Lecture Notes on Data Analysis

Glen D. Cowan
Universität Siegen

April 25, 1996

Preface

The following book is an introduction to the practical application of statistics in data analysis as typically encountered in the physical sciences, and in particular in high energy physics. Students entering this field do not usually go through a formal course in probability and statistics, despite having been exposed to many other advanced mathematical techniques. Statistical methods are invariably needed, however, in order to extract meaningful information from experimental data.

The book originally developed out of work with graduate students in the ALEPH collaboration at the European Organization for Nuclear Research (CERN). It is primarily aimed at graduate or advanced undergraduate students in the physical sciences engaged in research or laboratory courses which involve data analysis. It is desirable that the reader have access to a computer with mathematical and statistical program libraries (e.g. the CERN libraries), so as to be able to try out the various techniques. A number of the methods are widely used in the physical sciences but less widely understood, and it is therefore hoped that more advanced researchers can also profit from the material.

It is assumed that the reader has an understanding of linear algebra, multivariable calculus and some knowledge of complex analysis. This is essentially always the case for students in physics, engineering and other physical sciences, and thus the book should pose no serious difficulties in terms of assumed prior knowledge. Roughly speaking, the present book is somewhat less theoretically oriented than that of Eadie *et al.,* [Ead71], and somewhat more so than those of Lyons [Lyo86] and Barlow [Bar89].

An attempt has been made to present the most important concepts and tools in a manageably short space. As a consequence, many results are given without proof and the reader is often referred to the literature for more detailed explanations. It is thus considerably more compact than several other works on similar topics, e.g. those by Brandt [Bra92] and Frodeson *et al.* [Fro79]. Most chapters employ concepts introduced in previous ones. Since the book is relatively short, however, it is hoped that readers will look at least briefly at the earlier chapters before skipping to the topic needed.

The bulk of the material here was presented as a half-semester course at the University of Siegen in 1995. Given the material added since then, most of the book could be covered in 20 to 30 one-hour lectures. A major problem concerning use as a textbook is the question of exercises, since to be realistic these require a computer. Although no exercises are presented here, the reader interested in practicing the techniques is encouraged to implement the examples on a computer. By modifying the various parameters and the

input data, one can gain experience with the methods presented. This is particularly instructive in conjunction with the Monte Carlo method (Chapter 3), which allows one to generate simulated data sets with known properties. These can then be used as input for the various statistical techniques.

The topics include basic aspects of probability and statistical inference, Monte Carlo techniques, statistical tests, and methods of parameter estimation. The concept of probability plays, of course, a fundamental role. In addition to the interpretation of probability as a relative frequency as used in classical statistics, the Bayesian approach using subjective probability is discussed as well. Although the frequency interpretation tends to dominate in most of the commonly applied methods, it was felt that certain applications can be better handled with Bayesian statistics, and that a brief discussion of this approach was therefore justified.

The important topic of numerical minimization is not treated, since computer routines that perform this task are widely available in program libraries. Also omitted are techniques that are widely used in the biological sciences and economics, such as analysis of variance and time series analysis, since these are not as often applicable to problems encountered in the physical sciences.

In the last chapter[1], a number of examples are presented which demonstrate various concepts developed throughout the book. This chapter also includes a discussion of practical considerations that must be dealt with in a "real" data analysis, such as systematic and theoretical errors, data reduction, and ease of implementation of a method.

[1] In preparation.

Contents

Chapter 1

Fundamental Concepts

1.1 Probability and Random Variables

The aim of this book is to present the most important concepts and methods used in data analysis. Among these concepts, uncertainty plays a central role, since this is inevitably present in experimentally obtained information. For example, one is often faced with a situation where the outcome of a repeated measurement varies unpredictably upon repetition of the experiment. Such behaviour can stem from errors related to the measuring device, or it could be the result of a more fundamental (e.g. quantum mechanical) unpredictability inherent to the system. The uncertainty might stem from various undetermined factors which in principle could be known but in fact are not. A characteristic of a system is said to be *random* when some hypothesis concerning its nature is not known with complete certainty.

The degree of randomness can be quantified with the concept of *probability*. The mathematical theory of probability has a history dating back at least to the 17th century, and several different definitions of probability have been developed. We will use the definition in terms of set theory as formulated in 1933 by Kolmogorov [Kol33]. Consider a set S called the *sample space* consisting of a certain number of elements, the interpretation of which is left open for the moment. To each subset A of S one assigns a real number $P(A)$ called a probability, defined by the following three axioms:

(1) For every subset A in S, $P(A) \geq 0$.

(2) For any two subsets A and B that are *disjoint* (i.e. mutually exclusive, $A \cap B = \emptyset$) the probability assigned to the union of A and B is the sum of the two corresponding probabilities, $P(A \cup B) = P(A) + P(B)$.

(3) The probability assigned to the sample space is one, $P(S) = 1$.

From these axioms further properties of probability functions can be derived, e.g.

$$
\begin{aligned}
&P(\overline{A}) = 1 - P(A) \text{ where } \overline{A} \text{ is the complement of } A \\
&P(A \cup \overline{A}) = 1 \\
&0 \leq P(A) \leq 1 \\
&P(\emptyset) = 0 \\
&\text{if } A \subset B, \text{ then } P(A) \leq P(B) \\
&P(A \cup B) = P(A) + P(B) - P(A \cap B)
\end{aligned}
\tag{1.1}
$$

For proofs and further properties see e.g. [Bra92, Dud88].

A variable that takes on a specific value for each element of the set S is called a *random variable*. The individual elements may each be characterized by several quantities, in which case the random variable is a multidimensional vector.

Suppose one has a sample space S which contains subsets A and B. Provided $P(B) \neq 0$ one defines the *conditional probability* $P(A|B)$ (read P of A given B) as

$$
P(A|B) = \frac{P(A \cap B)}{P(B)} \, . \tag{1.2}
$$

Figure 1.1 shows the relationship between the sets A, B and S. One can easily show that conditional probabilities themselves satisfy the axioms of probability, both with S as well as with the subset B taken as the sample space. Note that the usual probability $P(A)$ can be regarded as the conditional probability for A given S: $P(A) = P(A|S)$.

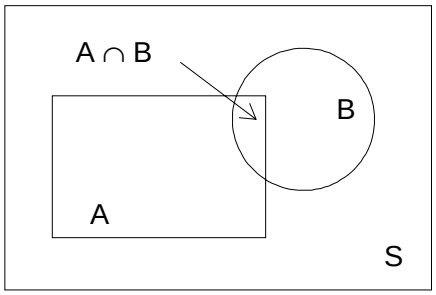

Figure 1.1: Relationship between the sets A, B and S in the definition of conditional probability.

Two subsets A and B are said to be *independent* if

$$
P(A \cap B) = P(A)\, P(B) \, . \tag{1.3}
$$

For A and B independent, it follows from the definition of conditional probability that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. (Do not confuse independent subsets according to (1.3) with *disjoint* subsets, i.e. $A \cap B = \emptyset$.)

From the definition of conditional probability one also has the probability of B given A (assuming $P(A) \neq 0$)

$$
P(B|A) = \frac{P(B \cap A)}{P(A)} \, . \tag{1.4}
$$

Since $A \cap B$ is the same as $B \cap A$, by combining equations (1.2) and (1.4) one has

$$P(B \cap A) = P(A|B)\, P(B) = P(B|A)\, P(A) \ , \tag{1.5}$$

or,

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)} \ . \tag{1.6}$$

Equation (1.6) relating the conditional probabilities $P(A|B)$ and $P(B|A)$ is called *Bayes' theorem* [Bay63].

Suppose the sample space S can be broken into disjoint subsets A_i, i.e. $S = \cup_i A_i$ with $A_i \cap A_j = \emptyset$ for $i \neq j$. Assume further that $P(A_i) \neq 0$ for all i. An arbitrary subset B can be expressed as $B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$. Since the subsets $B \cap A_i$ are disjoint, their probabilities add, giving

$$\begin{aligned} P(B) &= P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) \\ &= \sum_i P(B|A_i) P(A_i) \ . \end{aligned} \tag{1.7}$$

The last line comes from the definition (1.4) for the case $A = A_i$. This expression for the probability is useful if one can break the sample space into subsets A_i for which the probabilities are easy to calculate. It is often used with Bayes' theorem (1.6) to give

$$P(A|B) = \frac{P(B|A)\, P(A)}{\sum_i P(B|A_i) P(A_i)} \ . \tag{1.8}$$

Here A can be any subset of S, including e.g. one of the A_i.

As an example, consider a disease which is known to be carried by 0.1% of the population, i.e. the *prior* probabilities to have the disease or not are

$$\begin{aligned} P(\text{disease}) &= 0.001 \ , \\ P(\text{no disease}) &= 0.999 \ . \end{aligned}$$

A test is developed which yields a positive result with a probability of 98% given that the person carries the disease, i.e.

$$\begin{aligned} P(+|\text{disease}) &= 0.98 \ , \\ P(-|\text{disease}) &= 0.02 \ . \end{aligned}$$

Suppose there is also a 3% probability, however, to obtain a positive result for a person without the disease,

$$P(+|\text{no disease}) = 0.03\ ,$$
$$P(-|\text{no disease}) = 0.97\ .$$

Suppose your test result is positive. What is the probability that you have the disease? According to Bayes' theorem (equation (1.8)) this is given by

$$\begin{aligned} P(\text{disease}|+) &= \frac{P(+|\text{disease})\,P(\text{disease})}{P(+|\text{disease})\,P(\text{disease}) + P(+|\text{no disease})\,P(\text{no disease})} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\ &= 0.032\ . \end{aligned}$$

The probability that you have the disease given a positive test result is only 3.2%. This may be surprising, since the probability of having a wrong result is only 2% if you carry the disease and 3% if you do not. But the prior probability is extremely low, 0.1%, which leads to a *posterior* probability of only 3.2%. An important point that we have skipped over up to now is what it really means when we say $P(\text{disease}|+) = 0.032$, i.e. how exactly the probability should be interpreted. This question is examined in the next section.

1.2 Interpretation of Probability

Although any function satisfying the axioms above can be called by definition a probability function, one must still specify how to interpret the set elements and how to assign and interpret the probability values. There are two main interpretations of probability commonly used in data analysis. The most important is that of *relative frequency*, used among other things for assigning statistical errors to measurements. Another interpretation called *subjective* probability is also used, however, e.g. to quantify systematic uncertainties. These two interpretations are described in more detail below.

Probability as a Relative Frequency

In data analysis, probability is most commonly interpreted as a *limiting relative frequency*. Here the elements of the set S correspond to the possible outcomes of a measurement, assumed to be (at least hypothetically) repeatable. A subset A of S corresponds to the

occurrence of any of the outcomes in the subset. Such a subset is called an *event*, which is said to occur if the outcome of a measurement is in the subset.

A subset of S consisting of only one element denotes a single *elementary* outcome. One assigns for the probability of an elementary outcome A the fraction of times that A occurs in the limit that the measurement is repeated an infinite number of times:

$$P(A) = \lim_{n\to\infty} \frac{number\ of\ occurrences\ of\ outcome\ A\ in\ n\ measurements}{n} . \qquad (1.9)$$

The probabilities for the occurrence of any out of a set of outcomes (i.e. for a non-elementary subset A) are determined from those for individual outcomes by the addition rule given in the axioms of probability. These correspond in turn to relative frequencies of occurrence.

The relative frequency interpretation is clearly consistent with the axioms of probability, since the fraction of occurrences is always greater than or equal to zero, the frequency of any out of a set of independent outcomes is the sum of the individual frequencies, and the measurement must by definition yield some outcome (i.e. $P(S) = 1$). The conditional probability $P(A|B)$ is thus the number of cases where both A and B occur divided by the number of cases in which B occurs, regardless of whether A occurs. That is, $P(A|B)$ gives the frequency of A with the subset B taken as the sample space.

Clearly the probabilities based on such a model can never be determined experimentally with perfect precision. The basic tasks of *classical statistics* are to estimate the probabilities (assumed to have some definite but unknown values) given a finite amount of experimental data, and to test to what extent a particular model or theory that predicts probabilities is compatible with the observed data.

The relative frequency interpretation is straightforward when studying physical laws, which are assumed to act the same way in repeated experiments. The validity of the assigned probability values can be experimentally tested. The concept of relative frequency is more problematic for unique phenomena such as the Big Bang, or for the probability that the billionth digit of π is a 7. In such cases the repeatability must be regarded as an idealized property of the model only, not of the system it is supposed to describe.

Subjective Probability

Another probability interpretation is that of *subjective* (also called *Bayesian*) probability. Here the elements of the sample space[1] correspond to *hypotheses* or *propositions*, i.e. statements that are either true or false. One interprets the probability associated with a hypothesis as a measure of degree of belief:

[1] When using subjective probability the sample space is often called the hypothesis space.

$$P(A) = \textit{degree of belief that hypothesis } A \textit{ is true}\,. \tag{1.10}$$

The sample space S must be constructed such that the elementary hypotheses are mutually exclusive, i.e. only one of them is true. A subset consisting of more than one hypothesis is true if any of the hypotheses in the subset is true. That is, the union of sets corresponds to the Boolean *or* operation and the intersection corresponds to *and*. One of the hypotheses must necessarily be true, i.e. $P(S) = 1$.

The statement that a measurement will yield a given outcome a certain fraction of the time can be regarded as a hypothesis, so the framework of subjective probability includes the relative frequency interpretation. In addition, however, subjective probability can be associated with, for example, the value of an unknown constant, reflecting one's confidence that its value lies in a certain fixed interval. A probability for an unknown constant is not meaningful with the limiting frequency interpretation, since if one repeats an experiment depending on a physical parameter whose exact value is not certain (e.g. the mass of the electron) its value is either never or always in a given fixed interval. The corresponding probability would be either zero or one, but it is not known which. With subjective probability, however, a probability of 95% that the electron mass is contained in a given interval is a reflection of one's state of knowledge.

The use of subjective probability is closely related to Bayes' theorem and forms the basis of *Bayesian* (as opposed to classical) statistics. Consider again the probability to have a disease given a positive test result. From the standpoint of someone studying a large number of potential carriers of the disease, the probabilities in this problem can be interpreted as relative frequencies. The prior probability $P(\text{disease})$ is the overall fraction of people who carry the disease, and the posterior probability $P(\text{disease}|+)$ gives the fraction of people with a positive test result who are carriers. A central problem of classical statistics is to estimate the probabilities that are assumed to describe the population as a whole by examining a finite sample of data, e.g. a subsample of the population.

A specific individual, however, may be interested in the *subjective* probability that he or she has the disease given a positive test result. If no other information is available, one would usually take the prior probability $P(\text{disease})$ to be equal to the overall fraction of carriers, i.e. the same as in the relative frequency interpretation. Here, however, it is taken to mean the degree of belief that one has the disease before taking the test. If other information is available, different prior probabilities could be assigned; this aspect of Bayesian statistics is, as the name implies, subjective. Once $P(\text{disease})$ has been assigned, however, Bayes' theorem then tells how the probability to have the disease, i.e. the degree of belief in this hypothesis, changes in light of a positive test result. The use of subjective probability is discussed further in Sections 9.8 and 9.9.

1.3 Probability Density Functions

Consider a repeatable experiment whose outcome is characterized by a single continuous variable x. The sample space corresponds to the set of possible values that x can assume, and one can ask for the probability of observing a value within an infinitesimal interval $[x, x + dx]$.[2] This is given by the *probability density function* (p.d.f.) $f(x)$:

$$\textit{probability that } x \textit{ observed in the interval } [x, x + dx] = f(x)dx \ . \tag{1.11}$$

In the relative frequency interpretation, $f(x)dx$ gives the fraction of times that x is observed in the interval $[x, x + dx]$ in the limit that the total number of observations is infinitely large. The p.d.f. $f(x)$ is normalized such that the total probability (probability of some outcome) is one,

$$\int_\Omega f(x)dx = 1 \ , \tag{1.12}$$

where the region of integration Ω refers to the entire range of x, i.e. to the entire sample space.

Although finite data samples will be dealt with more thoroughly in Chapter 5, it is illustrative here to point out the relationship between a p.d.f. $f(x)$ and a set of n observations of x, $x_1, \ldots, x_n$. A set of such observations can be displayed graphically as a *histogram* as shown in Fig. 1.2. The x axis of the histogram is divided into m subintervals or *bins* of width $\Delta x_i, i = 1, \ldots, m$, where Δx_i is usually but not necessarily the same for each bin. The number of occurrences k_i of x in subinterval i, i.e. the number of entries in the bin, is given on the vertical axis. The area under the histogram is equal to the total number of entries n multiplied by Δx (or for unequal bin widths, $area = \sum_{i=1}^{m} k_i \cdot \Delta x_i$). Thus the histogram can be normalized to unit area by dividing each k_i by the corresponding bin width Δx_i and by the total number of entries in the histogram n. The p.d.f. $f(x)$ corresponds to a histogram of x normalized to unit area in the limit of zero bin width and an infinitely large total number of entries, as illustrated in Fig. 1.2(d).

One can consider cases where the variable x only takes on discrete values x_i, for $i = 1, \ldots, N$, where N can be infinite. The corresponding probabilities can be expressed as

$$\textit{probability to observe value } x_i = P(x_i) = f_i \ , \tag{1.13}$$

where $i = 1, \ldots, N$ and the normalization condition is

[2] A possible confusion can arise from the notation used here, since x refers both to the random variable and also to a value that can be assumed by the variable. Many authors use upper case for the random variable, and lower case for the value, i.e. one speaks of X taking on a value in the interval $[x, x + dx]$. This notation is avoided here for simplicity; the distinction between variables and their values should be clear from context.

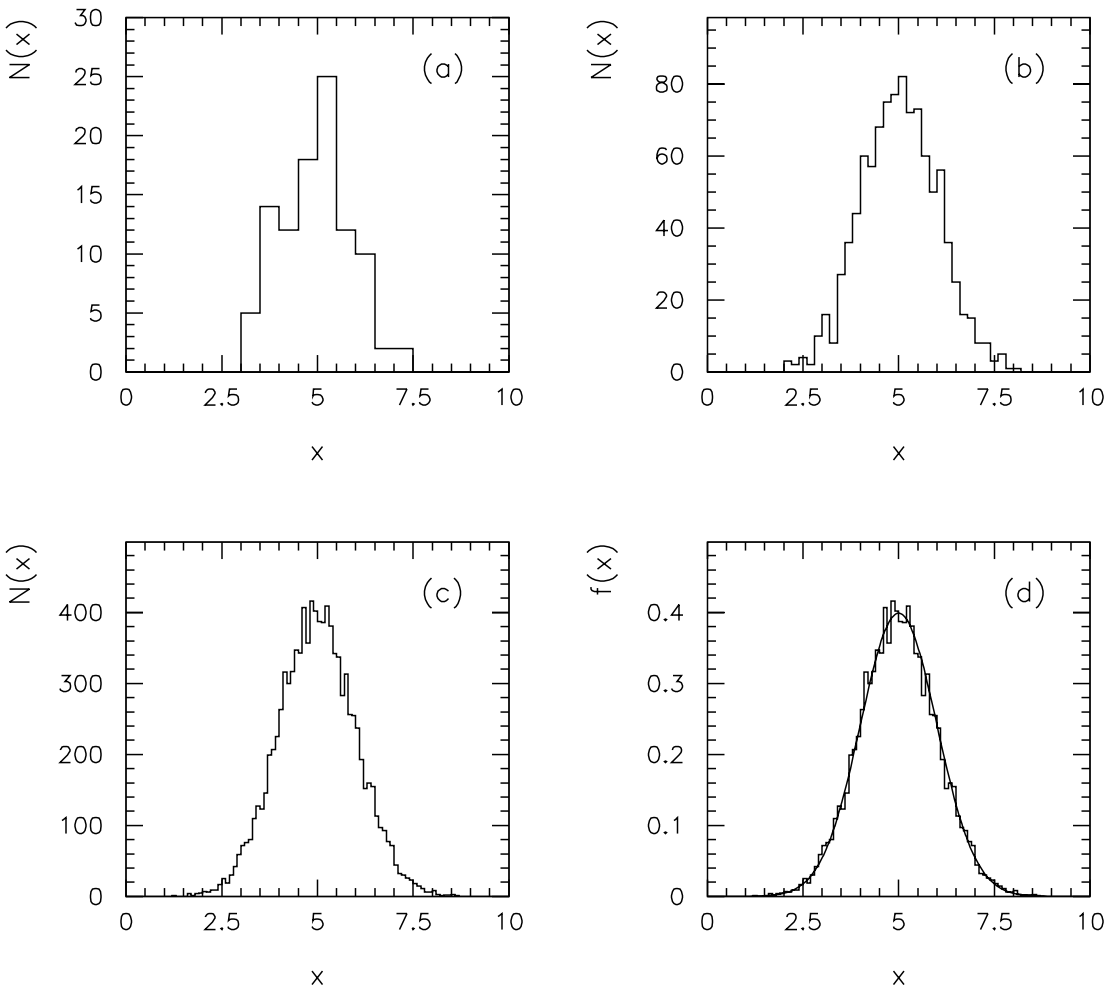

Figure 1.2: Histograms of various numbers of observations of a random variable x based on the same p.d.f. (a) $n = 100$ observations and a bin width of $\Delta x = 0.5$. (b) $n = 1000$ observations, $\Delta x = 0.2$. (c) $n = 10000$ observations, $\Delta x = 0.1$. (d) The same histogram as in (c), but normalized to unit area. Also shown as a smooth curve is the p.d.f. according to which the observations are distributed. For (a-c), the vertical axis $N(x)$ gives the number of entries in a bin containing x. For (d), the vertical axis is $f(x) = \frac{N(x)}{n\,\Delta x}$.

$$\sum_{i=1}^{N} f_i = 1 \ . \tag{1.14}$$

Although most of the examples in the following are done with continuous variables, the transformation to the discrete case is a straightforward correspondence between integrals and sums.

