e.g., M) are reasonable:

$$\text{pull}_i = \frac{x_i - g_i(\hat{\theta})}{\sqrt{M_{ii} - (FH^{-1}F^T)_{ii}}}.$$

If all is well, the pulls should be $N(0, 1)$ distributed. (Exercise)

3.7 Bayes Estimation

The Bayesian approach to parameter estimation is very similar in method to maximum likelihood estimation, with one important difference: In Bayesian estimation, the likelihood function is multiplied by a prior distribution to obtain the “posterior distribution”. The maximum of this posterior distribution is then taken to be the estimate of the parameter. Often in practice, the posterior is taken as a constant, in which case the Bayesian estimator is the same as the MLE.

3.8 Exercises

1. Prove the theorem that an efficient (perhaps biased) estimator for θ exists iff:

$$\frac{\partial \ln L(\mathbf{x}; \theta)}{\partial \theta} = [f(\mathbf{x}) - h(\theta)]g(\theta)$$

- . An unbiased efficient estimator exists iff we further have:

$$h(\theta) = \theta.$$

Hint: The RCF bound made use of the linear correlation coefficient, in which equality holds iff there is a linear relation:

$$\partial_\theta \ln L(\mathbf{x}; \theta) = a(\theta)\hat{\theta} + b(\theta).$$

2. Show that $\hat{\theta} = D(\mathbf{x})$ is an efficient estimator for θ , if x is sampled from the exponential family:

$$L(\mathbf{x}; \theta) = \exp[A(\theta)D(\mathbf{x}) + B(\theta) + C(\mathbf{x})].$$

3. If x is a sample from a normal distribution of known variance, show that x is an unbiased efficient estimator for the mean.
4. What is the bias of this estimator for \hat{a} in 3.29? Compare its efficiency with the minimum bound.

5. Generalize the moment method example to an estimator for the strength of an arbitrary $Y_{\ell m}$ moment.
6. Is the moment method always consistent?
7. Consider the simple angular distribution problem we have discussed in terms of the moment method already, with pdf:

$$\frac{d\sigma}{d\Omega} = A(1 + a \cos \vartheta),$$

where our measurement consists of the n samplings, $\{x_1, \dots, x_n\}$ of $x = \cos \vartheta$.

Find the LSE for a , and compare its properties with the estimator from the other methods.

8. Consider the simple angular distribution problem we have discussed in terms of the moment method already, with pdf:

$$\frac{d\sigma}{d\Omega} = A(1 + a \cos \vartheta),$$

where our measurement consists of the n samplings, $\{x_1, \dots, x_n\}$ of $x = \cos \vartheta$.

Find the MLE for a , and compare its properties with the estimator from the moment method.