The *cumulative distribution* $F(x)$ is given in terms of the p.d.f. $f(x)$ as

$$F(x) = \int_{-\infty}^{x} f(x')dx' \ , \tag{1.15}$$

i.e. $F(x)$ is the probability for the random variable to take on a value less than or equal to x.[3] In fact, $F(x)$ is usually *defined* as the probability to obtain an outcome less than or equal to x, and the p.d.f. $f(x)$ is then defined as $\partial F/\partial x$. For the "well-behaved" distributions (i.e. $F(x)$ everywhere differentiable) typically encountered in data analysis the two approaches are equivalent. Figure 1.3 illustrates the relationship between the probability density $f(x)$ and the cumulative distribution $F(x)$.

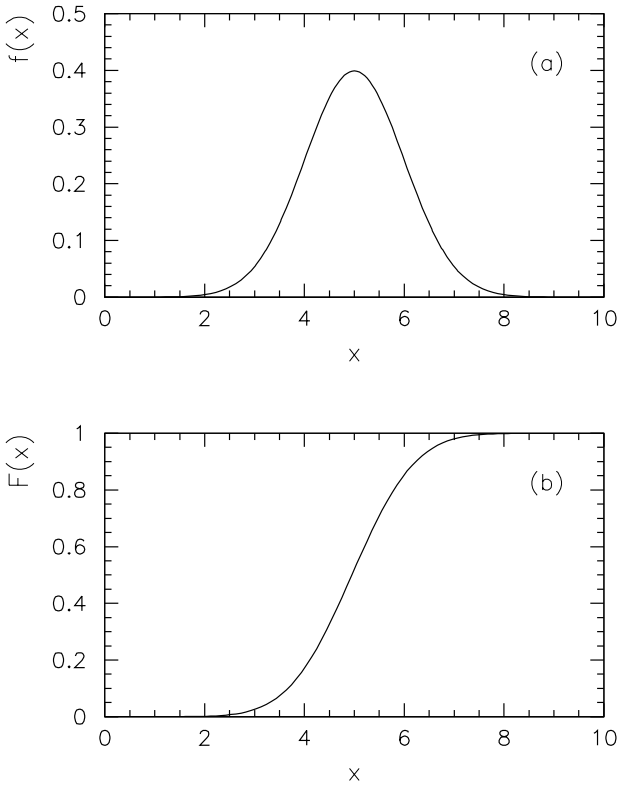

Figure 1.3: (a) A probability density function $f(x)$. (b) The corresponding cumulative distribution function $F(x)$.

For a discrete random variable x_i with probabilities $P(x_i)$ the cumulative distribution is defined to be the probability to observe values less than or equal to the value x,

[3]Mathematicians call $F(x)$ the "distribution" function, while physicists often use the word distribution to refer to the probability density function. To avoid confusion we will use the terms cumulative distribution and probability density (or p.d.f.).

$$F(x) = \sum_{x_i \leq x} P(x_i) \, . \tag{1.16}$$

A useful concept related to the cumulative distribution is the so-called *quantile of order* α or α-point. The quantile x_α is defined as the value of the random variable x such that $F(x_\alpha) = \alpha$, with $0 \leq \alpha \leq 1$. That is, the quantile is simply the inverse function of the cumulative distribution,

$$x_\alpha = F^{-1}(\alpha) \, . \tag{1.17}$$

A commonly used special case is $x_{1/2}$, called the *median* of x.

Consider now the case where the result of a measurement is characterized not by one but by several quantities, which may be regarded as a multidimensional random vector. If one is studying people, for example, one might measure for each person their height, weight, age, etc. Suppose a measurement is characterized by two continuous random variables x and y. The *joint* p.d.f. $f(x, y)$ is defined by

$$\textit{probability of } x \textit{ in } [x, x + dx] \textit{ and } y \textit{ in } [y, y + dy] = f(x, y) dx dy \, . \tag{1.18}$$

Since x and y must take on some values, one has

$$\int \int_\Omega f(x, y) dx dy = 1 \, . \tag{1.19}$$

Speaking again in terms of sets as in Section 1.1, let the event A be "x observed in $[x, x + dx]$" and let B be "y in $[y, y + dy]$". One then has $f(x, y)dxdy = P(A \cap B)$. In the relative frequency interpretation of probability, $f(x, y)$ corresponds to the density of points on a scatter plot of x and y in the limit of infinitely many points, as shown in Fig. 1.4.

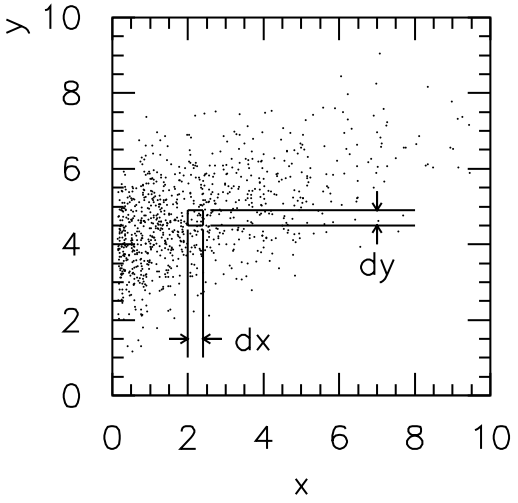

Figure 1.4: A scatter plot of two random variables x and y based on 1000 observations. The probability for a point to be in the square shown at (x, y) is given by the joint p.d.f. times the area element, $f(x, y)dxdy$.

Suppose the x axis is broken into intervals of width dx labeled by the index i. Let event A_i correspond to observing x in the interval i, and let B refer to observing y in a given interval $[y, y+dy]$, i.e. $P(A_i \cap B) = f(x_i, y)dxdy$. Since the events A_i are mutually exclusive, by summing over all intervals i one obtains

$$P(B) = \sum_i P(A_i \cap B) = f_y(y)dy \tag{1.20}$$

for the probability of observing the probability of y in $[y, y+dy]$ regardless of the value of x. The function $f_y(y)$ is called the *marginal* p.d.f. for y and is related to the joint p.d.f. by

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y)dx \ . \tag{1.21}$$

This corresponds to the normalized histogram of y obtained by projecting a scatter plot of x and y onto the y axis. Similarly, one obtains the marginal p.d.f. $f_x(x)$ by integrating $f(x, y)$ over y. The relationship between the marginal and joint p.d.f.'s are illustrated in Fig. 1.5.

From the definition of conditional probability (1.2), the probability for y to be in $[y, y+dy]$ (event A) given that x is in $[x, x+dx]$ (event B) is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{f(x, y)dxdy}{f_x(x)dx} \ . \tag{1.22}$$

The conditional p.d.f. for y given x, $h(y|x)$, is thus defined as

$$h(y|x) = \frac{f(x, y)}{f_x(x)} \ . \tag{1.23}$$

This corresponds to the normalized histogram of y obtained from the projection onto the y axis of a thin band in x (i.e. with infinitesimal width dx) from an (x, y)-scatter plot. This is illustrated in Fig. 1.6 for two values of x, leading to two different conditional p.d.f.'s, $h(y|x_1)$ and $h(y|x_2)$. Note that $h(y|x_1)$ and $h(y|x_2)$ in Fig. 1.6(b) are both normalized to unit area, as required by the definition of a probability density.

Similarly, the conditional p.d.f. for x given y is

$$g(x|y) = \frac{f(x, y)}{f_y(y)} \ . \tag{1.24}$$

Combining equations (1.23) and (1.24) gives the relationship between $g(x|y)$ and $h(y|x)$,

$$g(x|y) = \frac{h(y|x)f_x(x)}{f_y(y)} \ , \tag{1.25}$$

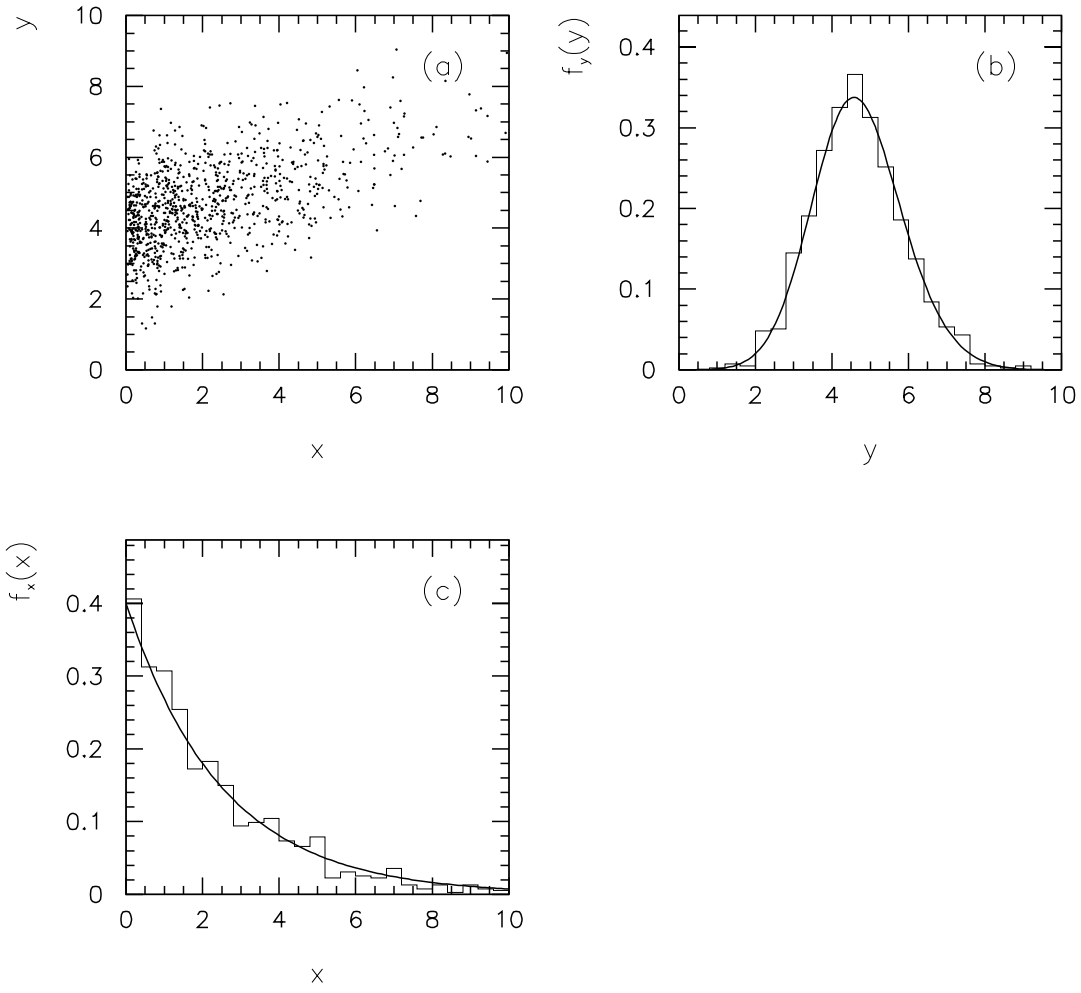

Figure 1.5: (a) The density of points on the scatter plot is given by the joint p.d.f. $f(x, y)$. (b) Normalized histogram from projecting the points onto the y axis with the corresponding marginal p.d.f. $f_y(y)$. (c) Projection onto the x axis giving $f_x(x)$.

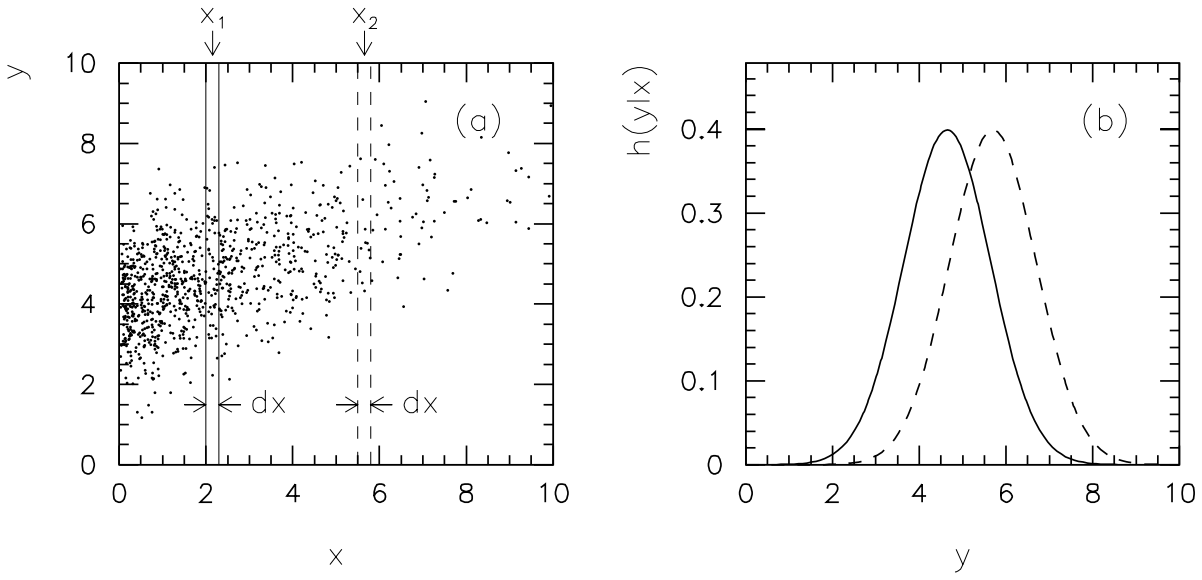

Figure 1.6: (a) A scatter plot of random variables x and y indicating two infinitesimal bands in x with width dx at x_1 (solid band) and x_2 (dashed band). (b) The conditional p.d.f.'s $h(y|x_1)$ and $h(y|x_2)$ corresponding to the projections of the bands onto the y axis.

which is Bayes' theorem for the case of continuous variables (cf. equation (1.6)).

By using $f(x,y) = h(y|x)\, f_x(x) = g(x|y)\, f_y(y)$, one can express the marginal p.d.f.'s as

$$\begin{aligned} f_x(x) &= \int_{-\infty}^{\infty} g(x|y) f_y(y) dy \\ f_y(y) &= \int_{-\infty}^{\infty} h(y|x) f_x(x) dx \, . \end{aligned} \tag{1.26}$$

These correspond to the expansion of $P(B)$ given by equation (1.7), generalized to the case of continuous random variables.

If "x in $[x, x+dx]$" (event A) and "y in $[y+dy]$" (event B) are independent, i.e. $P(A \cap B) = P(A)\, P(B)$, then the corresponding joint p.d.f. for x and y factorizes:

$$f(x,y) = f_x(x)\, f_y(y) \, . \tag{1.27}$$

From equations (1.23) and (1.24) one sees that for independent random variables x and y the conditional p.d.f. $g(x|y)$ is the same for all y, and similarly $h(y|x)$ does not depend on x. In other words, having knowledge of one of the variables does not change the probabilities for the other. The variables x and y shown in Fig. 1.6, for example, are not independent, as can be seen from the fact that $h(y|x)$ depends on x.

1.4 Functions of Random Variables

Functions of random variables are themselves random variables. Suppose $a(x)$ is a continuous function of a continuous random variable x, where x is distributed according to the p.d.f. $f(x)$. What is the p.d.f. $g(a)$ that describes the distribution of a? This is determined by requiring that the probability for x to occur between x and $x + dx$ be equal to the probability for a to be between a and $a + da$. That is,

$$g(a')da' = \int_{d\Omega} f(x)dx \ , \tag{1.28}$$

where the integral is carried out over the infinitesimal volume element $d\Omega$ defined by the region in x-space between $a(x) = a'$ and $a(x) = a' + da'$, as shown in Fig. 1.7(a). If the function $a(x)$ can be inverted to obtain $x(a)$, equation (1.28) gives

$$g(a)da = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right| = \int_{x(a)}^{x(a)+\left|\frac{dx}{da}\right|da} f(x')dx' \ , \tag{1.29}$$

or

$$g(a) = f(x(a)) \left| \frac{dx}{da} \right| \ . \tag{1.30}$$

The absolute value of dx/da insures that the integral is positive. If the function $a(x)$ does not have a unique inverse, one must include in $d\Omega$ contributions from all regions in x-space between $a(x) = a'$ and $a(x) = a' + da'$, as shown in Fig. 1.7(b).

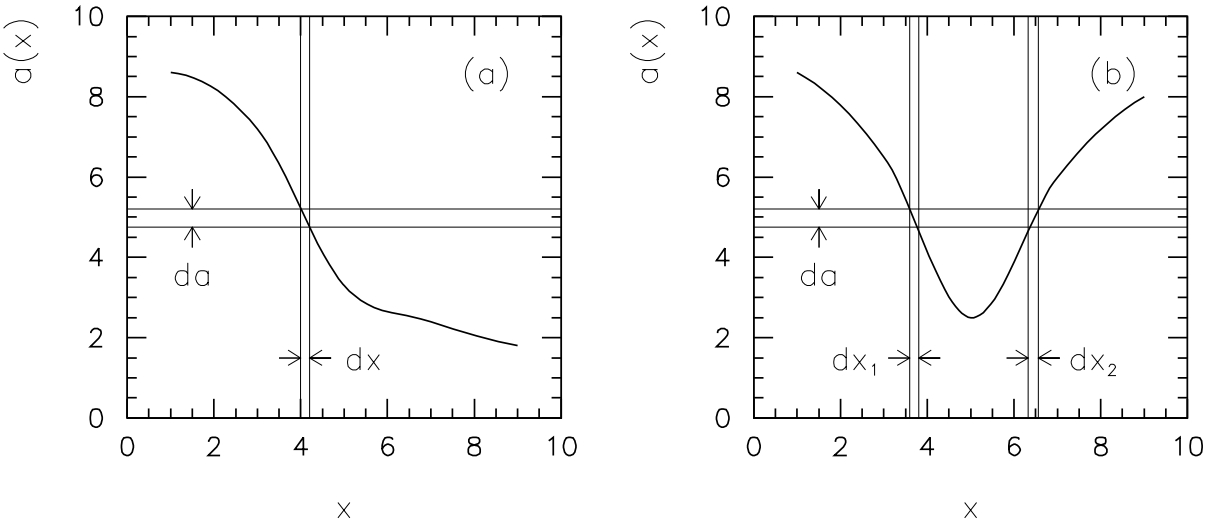

Figure 1.7: Transformation of variables for (a) a function $a(x)$ with a single valued inverse $x(a)$ and (b) a function for which the interval da corresponds to two intervals dx_1 and dx_2.

The p.d.f. $g(a)$ of a function $a(x_1, \ldots, x_n)$ of n random variables $x_1, \ldots, x_n$ with the joint p.d.f. $f(x_1, \ldots, x_n)$ is determined by

$$g(a')da' = \int \cdots \int_{d\Omega} f(x_1, \ldots, x_n) dx_1 \cdots, dx_n \ , \tag{1.31}$$

where the infinitesimal volume element $d\Omega$ is the region in $x_1, \ldots, x_n$-space between the two (hyper)surfaces defined by $a(x_1, \ldots, x_n) = a'$ and $a(x_1, \ldots, x_n) = a' + da'$.

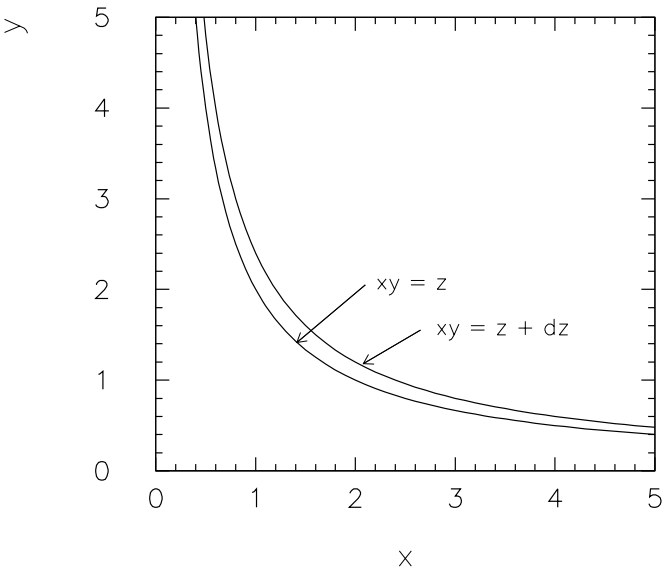

Figure 1.8: The region of integration $d\Omega$ contained between the two curves $xy = z$ and $xy = z + dz$. Occurrence of (x, y) values between the two curves results in occurrence of z values in the corresponding interval $[z, z + dz]$.

As an example of this technique, consider two independent random variables, x and y, distributed according to $g(x)$ and $h(y)$, and suppose we would like to find the p.d.f. of their product $z = xy$. Since x and y are assumed to be independent, their joint p.d.f. is given by $g(x)h(y)$. Equation (1.31) then gives for the p.d.f. of z, $f(z)$,

$$f(z)dz = \int \int_{d\Omega} g(x)h(y)dxdy = \int_{-\infty}^{\infty} g(x)dx \int_{z/x}^{(z+dz)/x} h(y)dy \ , \tag{1.32}$$

where $d\Omega$ is given by the region between $xy = z$ and $xy = z + dz$, as shown in Fig 1.8. This yields

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z/x)\frac{dx}{x} \\ &= \int_{-\infty}^{\infty} g(z/y)h(y)\frac{dy}{y} \ , \end{aligned} \tag{1.33}$$

where the second equivalent expression is obtained by reversing the order of integration. Equation (1.33) is often written $f = g \otimes h$, and the function f is called the *Mellin convolution* of g and h.

Similarly, the p.d.f. $f(z)$ of the sum of two random variables $z = x + y$ is found to be

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} g(x)h(z-x)dx \\ &= \int_{-\infty}^{\infty} g(z-y)h(y)dy \ . \end{aligned} \tag{1.34}$$

Equation (1.34) is also often written $f = g \otimes h$, and f is called the *Fourier convolution* of g and h. In most cases the names Fourier and Mellin are dropped and one must infer from context what kind of convolution is meant.

Another technique for determining the p.d.f. of a function of random variables is the following. Given n random variables $x_1, \ldots, x_n$ one can form n linearly independent functions $a_i(x_1, \ldots, x_n)$, $i = 1, \ldots, n$. Assuming the functions $a_1, \ldots, a_n$ can be inverted to give $x_i(a_1, \ldots, a_n), i = 1, \ldots, n$, the joint p.d.f. for the a_i is given by

$$g(a_1, \ldots, a_n) = f(x_1, \ldots, x_n)|J| \ , \tag{1.35}$$

where $|J|$ is the absolute value of the Jacobian determinant for the transformation,

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial a_1} & \frac{\partial x_1}{\partial a_2} & \cdots & \frac{\partial x_1}{\partial a_n} \\ \frac{\partial x_2}{\partial a_1} & \frac{\partial x_2}{\partial a_2} & \cdots & \frac{\partial x_2}{\partial a_n} \\ \vdots & & & \vdots \\ & & \cdots & \frac{\partial x_n}{\partial a_n} \end{vmatrix} . \tag{1.36}$$

In this procedure one maps n variables $x_1, \ldots, x_n$ onto n functions, $a_1, \ldots, a_n$, for which the joint p.d.f. is obtained. To determine the marginal p.d.f. for one of the functions (say $g_1(a_1)$) the joint p.d.f. $g(a_1, \ldots, a_n)$ must be integrated over the remaining a_i.

In many cases the techniques given above are too difficult to solve analytically. For example, if one is interested in a single function of n random variables, where n is some large and itself possibly variable number, it is rarely practical to come up with $n - 1$ additional functions and then integrate the transformed joint p.d.f. over the unwanted ones. In such cases a numerical solution can usually be found using the Monte Carlo techniques discussed in Chapter 3. If only the mean and variance of a function are needed, the so-called "error propagation" procedures described in Section 1.6 can be applied.

For certain cases the p.d.f. of a function of random variables can be found using integral transform techniques, specifically, Fourier transforms of the p.d.f.'s for sums of random variables and Mellin transforms for products. The basic idea is to take the Mellin or Fourier transform of equation (1.33) or (1.34) respectively. The equation $f = g \otimes h$ is then converted into the product of the transformed density functions, $\tilde{f} = \tilde{g} \cdot \tilde{h}$. The p.d.f. f is obtained by finding the inverse transform of $\tilde{f}$. A complete discussion of these methods is beyond the scope of this book; see e.g. reference [Spr79]. An example of a sum of random variables using Fourier transforms is given in Chapter 11.

1.5 Expectation Values

The *expectation value* $E[x]$ of a random variable x distributed according to the p.d.f. $f(x)$ is defined as

$$E[x] = \int_{-\infty}^{\infty} x f(x) dx = \mu \ . \tag{1.37}$$

The expectation value of x (also called the *population mean* or simply the mean of x) is often denoted by μ. Since $f(x)dx$ is the fraction of measurements with x in $[x, x+dx]$, $E[x]$ is the average value (arithemetic mean) of x one would obtain after infinitely many measurements. Note that $E[x]$ is not a function of x, but depends rather on the form of the p.d.f. $f(x)$. For a function $a(x)$, the expectation value is

$$E[a] = \int_{-\infty}^{\infty} a g(a) da = \int_{-\infty}^{\infty} a(x) f(x) dx \ . \tag{1.38}$$

The second integral is equivalent as can be seen by multiplying both sides of equation (1.28) by a and extending the region of integration to cover the entire space. The expectation value $E[a(x)]$ is not a function of x, but depends on the functional form of $a(x)$ and the p.d.f. $f(x)$.

Some more expectation values of interest are:

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x) dx = \mu'_n \ , \tag{1.39}$$

called the nth algebraic moment of x, for which $\mu = \mu'_1$ is a special case, and

$$E[(x - E[x])^n] = \int_{-\infty}^{\infty} (x-\mu)^n f(x) dx = \mu_n \ , \tag{1.40}$$

called the nth central moment of x. In particular, the second central moment,

$$E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2 = V[x] \ , \tag{1.41}$$

is called the *population variance* (or simply the variance) of x, written σ^2 or $V[x]$. Note that $E[(x - E[x])^2] = E[x^2] - \mu^2$. The variance is a measure of how widely x is spread about its mean value. The square root of the variance σ is called the *standard deviation* of x, which is often useful because it has the same dimension as x.

For the case of a function of more than one random variable, e.g. $a(x_1, \ldots, x_n)$ the expectation value is

$$\begin{aligned} E[a(x_1, \ldots, x_n)] &= \int_{-\infty}^{\infty} a g(a) da \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n = \mu_a \ , \end{aligned} \tag{1.42}$$

where $g(a)$ is the p.d.f. for a and $f(x_1, \ldots, x_n)$ is the joint p.d.f. for $x_1, \ldots, x_n$. In the following the notation $\mu_a = E[a]$ will often be used. As in the single variable case the two integrals in (1.42) are equivalent, as can be seen by multiplying both sides of equation (1.31) by a and extending the regions of integration to cover the entire space. The variance of a is

$$V[a] = E[(a - \mu_a)^2] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (a(x_1, \ldots, x_n) - \mu_a)^2 f(x_1, \ldots, x_n) dx_1 \cdots dx_n = \sigma_a^2 \ , \tag{1.43}$$

and is denoted by σ_a^2 or $V[a]$. The *covariance* of two random variables x and y is defined as

$$\begin{aligned} V_{xy} &= E[(x - \mu_x)(y - \mu_y)] = E[xy] - \mu_x \mu_y \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, y\, f(x, y)\, dx\, dy \ - \ \mu_x \mu_y \ , \end{aligned} \tag{1.44}$$

where $\mu_x = E[x]$ and $\mu_y = E[y]$. More generally, for two functions of n random variables $a(x_1, \ldots, x_n)$ and $b(x_1, \ldots, x_n)$ the covariance V_{ab} is given by

$$\begin{aligned} V_{ab} &= E[(a - \mu_a)(b - \mu_b)] \\ &= E[ab] - \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a\, b\, g(a, b)\, da\, db \ - \ \mu_a \mu_b \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} a(x_1, \ldots, x_n)\, b(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n) dx_1 \cdots dx_n \ - \ \mu_a \mu_b \ , \end{aligned} \tag{1.45}$$

where $g(a, b)$ is the joint p.d.f. for a and b and $f(x_1, \ldots, x_n)$ is the joint p.d.f. for the x_i. As in equation (1.42), the two integral expressions for V_{ab} are equivalent. Note that by construction the covariance matrix V_{ab} (sometimes called the error matrix) is symmetric in a and b and that the diagonal elements $V_{aa} = \sigma_a^2$ (i.e. the variances) are positive. V_{ab} is sometimes denoted by $\mathrm{cov}[a, b]$.

In order to give a dimensionless measure of the level of correlation between two random variables x and y, one often uses the *correlation coefficient*, defined by

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y} \ . \tag{1.46}$$

It can be shown (see e.g. [Fro79], [Bra92]) that the correlation coefficient lies in the range $-1 \leq \rho_{xy} \leq 1$.

One can roughly understand the covariance of two random variables x and y in the following way. V_{xy} is the expectation value of $(x - \mu_x)(y - \mu_y)$, the product of the

deviations of x and y from their means, μ_x and μ_y. Suppose that whenever x is observed to be greater than μ_x one has an enhanced probability for y also to be greater than μ_y, and x less than μ_x gives an enhanced probability to have y less than μ_y. Then V_{xy} is clearly greater than zero, and the variables are said to be positively correlated. Such a situation is illustrated in Fig. 1.9 (a), (c) and (d), for which the correlation coefficients ρ_{xy} are 0.75, 0.95 and 0.25 respectively. Similarly, $V_{xy} < 0$ is called a negative correlation: having $x > \mu_x$ increases the probability to observe $y < \mu_y$. An example is shown in Fig. 1.9(b), for which $\rho_{xy} = -0.75$.

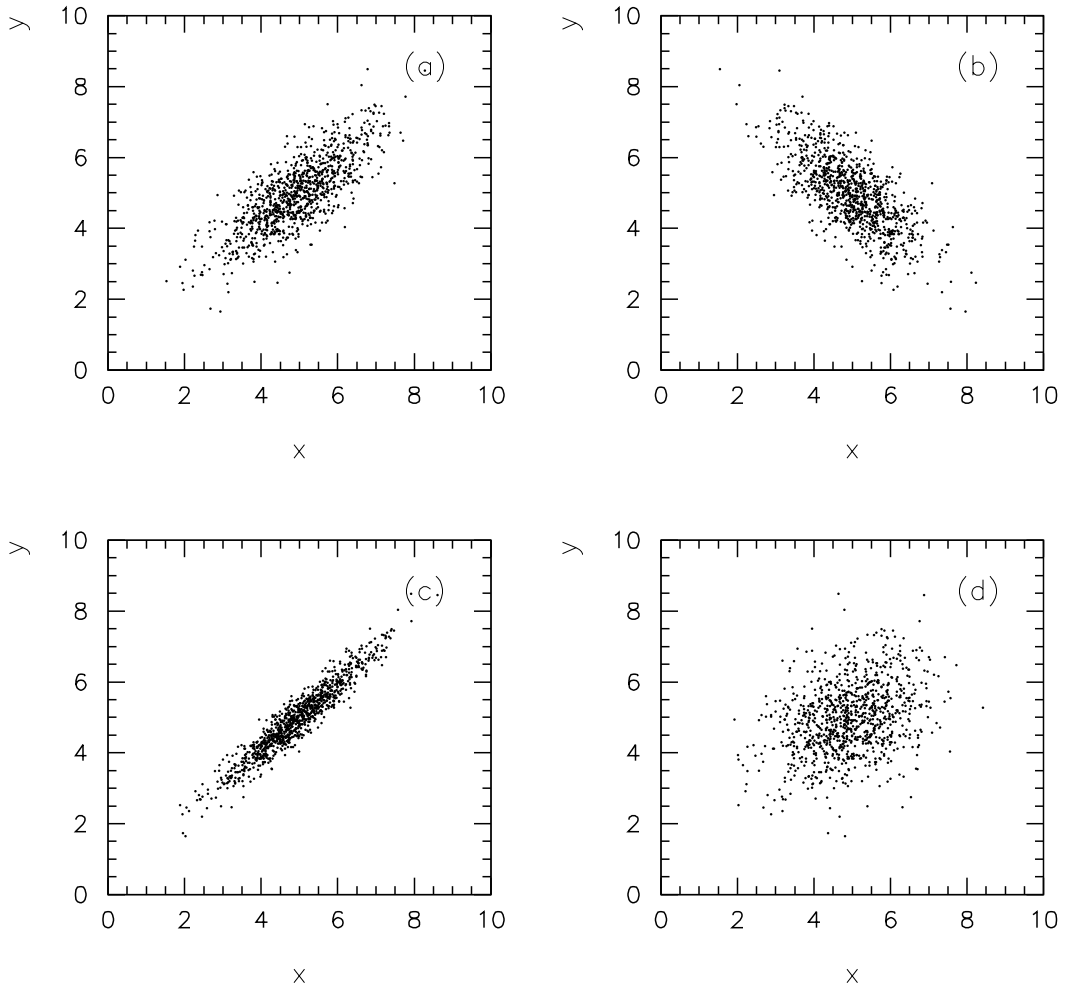

Figure 1.9: Scatter plots of random variables x and y with (a) a positive correlation, $\rho = 0.75$, (b) a negative correlation, $\rho = -0.75$, (c) $\rho = 0.95$, and (d) $\rho = 0.25$. For all four cases the standard deviations of x and y are $\sigma_x = \sigma_y = 1$.

From equations (1.27), (1.37) and (1.42) one sees that for independent random variables x and y one has

$$E[xy] = E[x]E[y] = \mu_x \mu_y \tag{1.47}$$

(and hence by equation (1.44) $V_{xy} = 0$) although the converse is not necessarily true.

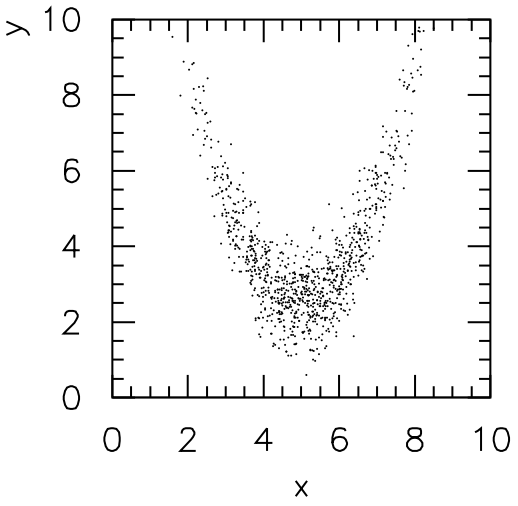

Figure 1.10: Scatter plot of random variables x and y which are not independent (i.e. $f(x,y) \neq f_x(x) f_y(y)$) but for which $V_{xy} = 0$ because of the particular symmetry of the distribution.

Figure 1.10, for example, shows a two-dimensional scatter plot of a p.d.f. for which $V_{xy} = 0$, but where x and y are not independent. That is, $f(x,y)$ does not factorize according to equation (1.27), and hence knowledge of one of the variables affects the conditional p.d.f. of the other. The covariance V_{xy} vanishes, however, because $f(x,y)$ is symmetric in x about the mean μ_x.

1.6 Error Propagation

Suppose one has a set of n random variables $\vec{x} = (x_1, \ldots, x_n)$ distributed according to some joint p.d.f. $f(\vec{x})$. Suppose that the p.d.f. is not completely known, but the mean values of the x_i, $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ and the covariance matrix, V_{ij} are known or have at least been estimated. (Methods for doing this are described in Chapters 6 – 8.)

Now consider a function of the n random variables $a(\vec{x})$. To determine the p.d.f. for a, one must in principle follow a procedure such as those described in Section 1.4 (e.g. equations (1.31) or (1.35)). We have assumed, however, that $f(\vec{x})$ is not completely known, only the means $\vec{\mu}$ and the covariance matrix V_{ij}, so this is not possible. One can, however, approximate the expectation value of a and the variance $V[a]$ by first expanding the function $a(\vec{x})$ to first order about the mean values of the x_i (assumed known):

$$a(\vec{x}) \approx a(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \; . \tag{1.48}$$

The expectation value of a is to first order

$$E[a(\vec{x})] \approx a(\vec{\mu}) \; , \tag{1.49}$$

since $E[x_i - \mu_i] = 0$. The expectation value of a^2 is

$$
\begin{aligned}
E[a^2(\vec{x})] &\approx a^2(\vec{\mu}) + 2a(\vec{\mu}) \cdot \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i] \\
&+ E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial a}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)\right)\right] \\
&= a^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial a}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij} \ , \qquad (1.50)
\end{aligned}
$$

so that the variance $V[a] = E[a^2] - (E[a])^2$ is given by

$$
V[a(\vec{x})] \approx \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial a}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij} \ . \qquad (1.51)
$$

Similarly, one obtains for the covariance of two functions $a(\vec{x})$ and $b(\vec{x})$

$$
V_{ab} \approx \sum_{i,j=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial b}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij} \ . \qquad (1.52)
$$

Equations (1.51) and (1.52) form the basis of *error propagation* (i.e. the variances, which are used as measures of statistical errors, are propagated from the x_i to the functions a, b, etc.). For the case where the x_i are not correlated, that is, $V_{ii} = \sigma_i^2$ and $V_{ij} = 0$ for $i \neq j$, equations (1.51) and (1.52) become

$$
V[a(\vec{x})] = \sigma_a^2 \approx \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i}\right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2 \qquad (1.53)
$$

and

$$
V_{ab} \approx \sum_{i=1}^{n} \left[\frac{\partial a}{\partial x_i}\frac{\partial b}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} \sigma_i^2 \ . \qquad (1.54)
$$

Equation (1.51) leads to the following special cases. If $a = x + y$, the variance of a is then

$$
\sigma_a^2 = \sigma_x^2 + \sigma_y^2 + 2V_{xy} \ . \qquad (1.55)
$$

For the product $a = xy$ one obtains

$$
\frac{\sigma_a^2}{a^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} + 2\frac{V_{xy}}{xy} \ . \qquad (1.56)
$$

If the variables x and y are not correlated ($V_{xy} = 0$), the relations above state that errors (i.e. standard deviations) add quadratically for the sum $a = x + y$, and that the *relative* errors add quadratically for the product $a = xy$.

In deriving the error propagation formulas we have assumed that the means and covariances of the original set of variables $x_1, \ldots, x_n$ are known (or at least estimated) and that the desired functions of these variables can be approximated by the first order Taylor expansion around the means $\mu_1, \ldots, \mu_n$. The latter assumption is of course only exact for a linear function. The approximation breaks down if the function $a(\vec{x})$ (or functions a, b) are significantly non-linear in a region around the means $\vec{\mu}$ of a size comparable to the standard deviations of the x_i, $\sigma_1, \ldots, \sigma_n$. Care must be taken, for example, with functions like $a(x) = 1/x$ when $E[x] = \mu$ is comparable to or smaller than the standard deviation of x. Such situations can be better treated with the Monte Carlo techniques described in Chapter 3, or using confidence intervals as described in Section 9.2.

Chapter 2

Examples of Probability Functions

In this chapter a number of commonly used probability distributions and density functions are presented. Properties such as mean and variance are given, mostly without proof. Additional p.d.f's and details on how to compute their means, variances, etc. can be found in e.g. [Fro79] Chapter 4, [Ead71] Chapter 4, [Bra92] Chapter 5.

2.1 Binomial and Multinomial Distributions

Consider a series of N independent trials or observations for which there are two possible outcomes, here called "success" and "failure", where the probability for success is some constant value, p. For example, one could define success if a measured quantity lands in a particular bin of a histogram, failure if not, with N total entries in the histogram. The set of trials can be regarded as a single measurement and is characterized by a discrete random variable k, defined to be the total number of successes. Note that here the entire set of observations is treated as a single random measurement, not each individual trial. That is, the sample space is defined to be the set of possible values of k successes given N observations. If one were to repeat the entire experiment many times with N trials each time, the resulting values of k would occur with relative frequencies given by the so-called binomial distribution.

The form of the binomial distribution can be derived in the following way: We have assumed that the probability of success in a single observation is p and the probability of failure is $1-p$. Since the individual trials are assumed to be independent, the probability for a series of successes and failures in a particular order is equal to the product of the individual probabilities. For example, the probability in five trials to have success, success, failure, success, failure in that order is $p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3(1-p)^2$. In general the probability for a particular sequence of k successes and $N-k$ failures is $p^k(1-p)^{N-k}$. We are not interested in the order, however, just in the final number of successes k. The number of sequences having k successes in N events is

$$\frac{N!}{k!(N-k)!}\ , \tag{2.1}$$

so the total probability to have k successes in N events is

$$f(k;N,p) = \frac{N!}{k!(N-k)!}\, p^k\,(1-p)^{N-k}\ , \tag{2.2}$$

for $k = 0,\ldots,N$. Note that $f(k;N,p)$ is itself a probability, not a probability density. The notation used is that the random variable (or variables) are listed as arguments of the probability function (or p.d.f.) to the left of the semicolon, and any parameters (in this case N and p) are listed to the right. Moments of k can be computed by using the binomial theorem, which states for arbitrary quantities p and q,

$$\sum_{k=0}^{N} \frac{N!}{k!(N-k)!}\, p^k\, q^{N-k} = (p+q)^N\ . \tag{2.3}$$

In order to compute the nth algebraic moment $E[k^n]$ one set $q = 1-p$, temporarily regard p and q as independent, and then set q again equal to $1-p$. This gives

$$\begin{aligned} E[k^n] &= \sum_{k=0}^{N} k^n\, \frac{N!}{k!(N-k)!}\, p^k\,(1-p)^{N-k} \\ &= \left(p\frac{\partial}{\partial p}\right)^n \sum_{k=0}^{N} \frac{N!}{k!(N-k)!}\, p^k\, q^{N-k}\Bigg|_{q=1-p} \\ &= \left(p\frac{\partial}{\partial p}\right)^n (p+q)^N\Big|_{q=1-p}\ . \end{aligned} \tag{2.4}$$

Using this one can compute the expectation value of k,

$$E[k] = Np\ , \tag{2.5}$$

and variance,

$$\begin{aligned} V[k] &= E[k^2] - (E[k])^2 \\ &= Np(1-p)\ . \end{aligned} \tag{2.6}$$

Recall that expectation values are not functions of the random variable, but they depend on the parameters of the probability function, in this case p and N. The binomial probability distribution is shown in Fig. 2.1 and Fig. 2.2 for various values of p and N.

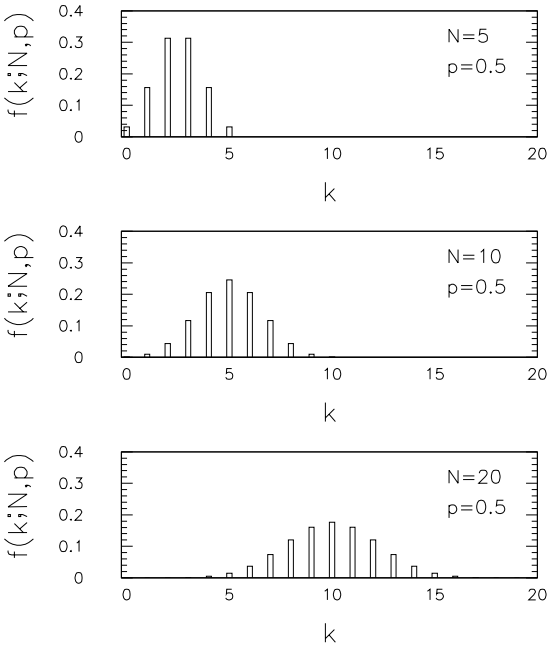

Figure 2.1: The binomial distribution for $p = 0.5$ and various values of N.

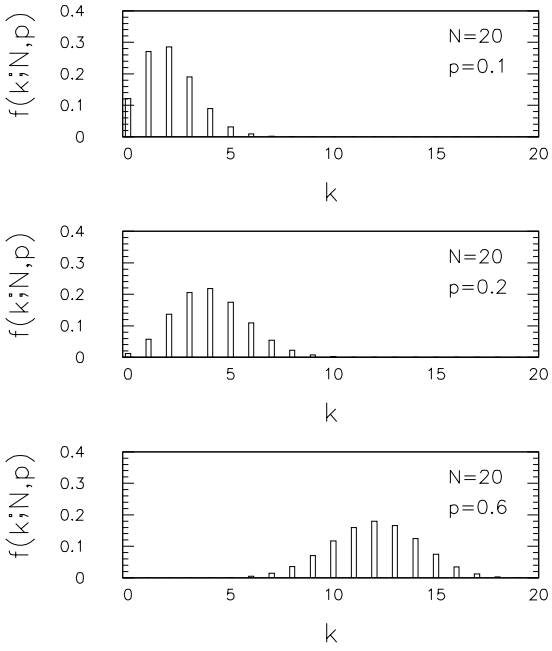

Figure 2.2: The binomial distribution for $N = 20$ and various values of p.

The *multinomial* distribution is the generalization of the binomial distribution to the case where there are not just two outcomes ("success" and "failure") but rather m different possible outcomes. For a particular trial the probability of outcome i is p_i, and since one of the outcomes must be realized, one has the normalization condition $\sum_{i=1}^{m} p_i = 1$.

Now consider a measurement consisting of N trials, each of which yields one of the possible m outcomes. The probability for a particular sequence of outcomes, e.g. i on the first trial, j on the second, and so on, in a particular order, is the product of the N corresponding probabilities, $p_i p_j \cdots p_k$. The number of such sequences that will lead to k_1 outcomes of type 1, k_2 outcomes of type two, etc., is

$$\frac{N!}{k_1! k_2! \cdots k_m!} \tag{2.7}$$

If we are not interested in the order of the outcomes, just the total numbers of each type, then the joint probability for k_1 outcomes of type 1, k_2 of type 2, etc. is given by the multinomial distribution,

$$f(k_1, \ldots, k_m; N, p_1, \ldots, p_m) = \frac{N!}{k_1! k_2! \cdots k_m!} p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m} \ . \tag{2.8}$$

Suppose one breaks the m possible outcomes into two categories: outcome i ("success") and not outcome i ("failure"). Since this is the same as the binomial process presented above, the number of occurrences of outcome i, k_i, must be binomially distributed. This

is of course true for all i. From equations (2.5) and (2.6) one has that the expectation value of k_i is $E[k_i] = Np_i$ and the variance is $V[k_i] = Np_i(1-p_i)$.

Consider now the three possible outcomes: i, j and everything else. The probability to have k_i outcomes of type i, k_j of type j and $N - k_i - k_j$ of everything else is

$$f(k_i, k_j; N, p_i, p_j) = \frac{N!}{k_i! k_j! (N-k_i-k_j)!} p_i^{k_i} \, p_j^{k_j} \, (1-p_i-p_j)^{N-k_i-k_j} \ , \tag{2.9}$$

so that the covariance $V_{ij} = \mathrm{cov}[k_i, k_j]$ is

$$\begin{aligned} V_{ij} &= \sum_{k_i=0}^{N} \sum_{k_j=0}^{N-k_i} (k_i - Np_i)(k_j - Np_j) \frac{N!}{k_i! k_j! (N-k_i-k_j)!} p_i^{k_i} \, p_j^{k_j} \, (1-p_i-p_j)^{N-k_i-k_j} \\ &= -Np_ip_j \end{aligned} \tag{2.10}$$

for $i \neq j$, otherwise $V_{ii} = \sigma_i^2 = N\,p_i\,(1-p_i)$.

An example of the multinomial distribution is the probability to obtain a particular result for a histogram constructed from N independent observations of a random variable, i.e. k_1 entries in bin 1, k_2 entries in bin 2, etc., with m bins and N total entries. Note from equation (2.10) that the number of entries in any two bins are negatively correlated. That is, if in N trials bin i contains a larger than average number of entries ($k_i > Np_i$) then the probability is increased that a different bin j will contain a smaller than average number.

2.2 Poisson Distribution

Consider the binomial distribution of Section 2.1 in the limit that N becomes very large, p becomes very small, but the product Np (i.e. the expectation value of the number of successes) remains some finite value λ. It can be shown that equation (2.2) leads in this limit to (see e.g. [Fro79, Bra92])

$$f(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \ , \tag{2.11}$$

which is called the Poisson distribution for the integer random variable k, where $k = 0, 1, \ldots, \infty$. The p.d.f. has one parameter, λ. Figure 2.3 shows the Poisson distribution for $\lambda = 2, 5, 10$.

The expectation value of the Poisson random variable k is

$$E[k] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \ , \tag{2.12}$$

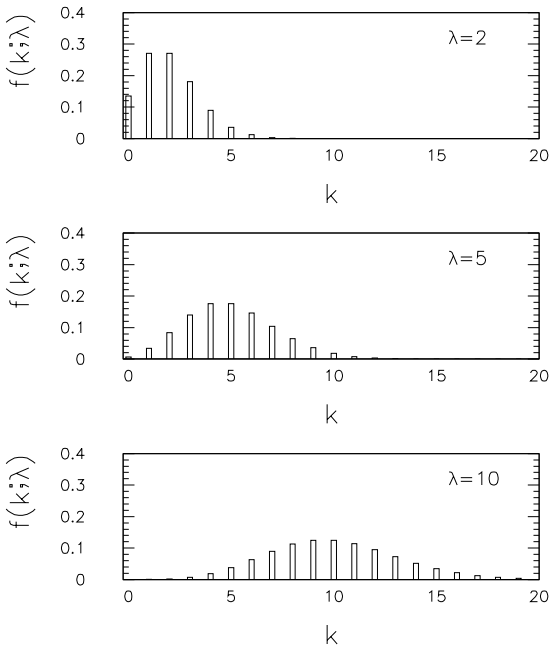

Figure 2.3: The Poisson probability distribution for various values of the parameter λ.

and the variance is given by

$$V[k] = \sum_{k=0}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \; . \tag{2.13}$$

An example of a Poisson distributed variable is the number of entries k in a bin of a histogram in the limit that the total number of entries, N is very large (and $k \ll N$), and providing that the individual entries are all independent. This is a useful approximation, since it allows one to estimate the variance of the number of entries in a bin directly from the number of entries. Another example of a Poisson random variable is the number of decays of a certain amount of radioactive material in a fixed time period, in the limit that the total number of possible decays (i.e. the total number of radioactive atoms) is very large and the probability for an individual decay within the time period is very small.

2.3 Uniform Distribution

The *uniform* p.d.f. for the continuous variable x ($-\infty < x < \infty$) is defined by

$$f(x;a,b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise,} \end{cases} \tag{2.14}$$

i.e. x is equally likely to be found anywhere between a and b. The mean and variance of x are given by

$$E[x] = \int_a^b \frac{x}{b-a}\,dx = \tfrac{1}{2}(a+b)\ , \tag{2.15}$$

$$V[x] = \int_a^b (x - \tfrac{1}{2}(a+b))^2\,\frac{1}{b-a}\,dx = \tfrac{1}{12}(b-a)^2\ . \tag{2.16}$$

The uniform distribution will be used frequently in Chapter 3 in connection with Monte Carlo techniques.

2.4 Exponential Distribution

The exponential probability density of the continuous variable x (with $0 \leq x < \infty$) is defined by

$$f(x;\xi) = \frac{1}{\xi}e^{-x/\xi}\ . \tag{2.17}$$

The p.d.f. is characterized by a single parameter ξ. The expectation value of x is

$$E[x] = \frac{1}{\xi}\int_0^\infty x e^{-x/\xi}dx = \xi\ , \tag{2.18}$$

and the variance of x is given by

$$V[x] = \frac{1}{\xi}\int_0^\infty (x-\xi)^2 e^{-x/\xi}dx = \xi^2\ . \tag{2.19}$$

An example of an exponential random variable is the decay time of an unstable particle measured in its rest frame. The parameter ξ then corresponds to the mean lifetime, usually denoted by τ. The exponential distribution is shown in Fig. 2.4 for different values of ξ.

2.5 Gaussian Distribution

The Gaussian (or normal) p.d.f. of the continuous random variable x (with $-\infty < x < \infty$) is defined by

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)\ , \tag{2.20}$$

which has two parameters, μ and σ^2. The names of the parameters are clearly motivated by the values of the mean and variance of x. These are found to be

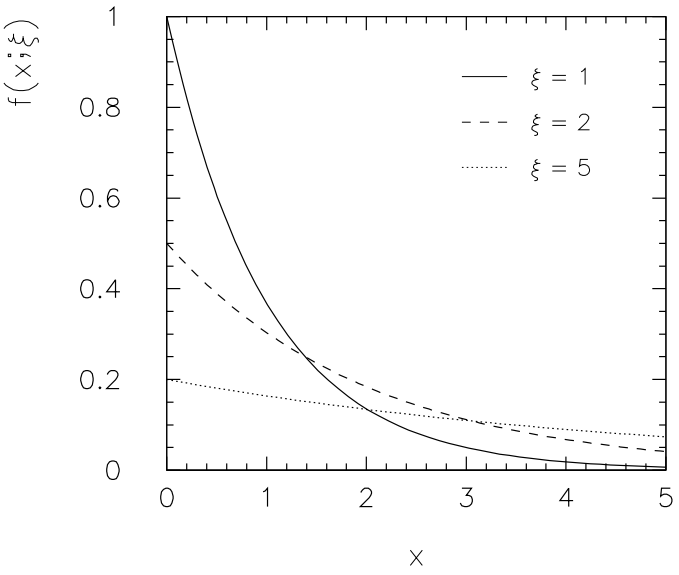

Figure 2.4: The exponential probability density for various values of the parameter ξ.

$$E[x] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx = \mu \ , \tag{2.21}$$

$$V[x] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx = \sigma^2 \ . \tag{2.22}$$

Recall that μ and σ^2 are often used to denote the mean and variance of any p.d.f. as defined by equations (1.37) and (1.41), not only those of a Gaussian. Note also that one may equivalently regard either σ or σ^2 as the parameter. The Gaussian p.d.f. is shown in Fig. 2.5 for different combinations of the parameters μ and σ.

A special case of the Gaussian p.d.f. is sufficiently important to merit its own notation. Using $\mu = 0$ and $\sigma = 1$ one defines the *standard Gaussian* p.d.f. $\varphi(x)$ as

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \ , \tag{2.23}$$

with the corresponding cumulative distribution $\Phi(x)$,

$$\Phi(x) = \int_{-\infty}^{x} \varphi(x')dx' \ . \tag{2.24}$$

One can easily show that if y is distributed according to a Gaussian p.d.f. with mean μ and variance σ^2, then the variable

$$x = \frac{y-\mu}{\sigma} \tag{2.25}$$

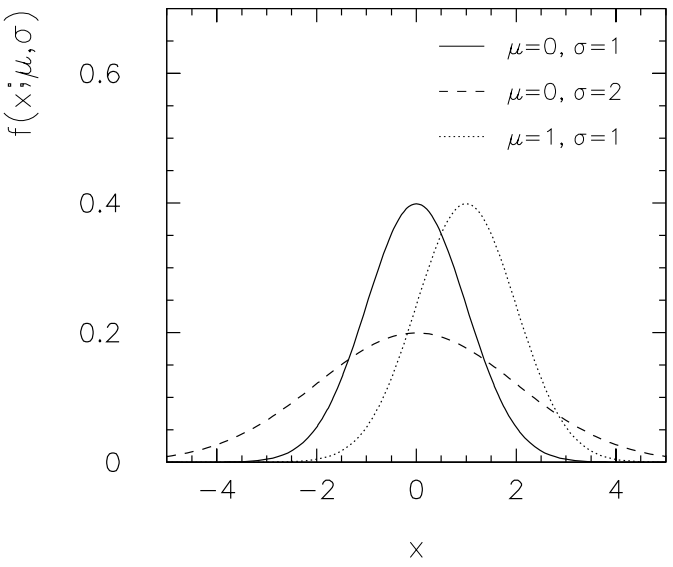

Figure 2.5: The Gaussian probability density for various values of the parameters μ and σ.

is distributed according to the standard Gaussian $\varphi(x)$, and the cumulative distributions are related by $F(y) = \Phi(x)$. The cumulative distribution $\Phi(x)$ cannot be expressed analytically and must be evaluated numerically. Values of $\Phi(x)$ as well as the quantiles $x_\alpha = \Phi^{-1}(\alpha)$ are tabulated in many reference books (e.g. [Bra92, Fro79, Dud88]) and are also available by means of computer routines [CER96].

The importance of the Gaussian distribution stems from the *Central Limit Theorem*. The theorem states that the sum of n independent continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \sum_{i=1}^n \mu_i$ and variance $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ in the limit that n approaches infinity. This holds (under fairly general conditions) regardless of the form of the individual p.d.f.'s of the x_i. This is the formal justification for treating measurement errors as Gaussian random variables, and holds to the extent that the total error is the sum of a large number of small contributions. The theorem can be proven using the Fourier transform techniques mentioned in Section 1.4; see e.g. [Bra92] Section 5.9.

The N-dimensional generalization of the Gaussian distribution is defined according to the following formula:

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})\right] , \tag{2.26}$$

where $\vec{x}$ and $\vec{\mu}$ are column vectors containing $x_1, \ldots, x_N$ and $\mu_1, \ldots, \mu_N$, $\vec{x}^T$ and $\vec{\mu}^T$ are the corresponding row vectors, and V is a symmetric $N \times N$ matrix, thus containing $N(N+1)/2$ free parameters. For now regard V as a label for the parameters of the Gaussian, although as with the one-dimensional case, the notation is motivated by what one obtains for the covariance matrix. The expectation values and (co)variances can be computed to be

$$\begin{aligned} E[x_i] &= \mu_i \\ V[x_i] &= V_{ii} \\ \mathrm{cov}[x_i, x_j] &= V_{ij} \ . \end{aligned} \tag{2.27}$$

For two dimensions the p.d.f. becomes

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right] \right\} , \tag{2.28}$$

where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient.

2.6 Chi-Square Distribution

The χ^2 (chi-square) distribution of the continuous variable z $(0 \leq z < \infty)$ is defined by

$$f(z; n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} \ , \ n = 1, 2, \ldots \ , \tag{2.29}$$

where the parameter n is called the number of degrees of freedom. The gamma function $\Gamma(x)$ is described e.g. in references [Arf70, Bra92].[1] The mean and variance of z are found to be

$$E[z] = \int_0^\infty z \ \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} = n \ , \tag{2.30}$$

$$V[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2} = 2n \ . \tag{2.31}$$

The χ^2-distribution is shown in Fig. 2.6 for several values of the parameter n.

The χ^2-distribution derives its importance from the following. Given N independent Gaussian random variables x_i with known mean μ_i and variance σ_i^2, it can be shown that the random variable

$$z = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \tag{2.32}$$

[1] For the purposes of computing the χ^2-distribution, one only needs to know that $\Gamma(n) = n!$ for integer n, $\Gamma(x+1) = x\Gamma(x)$, and $\Gamma(1/2) = \sqrt{\pi}$.

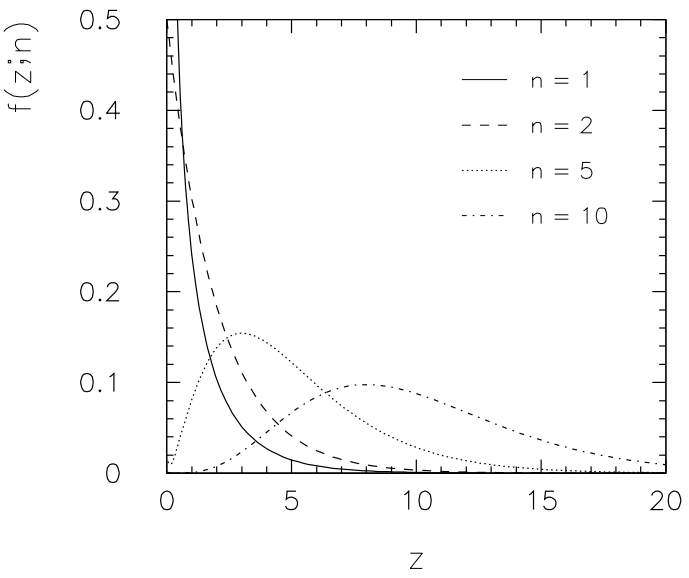

Figure 2.6: The χ^2 probability density for various values of the parameter n.

is distributed according to the χ^2-distribution for N degrees of freedom. (See e.g. [Fro79, Bra92].) More generally, if the x_i are not independent but are described by an N-dimensional Gaussian p.d.f. (equation (2.26)), the variable

$$z = (\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu}) \tag{2.33}$$

is a χ^2 random variable for N degrees of freedom. This and other similar examples will be discussed further in Chapter 7.

2.7 Cauchy (Breit-Wigner) Distribution

The Cauchy or Breit-Wigner p.d.f. of the continuous variable x $(-\infty < x < \infty)$ is defined by

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \, . \tag{2.34}$$

This is a special case of the Breit-Wigner distribution encountered in particle physics,

$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2} \, , \tag{2.35}$$

where the parameters x_0 and Γ correspond to the mass and width of a resonance particle. This is shown in Fig. 2.7 for several values of the parameters.

The expectation value of the Cauchy distribution is not well defined, since although the p.d.f. is symmetric about zero (or x_0 for (2.35)) the integrals $\int_{-\infty}^{0} x f(x) dx$ and $\int_{0}^{\infty} x f(x) dx$

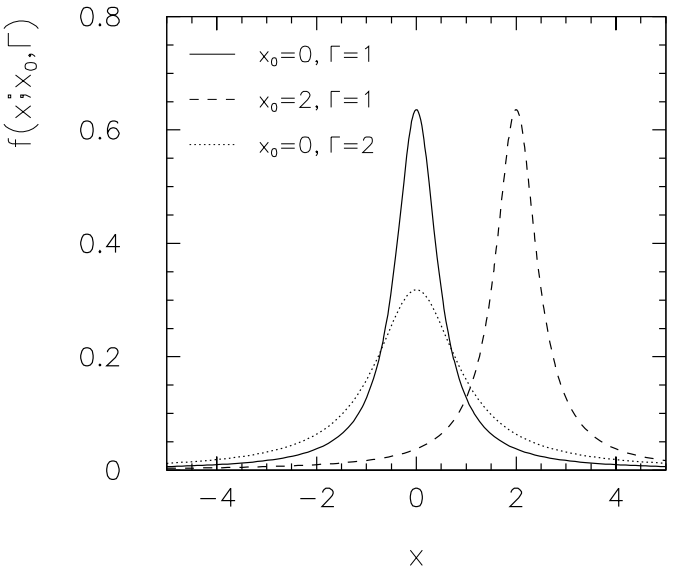

Figure 2.7: The Cauchy (Breit-Wigner) probability density for various values of the parameters x_0 and Γ.

are individually divergent. The variance and higher moments are also divergent. The parameters x_0 and Γ can nevertheless be used to give information about the position and width of the p.d.f., as can be seen from the figure; x_0 is the peak position (i.e. the most probable value, also called the *mode*) and Γ is the full-width of the peak at half of the maximum height.[2]

2.8 Landau Distribution

In nuclear and particle physics one often encounters the probability density $f(\Delta;\beta)$ for the energy loss Δ of a charged particle when traversing a layer of matter of a given thickness. This was first derived by Landau [Lan44], and is given by

$$f(\Delta;\beta) = \frac{1}{\xi}\,\phi(\lambda)\,, \quad 0 \leq \Delta < \infty\,, \tag{2.36}$$

where ξ is a parameter related to the properties of the material and the velocity of the particle $\beta = v/c$, (measured in units of the velocity of light c) and $\phi(\lambda)$ is the p.d.f. of the dimensionless random variable λ. The variable λ is related to the properties of the material, the velocity β, and the energy loss Δ. These quantities are given by

$$\xi = \frac{2\pi N_A e^4 z^2 \rho \sum Z}{m_e c^2 \sum A}\,\frac{d}{\beta^2}\,, \tag{2.37}$$

[2]The definition used here is standard in high energy physics where Γ is interpreted as the decay rate of a particle. In some references, e.g. [Ead71, Fro79], the parameter Γ is defined as the half-width at half maximum, i.e. the p.d.f. is given by equation (2.35) with the replacement $\Gamma \to 2\Gamma$.

$$\lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_E\right)\right] , \tag{2.38}$$

$$\epsilon' = \frac{I^2\,\exp(\beta^2)}{2m_e c^2\beta^2\gamma^2} , \tag{2.39}$$

where N_A is Avagadro's number, m_e and e are the mass and charge of the electron, z is the charge of the incident particle in units of the electron's charge, $\sum Z$ and $\sum A$ are the sums of the atomic numbers and atomic weights of the molecular substance, ρ is its density, d is the thickness of the layer, $I = I_0 Z$ with $I_0 \approx 13.5$ eV is an ionization energy characteristic of the material, $\gamma = 1/\sqrt{1-\beta^2}$, and $\gamma_E = 0.5772\ldots$ is Euler's constant. The function $\phi(\lambda)$ is given by

$$\phi(\lambda) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \exp(u\ln u + \lambda u)du , \tag{2.40}$$

where σ is infinitesimal and positive, or equivalently after a variable transformation by

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp[-u(\ln u + \lambda)]\,\sin\pi u\,du . \tag{2.41}$$

The integral must be evaluated numerically (see e.g. [Mac69], [CER96] routine G110). The energy loss distribution is shown in Fig. 2.8(a) for several values of the velocity $\beta = v/c$. Because of the long "Landau tail", the mean and higher moments of the Landau distribution do not exist, i.e. the integral $\int_0^\infty \Delta^n f(\Delta)d\Delta$ diverges for $n \geq 1$. As can be seen from the figure, however, the most probable value (mode) Δ_{mp} is sensitive to the particle's velocity. This has been computed numerically in [Mac69] to be

$$\Delta_{mp} = \xi\,[\ln(\xi/\epsilon') + 0.198] , \tag{2.42}$$

and is shown in Fig. 2.8(b).[3]

Although the mean and higher moments do not exist for the Breit-Wigner and Landau distributions, the probability densities actually describing physical processes must have finite moments. If, for example, one were to measure the energy loss Δ of a particle in a particular system many times, the average would eventually converge to some value, since Δ cannot exceed the energy of the incoming particle. Similarly, the mass of a resonance particle cannot be less than the sum of the rest masses of its decay products, and it cannot be more than the center-of-mass energy of the reaction in which it was created. The problem arises because the Cauchy and Landau distributions are only approximate models of the physical system. The models break down in the tails of the distributions, which is the part of the p.d.f. that causes the mean and higher moments to diverge.

[3]Equation (2.42) (the "Bethe-Bloch formula") forms the basis for identification of charged particles by measurement of ionization energy loss. An important effect not included here is the polarization of the medium, which leads to a saturation of the energy loss at high velocities (the density effect). See e.g. [All80].

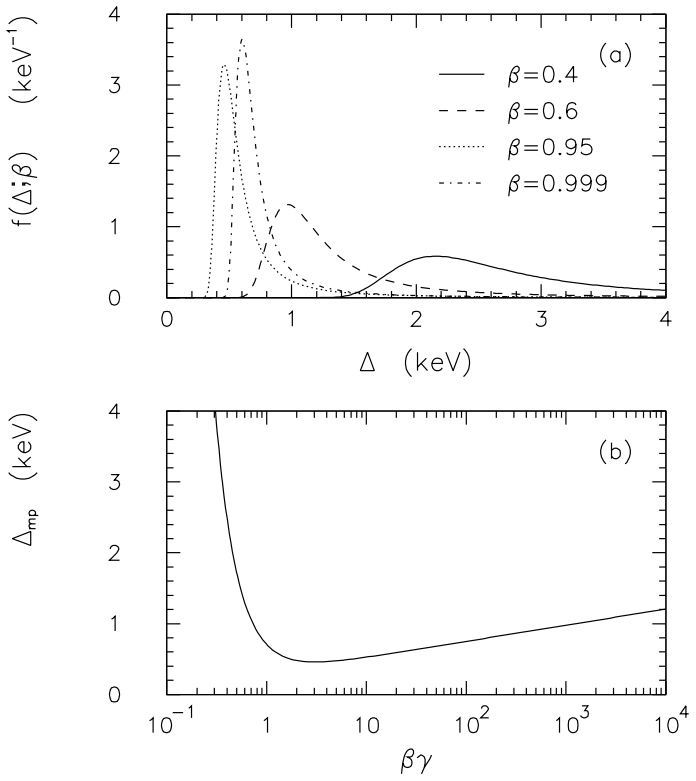

Figure 2.8: (a) The Landau probability density for the energy loss Δ of a charged particle traversing a 4 mm thick layer of argon gas for various values of the velocity β. (b) The peak position (mode) of the distributions in (a) as a function of $\beta\gamma$ as given by equation (2.42).

Chapter 3

The Monte Carlo Method

The Monte Carlo method is a numerical technique for calculating probabilities and related quantities by using sequences of random numbers generated according to known distributions. For the case of a single random variable, the procedure can be broken into the following stages. First, a series of random values $r_1, r_2 \ldots$ is generated according to a uniform distribution in the interval $0 < r < 1$. That is, the p.d.f. $g(r)$ is given by

$$g(r) = \begin{cases} 1 & 0 < r < 1 \,, \\ 0 & \text{otherwise} \,. \end{cases} \quad . \tag{3.1}$$

Next, the sequence $r_1, r_2, \ldots$ is used to determine another sequence $x_1, x_2 \ldots$ such that the x values are distributed according to a p.d.f. $f(x)$ in which one is interested. The values of x can then be treated as simulated measurements of a quantity x, and from them the probabilities for x to take on values in a certain region can be estimated. This may seem like a trivial exercise, since the function $f(x)$ was available to begin with, and could simply have been integrated over the region of interest. The true usefulness of the technique, however, becomes apparent in multidimensional problems, where integration of a joint p.d.f. $f(x, y, z, \ldots)$ over a complicated region of the sample space may not be feasible by other methods.

3.1 Uniformly Distributed Random Numbers

In order to generate a sequence of uniformly distributed random numbers one could in principle make use of a random physical process such as the repeated tossing of a coin. In practice, however, this task is almost always accomplished by a computer algorithm called a *random number generator*. Many such algorithms have been implemented as user-callable subprograms (e.g. the routine `RANMAR` in [CER96]) and are commonly available in computer program libraries. A detailed discussion of random number generators is beyond the scope of this book and the interested reader is referred to the more complete

treatments in [Bra92, Jam90]. Here a simple but effective algorithm will be presented in order to illustrate the general idea.

A commonly used type of random number generator is based on the so-called *multiplicative linear congruential* algorithm. Starting from an initial integer value n_0, one generates a sequence of integers $n_1, n_2, \ldots$ according to the rule,

$$n_{i+1} = a n_i \bmod m \; , \tag{3.2}$$

where the *multiplier* a and *modulus* m are integer constants and the mod (modulo) operator means that one takes the remainder of an_i divided by m. The values n_i follow a periodic sequence in the range $[1, m-1]$. In order to obtain values uniformly distributed in $[0, 1]$, one uses the transformation

$$r_i = n_i / m \; , \tag{3.3}$$

which excludes, however, the end-point values 0 and 1. (More sophisticated algorithms are able to overcome this minor defect.) The initial value n_0 (called the *seed*) and the two constants a and m determine the entire sequence, which, of course, is not truly random, but rather strictly determined. The resulting values are therefore more correctly called *pseudo-random*. For essentially all applications these can be treated as equivalent to true random numbers, with the exception of being reproducible, e.g. if one repeats the procedure with the same seed.

The values of m and a are chosen such that the generated numbers perform well with respect to various tests of randomness. Most important among these is a long period before the sequence repeats, since after this occurs the numbers can clearly no longer be regarded as random. In addition, one tries to attain the smallest possible correlations between pairs of generated numbers. For a 32-bit integer representation, for example, $m = 2147483647$ and $a = 39373$ have been shown to give good results, and with these one attains the maximum period of $m - 1 \approx 2 \times 10^9$ [Bra92, Lec88].

3.2 The Transformation Method

Given a sequence of random numbers $r_1, r_2, \ldots$ uniformly distributed in $[0, 1]$, the next step is to determine a sequence $x_1, x_2, \ldots$ distributed according to the p.d.f. $f(x)$ in which one is interested. In the *transformation method* this is accomplished by finding a suitable function $x(r)$ which directly yields the desired sequence when evaluated with the uniformly generated r values. The problem is clearly related to the transformation of variables discussed in section 1.4. There, an original p.d.f. $f(x)$ for a random variable x and a function $a(x)$ were specified, and the p.d.f. $g(a)$ for the function a was then found. Here the task is to find a function $x(r)$ that is distributed according to a specified $f(x)$, given that r follows a uniform distribution between 0 and 1.

The probability to obtain a value of r in the interval $[r, r+dr]$ is $g(r)dr$, and this should be equal to the probability to obtain a value of x in the corresponding interval $[x(r), x(r)+dx(r)]$, which is $f(x)dx$. In order to determine $x(r)$ such that this is true, one can require that the probability that r is less than some value r' be equal to the probability that x is less than $x(r')$. (We will see in the following example that this prescription is not unique.) Thus one must find a function $x(r)$ such that $F(x(r)) = G(r)$, where F and G are the cumulative distributions corresponding to the p.d.f.'s f and g. Since the cumulative distribution for the uniform p.d.f. is $G(r) = r$ with $0 \leq r \leq 1$, one has

$$\begin{aligned} F(x(r)) = \int_{-\infty}^{x(r)} f(x')dx' &= \int_{-\infty}^{r} g(r')dr' \\ &= r \end{aligned} \tag{3.4}$$

Equation (3.4) effectively says that the values of a cumulative distribution $F(x)$, treated as a random variable, are uniformly distributed between 0 and 1.

Depending on the $f(x)$ in question it may or may not be possible to solve for $x(r)$ using equation (3.4). Consider the exponential distribution discussed in section 2.4. Equation (3.4) becomes

$$\int_0^{x(r)} \frac{1}{\xi} e^{-x'/\xi} dx' = r \ . \tag{3.5}$$

Integrating and solving for x gives

$$x(r) = -\xi \log(1-r) \ . \tag{3.6}$$

If the variable r is uniformly distributed between 0 and 1 then $r' = 1 - r$ clearly is too, so that the function

$$x(r) = -\xi \log r \tag{3.7}$$

also has the desired property. That is, if r follows a uniform distribution between 0 and 1, then $x(r) = -\xi \log r$ will follow an exponential distribution between zero and infinity with mean ξ.

3.3 The Acceptance-Rejection Method

It turns out to be too difficult in many practical applications to solve equation (3.4) for $x(r)$ analytically. A useful alternative is the acceptance-rejection technique developed by von Neumann. Consider a p.d.f. $f(x)$ which can be completely surrounded by a box between x_{min} and x_{max} and having height f_{max}, as shown in Fig. 3.1. One can generate a series of numbers distributed according to $f(x)$ with the following algorithm:

(1) Generate a random number x, uniformly distributed between x_{min} and x_{max}. (i.e. $x = x_{min} + r_1(x_{max} - x_{min})$ where r_1 is uniformly distributed between 0 and 1.)

(2) Generate a second independent random number u uniformly distributed between 0 and f_{max}. (i.e. $u = r_2 f_{max}$.)

(3) If $u < f(x)$, then accept x. If not, reject x and repeat.

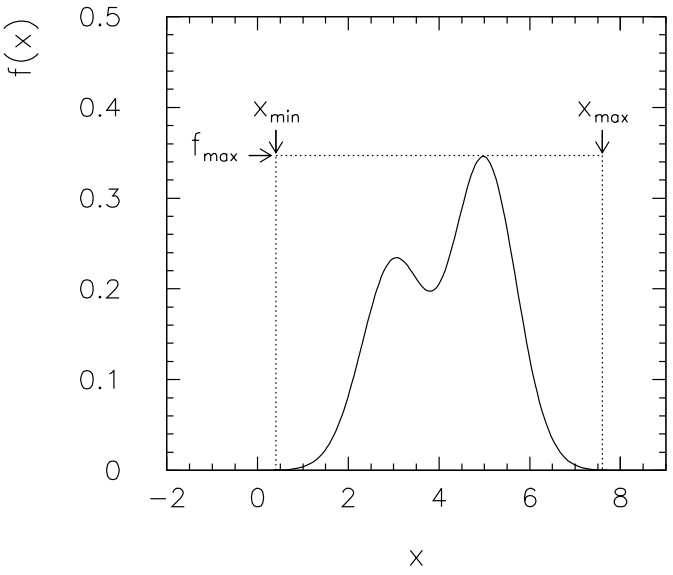

Figure 3.1: Probability density $f(x)$ enclosed by a box to generate random numbers using the acceptance-rejection technique.

The accepted x values will be distributed according to $f(x)$, since for each value of x obtained from step (1) above, the probability to be accepted is proportional to $f(x)$.

As an example consider the p.d.f.[1]

$$f(x) = \frac{3}{8}(1 + x^2)\,, \quad -1 \leq x \leq 1\,. \tag{3.8}$$

The p.d.f. has a maximum value at $x = \pm 1$ of $f_{max} = 3/4$. Figure 3.2(a) shows a scatter plot of the random numbers u and x generated according to the algorithm given above. The x values of the points that lie below the curve are accepted. Figure 3.2(b) shows a normalized histogram constructed from the accepted points.

The efficiency of the algorithm (i.e. the fraction of x values accepted) is the ratio of the areas of the p.d.f. (unity) to that of the enclosing box $f_{max} \cdot (x_{max} - x_{min})$. For a highly peaked density function this efficiency may be quite low, and the algorithm may be too slow to be practical. In cases such as these, one can improve the efficiency by enclosing the p.d.f. $f(x)$ in any other curve $g(x)$ for which random numbers can be generated according to $g(x)/\int g(x')dx'$, using e.g. the transformation method, equation (3.4). The more general algorithm is then:

[1]Equation (3.8) gives the distribution of the scattering angle θ in the reaction $e^+e^- \to \mu^+\mu^-$ with $x = \cos\theta$.

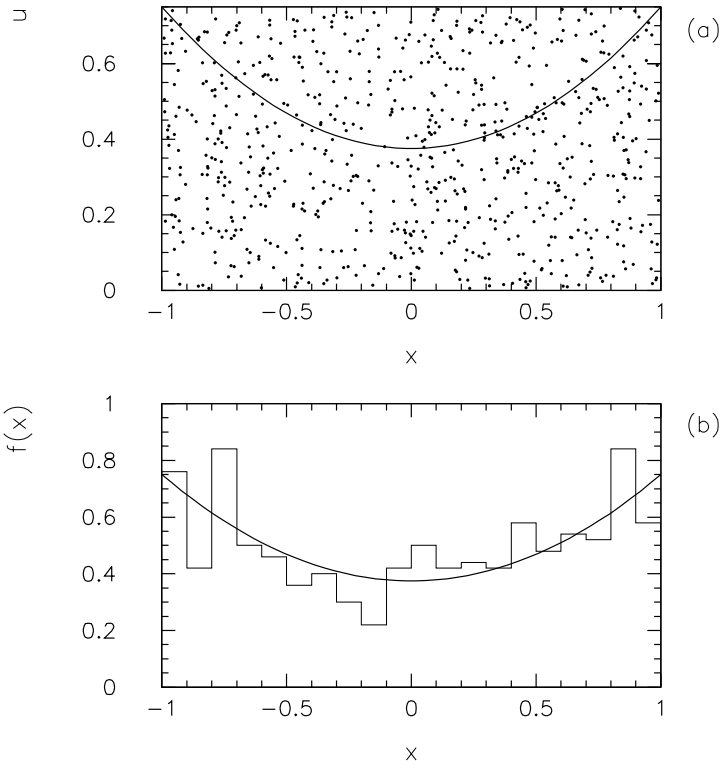

Figure 3.2: (a) Scatter plot of pairs of numbers (u, x), where x is uniformly distributed in $-1 \leq x \leq 1$, and u is uniform in $0 \leq u \leq f_{max}$. The x values of the points below the curve are accepted. (b) Normalized histogram of the accepted x values with the corresponding p.d.f.

(1) Generate a random number x according to the p.d.f. $g(x)/\int g(x')dx'$.

(2) Generate a second random number u uniformly distributed between 0 and $g(x)$.

(3) If $u < f(x)$, then accept x. If not, reject x and repeat.

Here the probability to generate a value x in step (1) is proportional to $g(x)$, and the probability to be retained after step (3) is equal to $f(x)/g(x)$, so that the total probability to obtain x is proportional to $f(x)$ as required.

3.4 Applications of the Monte Carlo Method

The Monte Carlo technique provides a method for determining the p.d.f.'s of functions of random variables. Suppose, for example, one has n independent random variables $x_1, \ldots, x_n$ distributed according to known p.d.f.'s $f_1(x_1), \ldots, f_n(x_n)$, and one would like to compute the p.d.f. $g(a)$ of some (possibly complicated) function $a(x_1, \ldots, x_n)$. The techniques described in section 1.4 are often only usable for relatively simple functions of a small number of variables. With the Monte Carlo method, a value for each x_i is generated according to the corresponding $f_i(x_i)$. The value of $a(x_1, \ldots, x_n)$ is then computed and recorded (e.g. in a histogram). The procedure is repeated until one has enough values of a to estimate the properties of its p.d.f. $g(a)$ (e.g. mean, variance) with

the desired statistical precision. Examples of this technique will be used in Chapters 6 through 8.

The Monte Carlo method is often used to simulate experimental data. In particle physics, for example, this is typically done in two stages: event generation and detector simulation. Consider for example an experiment in which an incoming particle such as an electron scatters off a target and is then detected. Suppose there exists a theory that predicts the probability for scattering to occur as a function of the scattering angle (i.e. the differential cross section). First one constructs a Monte Carlo computer program to generate values of the scattering angles and hence the momentum vectors of the final state particles. Such a program is called an *event generator*. In high energy physics, event generators are available to describe a wide variety of particle reactions.

The output of the event generator, i.e. the momentum vectors of the generated particles, are then used as input for a *detector simulation program*. Since the response of a detector to the passage of the scattered particles also involves random processes such as the production of ionization, multiple Coulomb scattering, etc., the detector simulation program is also implemented using the Monte Carlo method. Program packages such as GEANT [CER96] can be used to describe complicated detector configurations, and experimental collaborations typically spend considerable effort in achieving as complete a modelling of the detector as possible. This is especially important in order to optimize the detector's design for investigating certain physical processes before investing time and money in constructing the apparatus.

When the Monte Carlo method is used to simulate experimental data, one can most easily think of the procedure as a computer implementation of an intrinsically random process. Probabilities are naturally interpreted as relative frequencies of outcomes of a repeatable experiment, and the experiment is simply repeated many times on the computer. The Monte Carlo method can also be regarded, however, as providing a numerical solution to other problems involving probabilities, and the results are clearly independent of the probability interpretation. This is the case, for example, when the Monte Carlo method is used simply to carry out a transformation of variables or to compute integrals of p.d.f.'s.

Chapter 4

Statistical Tests

4.1 Hypotheses, Test Statistics, Significance Level, Power

In this chapter some basic concepts of statistical test theory are presented. Here the task is to make a statement about how well the observed data stand in agreement with given predicted probabilities, i.e. a *hypothesis*. The hypothesis under consideration is traditionally called the *null hypothesis*, H_0, which could specify, for example, a probability density $f(x)$ of a random variable x. If the hypothesis determines $f(x)$ uniquely it is said to be *simple*; if the form of the p.d.f. is defined but not the values of at least one free parameter θ, then $f(x;\theta)$ is called a *composite hypothesis*. In such cases the unknown parameter or parameters are estimated from the data using e.g. techniques discussed in Chapters 5 – 8. For now we will concentrate on simple hypotheses.

A statement about the validity of H_0 often involves a comparison with some *alternative* hypotheses, $H_1, H_2, \ldots$. Suppose one has n observations of a random variable x, $(x_1, \ldots, x_n)$, and a set of hypotheses, $H_0, H_1, \ldots$, each of which specifies a given p.d.f. $f(x|H_0), f(x|H_1), \ldots$[1]. In order to investigate the measure of agreement between the observed data and a given hypothesis, one constructs a function of the measured sample called a *test statistic* $T(x_1, \ldots, x_n)$. Each of the hypotheses will imply a given p.d.f. for the statistic T, i.e. $g(T|H_0)$, $g(T|H_1)$, etc.

The procedure for choosing the test statistic T depends in general on the hypotheses under consideration. In trying to distinguish between two hypotheses H_0 and H_1, the goal is clearly to construct T in such a way that the p.d.f.'s $g(T|H_0)$ and $g(T|H_1)$ overlap as little as possible. Procedures for constructing test statistics will be taken up again in sections 6.10 and 7.5. Let us suppose for the moment that we have chosen such a function

[1]For the p.d.f. of x given the hypothesis H the notation of conditional probability $f(x|H)$ is used (section 1.3), even though in the context of classical statistics a hypothesis H is only regarded as a random variable if it refers to the outcome of a repeatable experiment. In Bayesian statistics both x and H are random variables, so there the notation is in any event appropriate.

$T(x_1, \ldots, x_n)$, which would have the p.d.f. $g(T|H_0)$ if H_0 were true, and $g(T|H_1)$ if H_1 were true, as shown in Fig. 4.1.

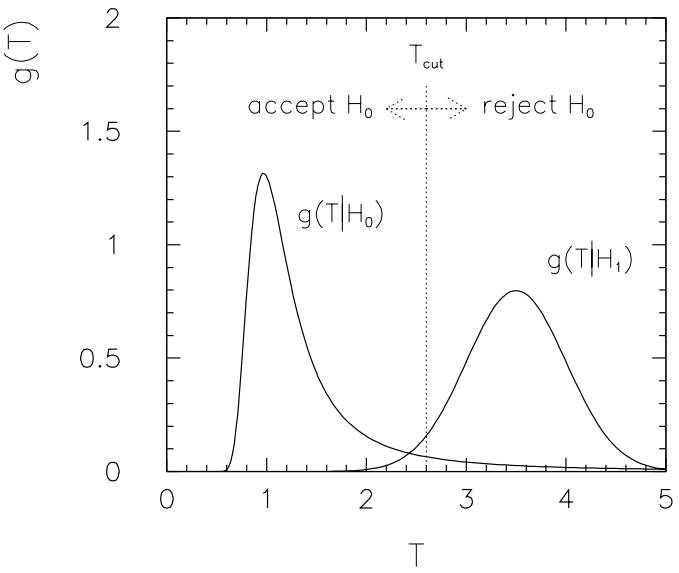

Figure 4.1: Probability densities for the test statistic T under assumption of the hypotheses H_0 and H_1. H_0 is rejected if T is observed in the critical region, here shown as $T > T_{cut}$.

Often one formulates the statement about the compatibility between the data and the various hypotheses in terms of a *decision* to accept or reject a given null hypothesis H_0. In practice this is done by defining a *critical region* for T. If the value of T actually observed is in the critical region, one rejects the hypothesis H_0; otherwise, H_0 is accepted. The critical region is chosen such that the probability for T to be observed there under assumption of the hypothesis H_0 is some value α, called the *significance level* of the test. For example, the critical region could consist of values of T greater than a certain cut-off T_{cut}, as shown in Fig. 4.1, so that

$$\alpha = \int_{T_{cut}}^{\infty} g(T|H_0)dT \ . \tag{4.1}$$

One would then decide that the hypothesis H_0 is true if the value of T observed is less than T_{cut}. The significance level α is thus the probability of rejecting H_0 if H_0 is true. This is called an *error of the first kind*. An *error of the second kind* takes place if the hypothesis H_0 is accepted (i.e. T is observed less than T_{cut}) but the true hypothesis was not H_0 but rather some alternative hypothesis H_1. The probability for this is

$$\beta = \int_{-\infty}^{T_{cut}} g(T|H_1)dT \ . \tag{4.2}$$

where $1-\beta$ is called the *power* of the test to discriminate against the alternative hypothesis H_1.

4.2 An Example with Particle Selection

As an example, the test statistic T could represent the measured ionization created by a charged particle of a known momentum traversing a detector. The amount of ionization is subject to fluctuations from particle to particle, and depends (for a fixed momentum) on the particle's mass. Thus the p.d.f. $g(T|H_0)$ in Fig. 4.1 could correspond to the hypothesis that the particle is an electron, and the $g(T|H_1)$ could be what one would obtain if the particle was a pion, i.e. $H_0 = e$, $H_1 = \pi$.

Suppose the particles in question are all known to be either electrons or pions, and that one would like to select a sample of electrons. (The electrons are regarded as "signal", and pions are considered as "background".) The probabilities to accept a particle of a given type, i.e. the efficiencies ϵ_e and ϵ_π, are thus

$$\epsilon_e = \int_{-\infty}^{T_{cut}} g(T|e)dT = 1 - \alpha \,, \tag{4.3}$$

$$\epsilon_\pi = \int_{-\infty}^{T_{cut}} g(T|\pi)dT = \beta \,. \tag{4.4}$$

Individually these can be made arbitrarily close to zero or unity simply by an appropriate choice of the critical region, (i.e. by making a looser or tighter cut on the ionization). The price one pays for a high efficiency for the signal is clearly an increased amount of contamination, i.e. the purity of the electron sample decreases because some pions are accepted as well.

If the relative fractions of pions and electrons are not known, the problem becomes one of parameter estimation (Chapters 5 – 8). That is, the test statistic T will be distributed according to $f(T; a_e) = a_e g(T|e) + a_\pi g(T|\pi)$, where a_e and $a_\pi = 1 - a_e$ are the fractions of electrons and pions, respectively. An estimate of a_e then gives the total number of electrons N_e in the original sample of N_{tot} particles, $N_e = a_e N_{tot}$.

Alternatively one may want to select a set of electron candidates by requiring $T < T_{cut}$, leading to N_{acc} accepted out of the N_{tot} particles. One is then often interested in determining the total number of electrons present before the cut on T was made. The number of accepted particles is given by

$$\begin{aligned} N_{acc} &= \epsilon_e N_e + \epsilon_\pi N_\pi \\ &= \epsilon_e N_e + \epsilon_\pi (N_{tot} - N_e) \,, \end{aligned} \tag{4.5}$$

which gives

$$N_e = \frac{N_{acc} - \epsilon_\pi N_{tot}}{\epsilon_e - \epsilon_\pi} \,. \tag{4.6}$$

From (4.6) one clearly sees that the number of accepted particles N_{acc} can only be used to determine the number of electrons N_e if the efficiencies ϵ_e and ϵ_π are different. If there are uncertainties in ϵ_π and ϵ_e, then these will translate into an uncertainty in N_e according to the error propagation techniques of section 1.6. One tries to select the critical region (i.e. the cut value for the ionization) in such a way that the total error in N_e is a minimum.

The probabilities that a particle with an observed value of T is an electron or a pion, $h(e|T)$ and $h(\pi|T)$, are obtained from the p.d.f.'s $g(T|e)$ and $g(T|\pi)$ using Bayes' theorem (1.8),

$$h(e|T) = \frac{a_e\, g(T|e)}{a_e\, g(T|e) \,+\, a_\pi\, g(T|\pi)}\,, \tag{4.7}$$

$$h(\pi|T) = \frac{a_\pi\, g(T|\pi)}{a_e\, g(T|e) \,+\, a_\pi\, g(T|\pi)}\,, \tag{4.8}$$

where a_e and $a_\pi = 1 - a_e$ are the *prior* probabilities for the hypotheses e and π. Thus in order to give the probability that a given selected particle is an electron one needs the prior probabilities for all of the possible hypotheses as well as the p.d.f.'s that they imply for the statistic T.

Although this is essentially the Bayesian approach to the problem, equations (4.7) and (4.8) also make sense in the framework of classical statistics. If one is dealing with a large sample of particles, then the hypotheses $H = e$ and $H = \pi$ refer to a characteristic that changes randomly from particle to particle. Using the relative frequency interpretation in this case, $h(e|T)$ gives the fraction of times a particle with a given T will be an electron. In Bayesian statistics using subjective probability, one would say that $h(e|T)$ gives the degree of belief that a given particle with a measured value of T is an electron.

Instead of the probability that an individual particle is an electron, one may be interested in the purity p_e of a sample of electron candidates selected by requiring $T < T_{cut}$. This is given by

$$\begin{aligned} p_e &= \frac{\text{number of electrons with } T < T_{cut}}{\text{number of all particles with } T < T_{cut}} \\ &= \frac{\int_{-\infty}^{T_{cut}} a_e g(T|e) dT}{\int_{-\infty}^{T_{cut}} (a_e g(T|e) + (1-a_e) g(T|\pi)) dT} \\ &= \frac{a_e \epsilon_e N_{tot}}{N_{acc}}\,. \end{aligned} \tag{4.9}$$

One can check using equation (4.7) that this is simply the mean electron probability $h(e|T)$, averaged over the interval $(-\infty, T_{cut}]$. That is,

$$p_e = \frac{\int_{-\infty}^{T_{cut}} h(e|T)\, f(T)\, dT}{\int_{-\infty}^{T_{cut}} f(T)\, dT}\,. \tag{4.10}$$

4.3 Goodness-of-Fit Tests

Frequently one wants to give a measure of how well a given null hypothesis H_0 is compatible with the observed data without specific reference to any alternative hypothesis. This is called a test of the *goodness-of-fit*, and can be done by constructing a test statistic whose value itself reflects the level of agreement between the observed measurements and the predictions of H_0. Procedures for constructing appropriate test statistics will be given in sections 6.10 and 7.5. For now we will give a short example to illustrate the main idea.

Suppose one tosses a coin N times and obtains n_h heads and $n_t = N - n_h$ tails. To what extent are n_h and n_t consistent with the hypothesis that the coin is "fair", i.e. that the probabilities for heads and tails are equal? As a test statistic one can simply use the number of heads n_h, which for a fair coin is assumed to follow a binomial distribution (equation (2.2)) with the parameter $p = 0.5$. That is,

$$f(n_h; N) = \frac{N!}{n_h!(N-n_h)!} \left(\frac{1}{2}\right)^{n_h} \left(\frac{1}{2}\right)^{N-n_h} . \tag{4.11}$$

Suppose that $N = 20$ tosses are made and $n_h = 17$ heads are observed. Since the expectation value of n_h (equation (2.5)) is $E[n_h] = Np = 10$, there is evidently a sizable discrepancy between the expected and actually observed outcomes. In order to quantify the significance of the difference one can give the probability of obtaining a result with the same level of discrepancy with the hypothesis or higher. In this case, this is the sum of the probabilities for $n_h = 0, 1, 2, 3, 17, 18, 19, 20$. Using equation (4.11) one obtains the probability $P = 0.0026$.

The result of the goodness-of-fit test is thus given by stating the so-called P-value, i.e. the probability P, under assumption of the hypothesis in question H_0, of obtaining a result as compatible or less with H_0 than the one actually observed. Equivalently one often gives the *confidence level* $CL = 1 - P$. In the classical approach one stops here, and does not attempt to give a probability for H_0 to be true, since a hypothesis is not treated as a random variable. (The confidence level is, however, often incorrectly interpreted as such a probability.) In Bayesian statistics one would use Bayes' theorem (1.6) to assign a probability to H_0, but this requires giving a prior probability, i.e. the probability that the coin is fair before having seen the outcome of the experiment. In some cases this is a practical approach, in others not. For the present we will remain within the classical framework and simply give the confidence level or the P-value.

The P-value is thus the fraction of times one would obtain data as compatible with H_0 or less so if the experiment (i.e. 20 coin tosses) were repeated many times under similar circumstances. By "similar circumstances" one means *always with 20 tosses*, or in general with the same number of observations in each experiment. Suppose the experiment had been designed to toss the coin until at least three heads and three tails were observed and then to stop, and in the real experiment this happened to occur after the 20th toss. Assuming such a design, one can show that the probability to stop after the 20th toss or later (i.e. to have an outcome as compatible or less with H_0) is not 0.26% but rather

0.072%, which would seem to lead to a significantly different conclusion about the validity of H_0. Maybe we do not even know how the experimenter decided to toss the coin; we are merely presented with the results afterwards. One way to avoid difficulties with the so-called *optional stopping* problem is simply to interpret the phrase "similar experiments" to always mean experiments with the same number of observations. Although this is an arbitrary convention, it allows for a unique interpretation of a reported confidence level. For further discussion of this problem see [Ber88, Oha94].

In the example with the coin tosses, the test statistic $T = n_h$ was reasonable since from the symmetry of the problem it was easy to identify the region of values of T that have an equal or lesser degree of compatibility with the hypothesis than the observed value. This is related to the fact that in the case of the coin, the set of all possible alternative hypotheses consists simply of all values of the parameter p not equal to 0.5, and all of these lead to an expected asymmetry between the number of heads and tails.

4.4 The Significance of a Peak

In preparation.

Chapter 5

General Concepts of Parameter Estimation

In this chapter some general concepts of parameter estimation are examined which apply to all of the methods discussed in Chapters 6 through 8. In addition, prescriptions for estimating properties of p.d.f.'s such as the mean and variance are given.

5.1 Samples, Estimators, Bias

Consider a random variable x described by a p.d.f. $f(x)$. Here, the sample space is the set of all possible values of x. A set of n independent observations of x is called a *sample* of size n. A new sample space can be defined as the set of all possible values for the n-dimensional vector $\vec{x} = (x_1, \ldots, x_n)$. That is, the entire experiment consisting of n measurements is considered to be a single random measurement, which is characterized by n numerical quantities, $x_1, \ldots, x_n$. Since it is assumed that the observations are all independent and that each x_i is described by the same p.d.f. $f(x)$, the joint p.d.f. for the sample $f_{sample}(x_1, \ldots, x_n)$ is given by

$$f_{sample}(x_1, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n) \; . \tag{5.1}$$

Although the dimension of the random vector (i.e. the number of measurements) can in practice be very large, the situation is greatly simplified by the fact that the joint p.d.f. for the sample is the product of n p.d.f.'s of identical form.

Consider now the situation where one has made n measurements of a random variable x, whose p.d.f. $f(x)$ is not known. The central problem of statistics is to infer the properties of $f(x)$ based on the observations $x_1, \ldots, x_n$. Specifically, one would like to construct functions of the x_i to estimate the various properties of the p.d.f. $f(x)$. Often one has a hypothesis for the p.d.f. $f(x; \theta)$ which depends on an unknown parameter θ (or parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$). One then tries to construct a function of the observed x_i to estimate the parameters.

A function of the observed measurements $x_1, \ldots, x_n$ which contains no unknown parameters is called a *statistic*. In particular, a statistic used to estimate some property of a p.d.f. (e.g. its mean, variance or other parameters) is called an *estimator*. The estimator for a quantity θ is usually written with a hat, $\hat{\theta}$, to distinguish it from the true value θ whose exact value is (and may forever remain) unknown.

If $\hat{\theta}$ converges to θ in the limit of large n, the estimator is said to be *consistent*. This is usually a minimum requirement for any useful estimator. In the following the limit of large n will be referred to as either the "large sample" or "asymptotic" limit, and is also sometimes called the "high statistics" limit. In situations where it is necessary to make the distinction, the term estimator will be used to refer to the function of the sample (i.e. its functional form) and an *estimate* will mean the value of the estimator evaluated with a particular sample. The procedure of estimating a parameter's value given the data $x_1, \ldots, x_n$ is usually called *parameter fitting*.

Since an estimator $\hat{\theta}(x_1, \ldots, x_n)$ is a function of the measured values, it is itself a random variable. That is, if the entire experiment were repeated many times, each time with a new sample $x_1, \ldots, x_n$ of size n, the estimator $\hat{\theta}(x_1, \ldots, x_n)$ would take on different values, being distributed according to some p.d.f. $g(\hat{\theta}; \theta)$, which depends in general on the true value of θ. The probability distribution of a statistic is called a *sampling distribution*. Much of what follows in the next several chapters concerns sampling distributions and their properties, especially expectation value and variance.

The expectation value of an estimator $\hat{\theta}$ with the sampling p.d.f. $g(\hat{\theta}; \theta)$ is

$$\begin{aligned} E[\hat{\theta}(x_1, \ldots, x_n)] &= \int \hat{\theta} g(\hat{\theta}; \theta) d\hat{\theta} \\ &= \int \cdots \int \hat{\theta}(x_1, \ldots, x_n) f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n \ , \end{aligned} \tag{5.2}$$

where equation (5.1) has been used for the joint p.d.f. of the sample. Recall that this is the expected mean value of $\hat{\theta}$ from an infinite number of similar experiments, each with a sample of size n. One defines the *bias* of an estimator $\hat{\theta}$ as

$$b = E[\hat{\theta}] - \theta \ . \tag{5.3}$$

Note that the bias does not depend on the measured values of the sample but rather on the sample size, the functional form of the estimator and on the true (and in general unknown) properties of the p.d.f. $f(x)$, including the true value of θ. A parameter for which the bias is zero independent of the sample size n is said to be unbiased; if the bias vanishes in the limit $n \to \infty$ then it is said to be asymptotically unbiased. Note also that an estimator $\hat{\theta}$ can be biased even if it is consistent. That is, even if $\hat{\theta}$ converges to the true value θ in a single experiment with an infinitely large number of measurements, it does not follow that the average of $\hat{\theta}$ from an infinite number of experiments, each with a finite number of measurements, will converge to the true value. Unbiased estimators are thus particularly valuable if one would like to combine the result with those of other

experiments. In most practical cases, the bias is small compared to the statistical error (i.e. the standard deviation) and one does not usually reject using an estimator with a small bias if there are other characteristics (e.g. simplicity) in its favor.

5.2 Estimators for Mean, Variance, Covariance

Suppose one has a sample of size n of a random variable x: $x_1, \ldots, x_n$. It is assumed that x is distributed according to some p.d.f. $f(x)$ which is not known, not even as a parameterization. We would like to construct a function of the x_i to be an estimator for the expectation value of x, μ. One possibility is the arithmetic mean of the x_i, defined by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{5.4}$$

The arithmetic mean of the elements of a sample is called the *sample mean*, and is denoted by a bar, e.g. $\overline{x}$. This should not be confused with the expectation value (population mean) of x, denoted by μ or $E[x]$, for which $\overline{x}$ is an estimator.

The expectation value of the sample mean $E[\overline{x}]$ is given by (see equation (5.2))

$$E[\overline{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \ , \tag{5.5}$$

since

$$E[x_i] = \int \cdots \int x_i f(x_1) \cdots f(x_n) dx_1 \cdots dx_n = \mu \tag{5.6}$$

for all i. One sees from equation (5.5) that the sample mean $\overline{x}$ is an unbiased estimator for the population mean μ.

Consider again a sample of size n of a random variable x from some unknown p.d.f. $f(x)$. The *sample variance* s^2 is defined by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ , \tag{5.7}$$

By computing the expectation value of s^2 as done with $\overline{x}$ one can show that the sample variance is an unbiased estimator for the population variance, σ^2. Similarly, one finds that the statistic S^2 defined by

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{5.8}$$

is an unbiased estimator of σ^2 for a p.d.f. $f(x)$ with known mean, μ, and the quantity

$$\widehat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{5.9}$$

is an unbiased estimator for the covariance V_{xy} of two random variables x and y of unknown mean. This can be normalized by the square root of the estimators for the sample variance to form an estimator r_{xy} for the correlation coefficient ρ_{xy} (see equation (1.46)):

$$r_{xy} = \frac{\widehat{V}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2\right)^{1/2}} \,. \tag{5.10}$$

Given an estimator $\hat{\theta}$ one can compute its variance $V[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$. Recall that $V[\hat{\theta}]$ (or equivalently its square root $\sigma_{\hat{\theta}}$) is a measure of the expected dispersion of $\hat{\theta}$ about its mean in a large number of similar experiments each with sample size n, and as such is often quoted as the statistical error of $\hat{\theta}$. For example, the variance of the sample mean $\overline{x}$ is

$$\begin{aligned} V[\overline{x}] &= E[\overline{x}^2] - (E[\overline{x}])^2 = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)\right] - \mu^2 \\ &= \frac{1}{n^2}\sum_{i,j=1,}^{n} E[x_i x_j] - \mu^2 \\ &= \frac{1}{n^2}\left[(n^2 - n)\mu^2 + n(\mu^2 + \sigma^2)\right] - \mu^2 = \frac{\sigma^2}{n} \,, \end{aligned} \tag{5.11}$$

where σ^2 is the variance of $f(x)$, and we have used the fact that $E[x_i x_j] = \mu^2$ for $i \neq j$ and $E[x_i^2] = \mu^2 + \sigma^2$. This expresses the well known result that the standard deviation of the mean of n measurements of x is equal to the standard deviation of $f(x)$ itself divided by $\sqrt{n}$.

The expectation value and variance of the estimator of the correlation coefficient r_{xy} depend on higher moments of the joint p.d.f. $f(x, y)$. For the case of the two-dimensional Gaussian p.d.f. (2.28) they are found to be (see [Mui82] and references therein)

$$E[r] = \rho - \frac{\rho(1 - \rho^2)}{2n} + O(n^{-2}) \tag{5.12}$$

$$V[r] = \frac{1}{n}(1 - \rho^2)^2 + O(n^{-2}) \,. \tag{5.13}$$

Although the estimator r given by equation (5.10) is only asymptotically unbiased, it is nevertheless widely used because of its simplicity. Note that although $\widehat{V}_{xy}$, s_x^2 and s_y^2 are unbiased estimators of V_{xy}, σ_x^2 and σ_y^2, the nonlinear function $\widehat{V}_{xy}/(s_x s_y)$ is not an unbiased estimator of $V_{xy}/(\sigma_x \sigma_y)$; cf. Section 6.2. Caution should be used when applying equation (5.13) to estimate the significance of an observed correlation; see Section 9.5.

Chapter 6

The Method of Maximum Likelihood

6.1 ML Estimators

Suppose a measurement of a certain random variable x has been repeated n times, yielding the values $(x_1, \ldots, x_n)$. (Here, x could also represent a multidimensional random vector, i.e. the outcome of each individual observation could be characterized by several quantities.) Suppose, in addition, one has a composite hypothesis for the underlying probability density function in the form of a parameterization, $f(x;\theta)$, where θ represents one or possibly several unknown parameters. The task is to estimate the values of the parameters. (This is in contrast to Chapter 5 where no parameterization of the unknown p.d.f. was assumed.)

Under the assumption of the hypothesis $f(x;\theta)$, including the value of θ, the probability for the first measurement to be in the interval $[x_1, x_1 + dx_1]$ is $f(x_1;\theta)dx_1$. Since the measurements are all assumed to be independent, the probability to have the first one in $[x_1, x_1 + dx_1]$, the second in $[x_2, x_2 + dx_2]$, and so forth is given by

$$\textit{Probability that } x_i \textit{ in } [x_i, x_i + dx_i] \textit{ for all } i \; = \prod_{i=1}^{n} f(x_i;\theta)dx_i \; . \tag{6.1}$$

If the hypothesized p.d.f. and parameter values are correct, one expects a high probability for the data that were actually measured. Conversely, a parameter value far away from the true value should yield a low probability for the measurements obtained. Since the dx_i do not depend on the parameters, the same reasoning also applies to the following function L,

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{6.2}$$

called the likelihood function. Note that this is just the joint p.d.f. for the x_i, although it is treated here as a function of the parameter, θ. The x_i, on the other hand, are treated as fixed (i.e. the experiment is over).

With this motivation one defines the maximum likelihood (ML) estimators for the parameters to be those which maximize the likelihood function. That is, for m parameters $\theta_1, \ldots, \theta_m$, one has the following set of m equations,

$$\frac{\partial L(\theta_1, \ldots, \theta_m)}{\partial \theta_i} = 0, \; i = 1, \ldots, m \; . \tag{6.3}$$

The equations can be solved for $\theta_1, \ldots, \theta_m$, which are called the ML estimators. If more than one maximum exists, the highest one is taken. As with other types of estimators, they are usually written with hats, $\hat{\theta}_1, \ldots, \hat{\theta}_m$, to distinguish them from the true parameters θ_i whose exact values remain unknown.

The general idea of maximum likelihood is illustrated in Fig. 6.1. A sample of 50 measurements (shown as tick marks on the horizontal axis) was generated according to a Gaussian p.d.f. with parameters $\mu = 0.2$, $\sigma = 0.1$. The solid curve in Fig. 6.1(a) was computed using the parameter values for which the likelihood function (and hence also its logarithm) are a maximum: $\hat{\mu} = 0.204$ and $\hat{\sigma} = 0.106$. Also shown as a dashed curve is the p.d.f. using the true parameter values. As is unavoidable because of random fluctuations, the estimates $\hat{\mu}$ and $\hat{\sigma}$ are not exactly equal to the true values μ and σ. The estimators $\hat{\mu}$ and $\hat{\sigma}$ and their variances, which reflect the size of the statistical errors, are derived below in Section 6.3. Figure 6.1(b) shows the p.d.f. for parameters far away from the true values, leading to lower values of the likelihood function.

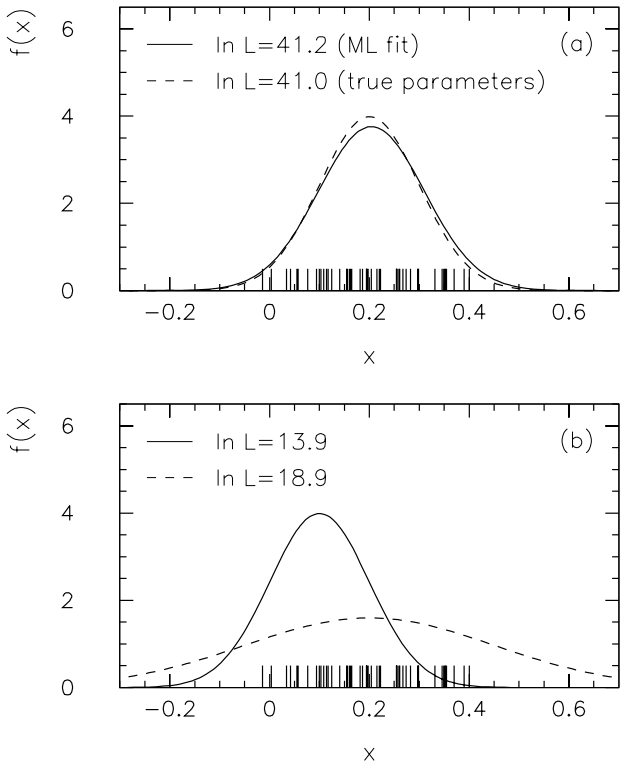

Figure 6.1: A sample of 50 observations of a Gaussian random variable with mean $\mu = 0.2$ and standard deviation $\sigma = 0.1$. (a) The p.d.f. evaluated with the parameters that maximize the likelihood function and with the true parameters. (b) The p.d.f. evaluated with parameters far from the true values, giving a lower likelihood.

The motivation for the ML principle presented above does not necessarily guaranty

any optimal properties for the resulting estimators. The ML method turns out to have many advantages, among them ease of use and the fact that all of the available information from the data is used (i.e. no binning is required). In the following the desirability of ML estimators will be investigated with respect to several criteria, most importantly variance and bias.

6.2 Example of ML Estimator: an Exponential Distribution

Suppose the proper decay times for unstable particles of a certain type have been measured for n decays, yielding values $t_i, \ldots, t_n$, and suppose one chooses as a hypothesis for the distribution of t an exponential p.d.f. with mean τ:

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \; . \tag{6.4}$$

The task here is to estimate the value of the parameter τ. Rather than using the likelihood function as defined in equation (6.2) it is usually more convenient to use its logarithm. From the condition

$$\left.\frac{\partial \log L}{\partial \tau}\right|_{\tau=\hat{\tau}} = \left.\frac{1}{L}\frac{\partial L}{\partial \tau}\right|_{\tau=\hat{\tau}} = 0 \tag{6.5}$$

one obtains the same estimators as from maximizing L, with the advantage that the product in L is converted into a sum, and exponentials in f are converted into simple factors. The *log-likelihood function* is then

$$\log L(\tau) = \sum_{i=1}^{n} \log f(t_i;\tau) = \sum_{i=1}^{n} \left(\log \frac{1}{\tau} - \frac{t_i}{\tau} \right) \; . \tag{6.6}$$

Maximizing $\log L$ with equation (6.5) gives the ML estimator $\hat{\tau}$,

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i \; . \tag{6.7}$$

In this case the ML estimator $\hat{\tau}$ is simply the sample mean of the measured time values. The expectation value of $\hat{\tau}$ is

$$\begin{aligned} E[\hat{\tau}(t_1,\ldots,t_n)] &= \int \cdots \int \hat{\tau}(t_1,\ldots,t_n)\, f_{joint}(t_1,\ldots,t_n;\tau)\, dt_1 \cdots dt_n \\ &= \int \cdots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \cdots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \cdots dt_n \end{aligned} \tag{6.8}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\int t_i \frac{1}{\tau}e^{-t_i/\tau}dt_i \prod_{j\neq i}\int \frac{1}{\tau}e^{-t_j/\tau}dt_j\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\tau = \tau \ ,$$

so $\hat{\tau}$ is an unbiased estimator for τ. We could have concluded this from the results of Sections 2.4 and 5.2, where it was seen that τ is the expectation value of the exponential p.d.f., and that the sample mean is an unbiased estimator of the expectation value for any p.d.f. (See Chapter 11 for a derivation of the p.d.f. for $\hat{\tau}$.)

As an example consider a sample of 50 Monte Carlo generated decay times t distributed according to an exponential p.d.f. as shown in Fig. 6.2. The values were generated using a true lifetime $\tau = 1.0$. Equation (6.7) gives the ML estimate $\hat{\tau} = 1.062$. The curve shows the exponential p.d.f. evaluated with the ML estimate.

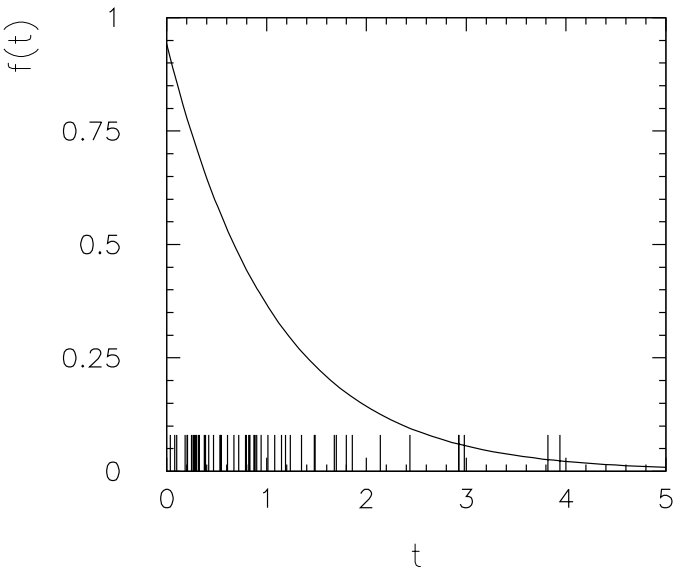

Figure 6.2: A sample of 100 Monte Carlo generated observations of an exponential random variable t with mean $\tau = 1.0$. The curve is the result of a maximum likelihood fit, giving $\hat{\tau} = 1.062$.

Suppose that one is interested not in the mean lifetime but in the decay constant $\lambda = 1/\tau$. How can we estimate λ? In general, given a function $a(\theta)$ of some parameter θ, one has

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial a}\frac{\partial a}{\partial \theta} = 0 \, . \tag{6.9}$$

Thus $\partial L/\partial\theta = 0$ implies $\partial L/\partial a = 0$ at $a = a(\theta)$ unless $\partial a/\partial\theta = 0$. As long as this is not the case, one obtains the ML estimator of a function simply by evaluating the function with the original ML estimator, i.e. $\hat{a} = a(\hat{\theta})$. The estimator for the decay constant is thus $\hat{\lambda} = 1/\hat{\tau} = n/\sum_{i=1}^{n} t_i$. The transformation invariance of ML estimators is a convenient property, but an unbiased estimator does not necessarily remain so under transformation. As will be derived in Section 10.1, the expectation value of $\hat{\lambda}$ is

$$E[\hat{\lambda}] = \lambda \frac{n}{n-1} = \frac{1}{\tau} \frac{n}{n-1} , \tag{6.10}$$

so $\hat{\lambda} = 1/\hat{\tau}$ is an unbiased estimator of $1/\tau$ only in the limit of large n, even though $\hat{\tau}$ is an unbiased estimator for τ for any value of n. To summarize, the ML estimator of a function a of a parameter θ is simply $\hat{a} = a(\hat{\theta})$. But if $\hat{\theta}$ is an unbiased estimator of θ ($E[\hat{\theta}] = \theta$) it does not necessarily follow that $a(\hat{\theta})$ is an unbiased estimator of $a(\theta)$. It can be shown, however, that the bias of ML estimators goes to zero in the large sample limit for essentially all practical cases. (An exception to this rule occurs if the allowed range of the random variable depends on the parameter; see [Ead71] Section 8.3.3.)

6.3 Example of ML estimators: Gaussian of Unknown μ and σ^2

Suppose one has n measurements of a random variable x assumed to be distributed according to a Gaussian p.d.f. of unknown μ and σ^2. The log-likelihood function is

$$\log L(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i; \mu, \sigma^2) = \sum_{i=1}^{n} \left(\log \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \log \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) . \tag{6.11}$$

Setting the derivative of $\log L$ with respect to μ equal to zero and solving gives

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i . \tag{6.12}$$

Computing the expectation value as done in equation (6.8) gives $E[\hat{\mu}] = \mu$, so $\hat{\mu}$ is unbiased. (As in the case of the mean lifetime estimator $\hat{\tau}$, $\hat{\mu}$ here happens to be a sample mean, so one knows already from Sections 2.5 and 5.2 that it is an unbiased estimator for the mean μ.) Repeating the procedure for σ^2 and using the result for $\hat{\mu}$ gives

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 . \tag{6.13}$$

Computing the expectation value of $\widehat{\sigma^2}$, however, gives

$$E[\widehat{\sigma^2}] = \frac{n-1}{n} \sigma^2 . \tag{6.14}$$

The ML estimator $\widehat{\sigma^2}$ is thus biased, but the bias vanishes in the limit of large n.

Recall, however, from Section 5.1 that the sample variance s^2 is an unbiased estimator for the variance of any p.d.f., so that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{6.15}$$

is an unbiased estimator for the parameter σ^2 of the Gaussian. To summarize, equation (6.13) gives the ML estimator for the parameter σ^2, and it has a bias that goes to zero as n approaches infinity. The statistic s^2 from equation (6.15) is not biased (which is good) but it is not the ML estimator.

6.4 Variance of ML Estimators: Analytic Method

Given a set of n measurements of a random variable x and a hypothesis for the p.d.f. $f(x;\theta)$ we have seen how to estimate its parameters. The next question is, what are the statistical errors on the estimates? That is, if we repeated the entire experiment a large number of times (with n measurements each time) each experiment would give different estimated values for the parameters. How widely spread will they be? One way of summarizing this is with the variance (or standard deviation) of the estimator.

For certain cases one can compute the variances of the ML estimators analytically. For the example of the exponential distribution with mean τ estimated by $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$, one has

$$\begin{aligned} V[\hat{\tau}] &= E[\hat{\tau}^2] - (E[\hat{\tau}])^2 \\ &= \int\cdots\int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \frac{1}{\tau}e^{-t_1/\tau}\cdots\frac{1}{\tau}e^{-t_n/\tau}dt_1\cdots dt_n - \\ &\quad \left(\int\cdots\int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)\frac{1}{\tau}e^{-t_1/\tau}\cdots\frac{1}{\tau}e^{-t_n/\tau}dt_1\cdots dt_n\right)^2 \\ &= \frac{\tau^2}{n}\,, \end{aligned} \tag{6.16}$$

This could have been guessed, since it was seen in Section 5.2 that the variance of the sample mean is $1/n$ times the variance of the p.d.f. of t (the time of an individual measurement), for which in this case the variance is τ^2, (Section 2.4) and the estimator $\hat{\tau}$ happens to be the sample mean.

Remember that the variance of $\hat{\tau}$ computed in equation (6.16) is a function of the true (and unknown) parameter τ. So what do we report for the statistical error of the experiment? Because of the transformation invariance of ML estimators (equation (6.9)) we can obtain the ML estimate for the variance $\sigma^2_{\hat{\tau}} = \tau^2/n$ simply by replacing τ with its own ML estimator $\hat{\tau}$, giving $\widehat{\sigma^2}_{\hat{\tau}} = \hat{\tau}^2/n$, or similarly for the standard deviation, $\hat{\sigma}_{\hat{\tau}} = \hat{\tau}/\sqrt{n}$.

When an experimenter then reports a result like $\hat{\tau} = 7.82 \pm 0.43$, it is meant that the estimate (e.g. from ML) is 7.82, and if the experiment were repeated many times with the same number of measurements per experiment, one would expect the standard deviation of the distribution of the estimates to be 0.43. This is one possible interpretation of the "statistical error" of a fitted parameter, and is independent of exactly how (according to what p.d.f.) the estimates are distributed. It is not, however, the standard interpretation in those cases where the distribution of estimates from many repeated experiments is not Gaussian. In such cases one usually gives the so-called 68.3% confidence interval, which will be discussed in Chapter 9. This is the same as plus or minus one standard deviation if the p.d.f. for the estimator is Gaussian. It can be shown (see e.g. reference [Ken79], [Fro79]) that in the large sample limit, ML estimates are in fact distributed according to a Gaussian p.d.f., so in this case the two procedures lead to the same result.

6.5 Variance of ML Estimators: Monte Carlo Method

For cases that are too difficult to solve analytically, the distribution of the ML estimates can be investigated with the Monte Carlo method. To do this one must simulate a large number of experiments, compute the ML estimates each time and look at how the resulting values are distributed. For the "true" parameter in the Monte Carlo program the estimated value from the real experiment can be used. As has been seen in the previous section, the quantity s^2 defined by equation (5.7) is an unbiased estimator for the variance of a p.d.f. Thus one can compute s for the ML estimates obtained from the Monte Carlo experiments and give this as the statistical error of the parameter estimated from the real measurement.

As an example of this technique, consider again the case of the mean lifetime measurement with the exponential distribution (Section 6.2). Using a true lifetime of $\tau = 1.0$, a sample of $n = 50$ measurements gave the ML estimate $\hat{\tau} = 1.062$ (see Fig. 6.2). Regarding the first Monte Carlo experiment as the "real" one, the procedure was then repeated 1000 times, with 50 measurements per experiment. The true value was taken to be $\tau = 1.062$, i.e. the ML estimate of the first experiment.

Figure 6.3 shows a histogram of the resulting ML estimates. The sample mean of the estimates is $\overline{\hat{\tau}} = 1.059$, which is close to the input value, as expected since the ML estimator $\hat{\tau}$ is unbiased. The sample standard deviation from the 1000 experiments is $s = 0.151$. This gives essentially the same error value as what one would obtain from equation (6.16), $\hat{\sigma}_{\hat{\tau}} = \hat{\tau}/\sqrt{n} = 1.062/\sqrt{50} = 0.150$. For the real measurement one would then report (for either method to estimate the error) $\hat{\tau} = 1.06 \pm 0.15$.

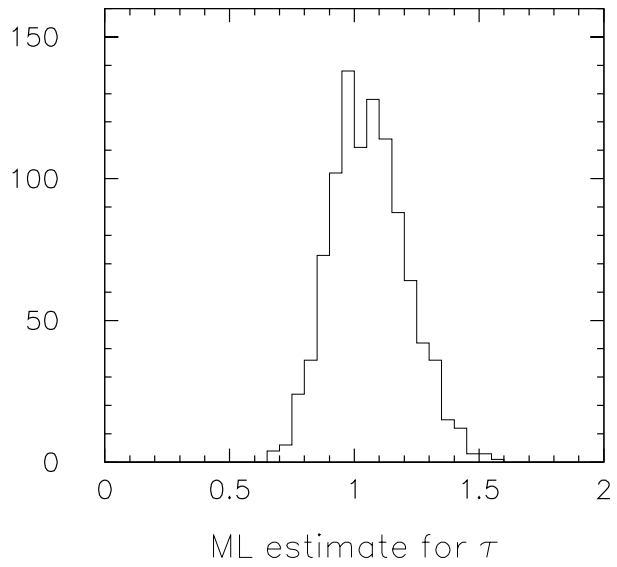

Figure 6.3: A histogram of the ML estimate $\hat{\tau}$ from 1000 Monte Carlo experiments with 50 observations per experiment. For the Monte Carlo "true" parameter τ, the result of Fig. 6.2 was used. The sample standard deviation is $s = 0.151$.

6.6 Variance of ML Estimators: the RCF Bound

It turns out in many applications to be too difficult to compute the variances analytically, and a Monte Carlo study usually involves a significant amount of work. In such cases one typically uses the Rao-Cramér-Frechet (RCF) inequality (see e.g. [Ken79], [Fro79], [Bra92]) which gives a lower bound on an estimator's variance. This inequality applies to any estimator, not just those constructed from the ML principle. For the case of a single parameter θ the limit is given by

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Bigg/ E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right] \ , \tag{6.17}$$

where b is the bias as defined in equation (5.3) and L is the likelihood function. Equation (6.17) is not, in fact, the most general form of the RCF inequality, but the conditions under which the form presented here hold are almost always met in practical situations; see e.g. [Ead71] Section 7.4.5. In the case of equality (i.e. minimum variance) the estimator is said to be *efficient*. It can be shown (see e.g. references [Ken79], [Fro79]) that if efficient estimators exist for a given problem, the maximum likelihood method will find them. Furthermore it can be shown that ML estimators are always efficient in the large sample limit. In practice, one often assumes efficiency and zero bias and hopes that this is a good approximation. In cases of doubt one should check the results with a Monte Carlo study. The general conditions for efficiency are discussed in e.g. [Ead71] Section 7.4.5, [Ken79].

For the example of the exponential distribution with mean τ one has from equation (6.6)

$$\frac{\partial^2 \log L}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2}{\tau}\frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right) \tag{6.18}$$

and $\partial b/\partial \tau = 0$ since $b = 0$ (see equation (6.8)). Thus the RCF bound for the variance (also called the minimum variance bound, or MVB) of $\hat{\tau}$ is

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2}(1 - \frac{2\hat{\tau}}{\tau})\right]} = \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2E[\hat{\tau}]}{\tau}\right)} = \frac{\tau^2}{n} \; , \tag{6.19}$$

where we have used equation (6.8) to obtain $E[\hat{\tau}]$. Since τ^2/n is also the variance obtained from the exact calculation (equation (6.16)) we see that equality holds and $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ is an efficient estimator for the parameter τ.

For the case of more than one parameter, $\theta_1, \ldots, \theta_m$, the corresponding formula for the inverse of the covariance matrix of their estimators $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ is (assuming efficiency and zero bias)

$$(V^{-1})_{ij} = E\left[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right] \; . \tag{6.20}$$

This is then inverted to find V_{ij}, which can then be estimated by evaluating it with the ML estimates $\theta = \hat{\theta}$. Note that equation (6.20) can be written

$$\begin{aligned} (V^{-1})_{ij} &= \int \cdots \int -\frac{\partial^2}{\partial \theta_i \partial \theta_j}\left(\sum_{k=1}^{n} \log f(x_k; \theta_1, \ldots, \theta_m)\right) \prod_{l=1}^{n} f(x_l; \theta_1, \ldots, \theta_m) dx_l \\ &= n \cdot \int -f(x; \theta_1, \ldots, \theta_m) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x; \theta_1, \ldots, \theta_m) dx \; , \end{aligned} \tag{6.21}$$

where $f(x; \theta_1, \ldots, \theta_m)$ is the p.d.f. for the random variable x, for which one has n measurements. That is, the inverse of the RCF bound for the covariance matrix (also called the *Fisher information matrix*, see [Ead71] Section 5.2 and [Bra92]) is proportional to the number of measurements in the sample, n. This expresses the well-known result that statistical errors (at least for efficient estimators) decrease in proportion to $1/\sqrt{n}$.

It turns out to be impractical in many situations to compute the RCF bound analytically, since this requires the expectation value of the second derivative of the log-likelihood function (i.e. an integration over the measured random variable x). In the case of a sufficiently large sample (large number of individual measurements in the experiment) one can estimate V^{-1} by evaluating the second derivative with the measured data and the ML estimates of $\hat{\theta}$:

$$\widehat{(V^{-1})}_{ij} = -\left.\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right|_{\theta = \hat{\theta}} \; . \tag{6.22}$$

For a single parameter θ this reduces to

$$\widehat{\sigma^2}_{\hat{\theta}} = \left(-1 \Big/ \frac{\partial^2 \log L}{\partial \theta^2} \right) \Bigg|_{\theta=\hat{\theta}} . \tag{6.23}$$

This is the usual method for estimating the covariance matrix when the likelihood function is numerically maximized with a computer.[1]

6.7 Variance of ML Estimators: Graphical Method

A simple extension of the previously discussed method using the RCF bound leads to a graphical technique for obtaining the variance of ML estimators. Consider the case of a single parameter θ, and expand the log-likelihood function in a Taylor series about the ML estimate $\hat{\theta}$:

$$\log L(\theta) = \log L(\hat{\theta}) + \left[\frac{\partial \log L}{\partial \theta} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) \; + \; \frac{1}{2!} \left[\frac{\partial^2 \log L}{\partial \theta^2} \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 \; + \ldots \; . \tag{6.24}$$

By definition of $\hat{\theta}$ we know that $\log L(\hat{\theta}) = \log L_{max}$ and that the second term in the expansion is zero. Estimating the variance of $\hat{\theta}$, $\widehat{\sigma^2}_{\hat{\theta}}$ by the RCF bound and ignoring higher order terms gives

$$\log L(\theta) = \log L_{max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}} , \tag{6.25}$$

or

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{max} - \frac{1}{2} . \tag{6.26}$$

That is, a change in the parameter θ of one standard deviation from its ML estimate leads to a decrease in the log-likelihood of 1/2 from its maximum value.

It can be shown that the log-likelihood function is a parabola (i.e. the likelihood function is a Gaussian curve) in the large sample limit [Fro79]. If this is not the case, one can nevertheless adopt equation (6.26) as the definition of the statistical error. The interpretation of such errors is discussed further in Chapter 9.

As an example of the graphical method for determining the variance of an estimator, consider again the examples of Sections 6.2 and 6.5 with the exponential distribution.

[1] For example, the routines MIGRAD and HESSE in the program MINUIT [Jam89] numerically compute the matrix of second derivatives of $\log L$ using finite differences, evaluate it at the ML estimates, and invert to find the covariance matrix.

Figure 6.4 shows the log-likelihood function $\log L(\tau)$ as a function of the parameter τ for a Monte Carlo experiment consisting of 50 measurements. The standard deviation of $\hat{\tau}$ is estimated by changing τ until $\log L(\tau)$ decreases by 1/2, giving $\Delta\hat{\tau}_- = 0.137$, $\Delta\hat{\tau}_+ = 0.165$ (see figure). In this case $\log L(\tau)$ is reasonably close to a parabola and one can approximate $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$. This leads to approximately the same answer as from the exact standard deviation $\tau/\sqrt{n}$ evaluated with $\tau = \hat{\tau}$. In Chapter 9 the interval $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ will be reinterpreted as an approximation for the 68.3% *central confidence interval* (cf. Section 9.6).

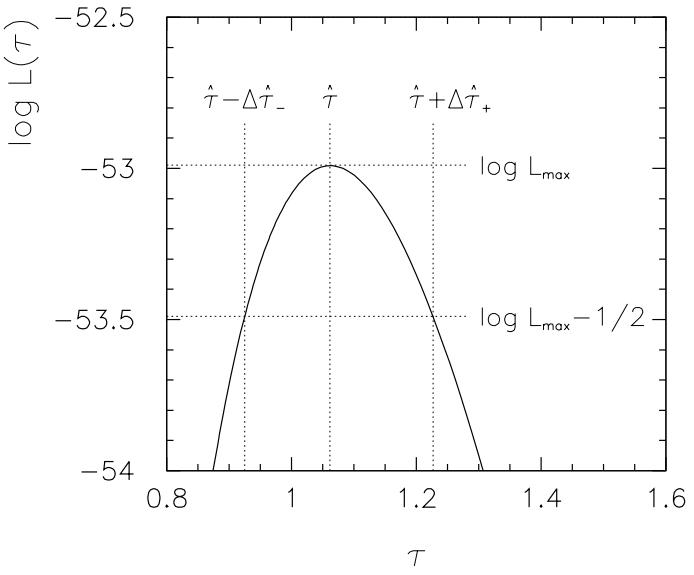

Figure 6.4: Log-likelihood function $\log L(\tau)$ as a function of τ. In the large sample limit the widths of the intervals $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau}]$ and $[\hat{\tau}, \hat{\tau} + \Delta\hat{\tau}_+]$ correspond to one standard deviation $\hat{\sigma}_{\hat{\tau}}$.

6.8 Example of ML with Two Parameters

As an example of the maximum likelihood method with two parameters, consider a particle reaction where each scattering event is characterized by a certain scattering angle θ (or equivalently $x = \cos\theta$). Suppose a given theory predicts the angular distribution

$$f(x;\alpha,\beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3} \,. \tag{6.27}$$

(e.g. $\alpha = 0$ and $\beta = 1$ for $e^+e^- \to \mu^+\mu^-$ in lowest order quantum electrodynamics [Qui83].) Note that the denominator $2 + 2\beta/3$ is necessary for $f(x;\alpha,\beta)$ to be normalized to one for $-1 \le x \le 1$.

To make the problem slightly more complicated (and more realistic) assume that the measurement is only possible in a restricted angular range, say $x_{min} \le x \le x_{max}$. This requires a recalculation of the normalization constant, giving

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{(x_{max}-x_{min})+\frac{\alpha}{2}(x^2_{max}-x^2_{min})+\frac{\beta}{3}(x^3_{max}-x^3_{min})} \; . \tag{6.28}$$

Figure 6.5 shows a histogram of a Monte Carlo experiment where 2000 events were generated using $\alpha = 0.5$, $\beta = 0.5$, $x_{min} = -0.95$ and $x_{max} = 0.95$. By numerically maximizing the log-likelihood function with the program MINUIT one obtains

$$\begin{aligned} \hat{\alpha} &= 0.508 \pm 0.052 \; , \\ \hat{\beta} &= 0.466 \pm 0.108 \; , \end{aligned} \tag{6.29}$$

where the statistical errors correspond to the square roots of the variance. These are estimated by the routine HESSE by numerically computing the matrix of second derivatives of the log-likelihood function with respect to the parameters and then inverting, as described in Section 6.6. In the same manner one obtains the estimate of the covariance $\widehat{\mathrm{cov}}[\hat{\alpha},\hat{\beta}] = 0.00257$ or equivalently the correlation coefficient $r_{\alpha\beta} = 0.458$. One sees that the estimators $\hat{\alpha}$ and $\hat{\beta}$ are positively correlated. Note that the histogram itself is not used in the procedure; each value of x is used to compute the likelihood function.

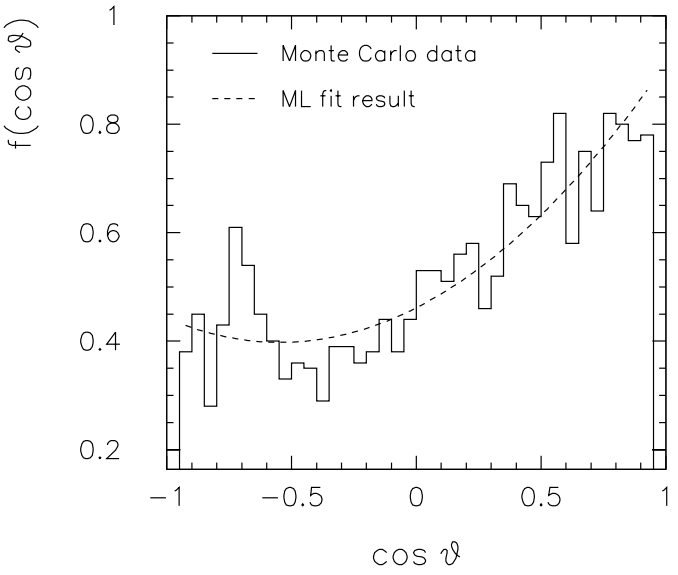

Figure 6.5: Histogram based on 2000 Monte Carlo generated values of $x(=\cos\theta)$ distributed according to equation (6.28) with $\alpha = 0.5$, $\beta = 0.5$. Also shown is the result of the ML fit, which gave $\hat{\alpha} = 0.508 \pm 0.052$ and $\hat{\beta} = 0.466 \pm 0.108$. The errors were computed numerically using equation 6.22.

To understand these results more intuitively it is useful to look at a Monte Carlo study of 500 similar experiments, all with 2000 events with $\alpha = 0.5$ and $\beta = 0.5$. A scatter plot of the ML estimates $\hat{\alpha}$ and $\hat{\beta}$ are shown in Fig. 6.6(a). The density of points corresponds to the joint p.d.f. for $\hat{\alpha}$ and $\hat{\beta}$. Also shown in Fig. 6.6 (b) and (c) are the normalized projected histograms for $\hat{\alpha}$ and $\hat{\beta}$ separately, corresponding to the marginal p.d.f.'s, i.e. the distribution of $\hat{\alpha}$ integrated over all values of $\hat{\beta}$, and vice versa. One sees that the marginal p.d.f.'s for $\hat{\alpha}$ and $\hat{\beta}$ are both approximately Gaussian in shape.

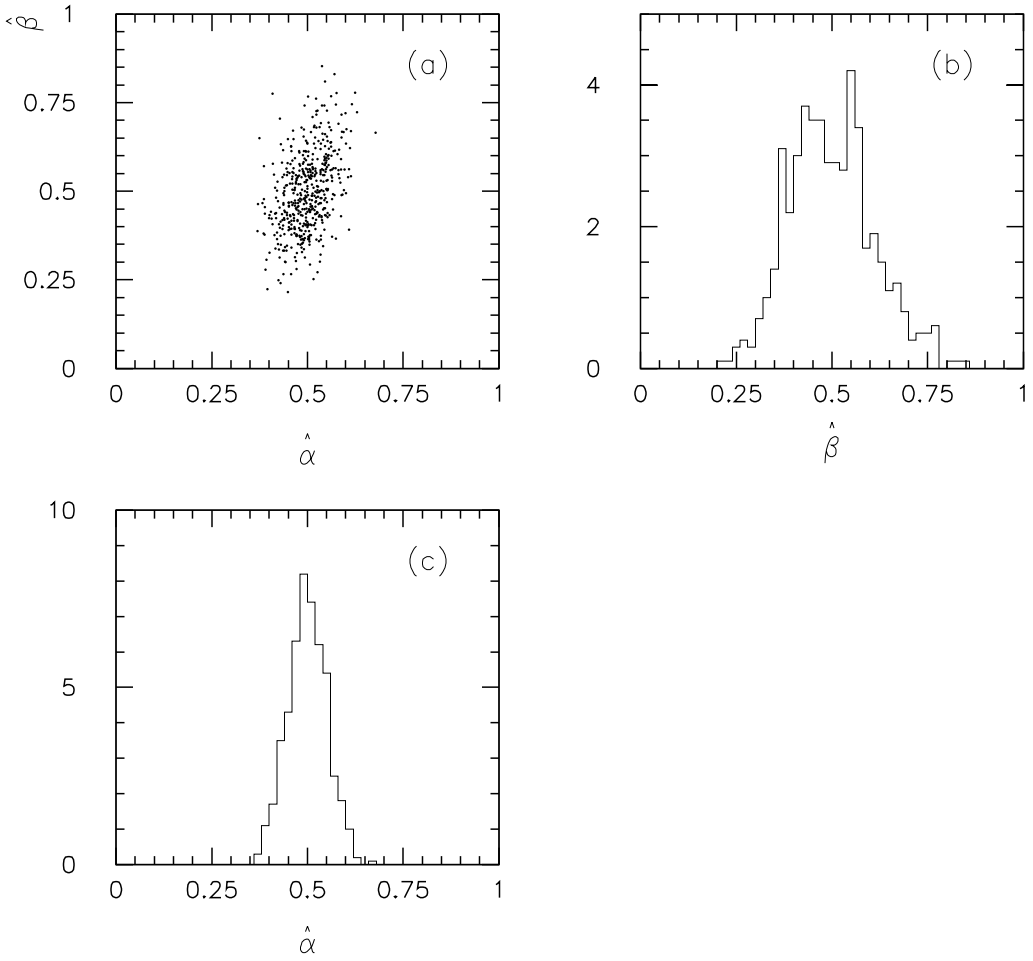

Figure 6.6: Results of ML fits to 500 Monte Carlo generated data sets. (a) The fitted values of $\hat{\alpha}$ and $\hat{\beta}$. (b) The marginal distribution of $\hat{\beta}$. (c) The marginal distribution of $\hat{\alpha}$.

The sample means, standard deviations and covariance (see Section 5.2) from the Monte Carlo experiments are:

$$\begin{array}{ll} \overline{\hat{\alpha}} = 0.499 & \overline{\hat{\beta}} = 0.498 \\ s_{\hat{\alpha}} = 0.051 & s_{\hat{\beta}} = 0.111 \\ \hat{V}_{\alpha\beta} = 0.00235 & r_{\alpha\beta} = 0.418 \end{array} \tag{6.30}$$

Note that $\overline{\hat{\alpha}}$ and $\overline{\hat{\beta}}$ are in good agreement with the "true" values put into the Monte Carlo, ($\alpha = 0.5$ and $\beta = 0.5$) and the sample (co)variances are in good agreement with the values estimated numerically from the RCF bound.

The fact that $\hat{\alpha}$ and $\hat{\beta}$ are correlated is easily seen from the fact that the band of points in the scatter plot is tilted. That is, if one required $\hat{\alpha} > \alpha$, this would lead to an enhanced probability to also find $\hat{\beta} > \beta$. In other words, the conditional p.d.f. for $\hat{\alpha}$ given $\hat{\beta} > \beta$ is centered at a higher mean value and has a smaller variance than the marginal p.d.f. for $\hat{\alpha}$.

Figure 6.7 shows the positions of the ML estimates in the parameter space (i.e. $\log L(\hat{\alpha}, \hat{\beta}) = \log L_{max}$) along with a contour corresponding to $\log L = \log L_{max} - 1/2$.

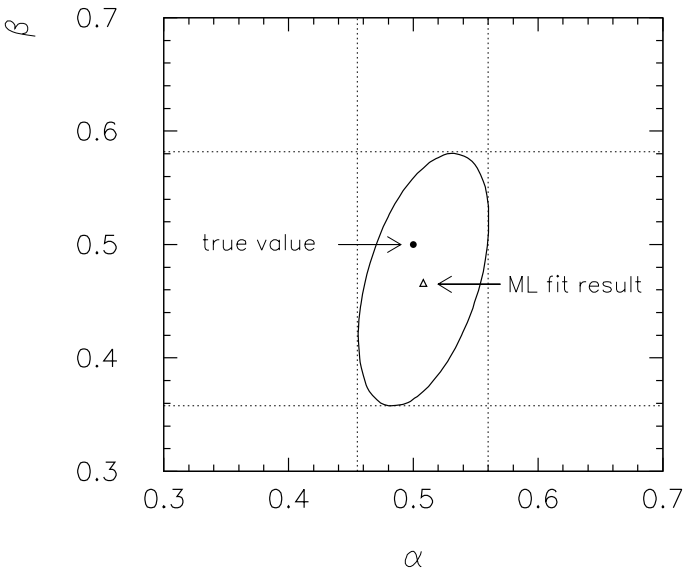

Figure 6.7: The contour $\log L = \log L_{max} - 1/2$ shown with the true values for the parameters (α, β) and the ML estimates $(\hat{\alpha}, \hat{\beta})$. In the large sample limit the tangents to the curve correspond to $\hat{\alpha} \pm \hat{\sigma}_{\hat{\alpha}}$ and $\hat{\beta} \pm \hat{\sigma}_{\hat{\beta}}$.

In the large sample limit the log-likelihood function takes on the form (see reference [Fro79])

$$\log L(\alpha, \beta) = \log L_{max} - \frac{1}{2(1-\rho_{\alpha\beta}^2)} \left[\left(\frac{\alpha - \hat{\alpha}}{\sigma_\alpha} \right)^2 + \left(\frac{\beta - \hat{\beta}}{\sigma_\beta} \right)^2 - 2\rho_{\alpha\beta} \left(\frac{\alpha - \hat{\alpha}}{\sigma_\alpha} \right) \left(\frac{\beta - \hat{\beta}}{\sigma_\beta} \right) \right] \tag{6.31}$$

The contour of $\log L(\alpha,\beta) = \log L_{max} - 1/2$ is thus given by

$$\frac{1}{1-\rho_{\alpha\beta}^2}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_\alpha}\right)^2 + \left(\frac{\beta-\hat{\beta}}{\sigma_\beta}\right)^2 - 2\rho_{\alpha\beta}\left(\frac{\alpha-\hat{\alpha}}{\sigma_\alpha}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_\beta}\right)\right] = 1 \qquad (6.32)$$

which is called the *covariance ellipse*. It is centered at the ML estimates $(\hat{\alpha},\hat{\beta})$ and has an angle ϕ with respect to the α axis given by

$$\tan 2\phi = \frac{2\rho\sigma_\alpha\sigma_\beta}{\sigma_\alpha^2 - \sigma_\beta^2} \; . \qquad (6.33)$$

Note in particular that the tangents to the ellipse are at $\alpha = \hat{\alpha} \pm \sigma_\alpha, \beta = \hat{\beta} \pm \sigma_\beta$ (see Fig. 6.7). If the estimators are correlated, then changing a parameter by one standard deviation corresponds in general to a decrease in the log-likelihood of more than 1/2. If one of the parameters, say β, were known, then the standard deviation of $\hat{\alpha}$ would be somewhat smaller, since this would then be given by a decrease of 1/2 in $\log L(\alpha)$. Similarly, if additional parameters $(\gamma,\delta,\ldots)$ are included in the fit, and if their estimators are correlated with $\hat{\alpha}$, then this will result in an increase in the standard deviation of $\hat{\alpha}$.

6.9 Maximum Likelihood with Binned Data

Consider n observations of a random variable x distributed according to a p.d.f. $f(x;\theta)$ for which one would like to estimate the unknown parameter θ (or parameters $\theta_1,\ldots,\theta_m$). For very large data samples the log-likelihood function becomes difficult to compute since one must sum $\log f(x_i;\theta)$ for each value x_i. In such cases instead of recording the value of each measurement one usually makes a histogram, yielding a certain number of entries $k_1,\ldots,k_N$ in N bins. The expectation value of the number of entries in bin i is given by

$$\lambda_i(\theta) = n\int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx \; , \qquad (6.34)$$

where x_i^{min} and x_i^{max} are the bin limits. One can regard the histogram (i.e. the N values k_i) as a single measurement of an N-dimensional random vector for which the joint p.d.f. is given by a multinomial distribution (equation (2.8))

$$f_{joint}(k_1,\ldots,k_N;\lambda_1,\ldots,\lambda_N) = \frac{n!}{k_1!\cdots k_N!}\left(\frac{\lambda_1}{n}\right)^{k_1}\cdots\left(\frac{\lambda_N}{n}\right)^{k_N} \; . \qquad (6.35)$$

The probability to be in bin i has been expressed as the expectation value λ_i divided by the total number of entries n. Taking the logarithm of the joint p.d.f. gives the log-likelihood function,

$$\log L(\theta) = \sum_{i=1}^{N} k_i \log \lambda_i(\theta) \ . \tag{6.36}$$

where additive terms not depending on the parameters have been dropped. (This is allowed since the estimators depend only on derivatives of $\log L$.) The estimators $\hat{\theta}$ are found by maximizing $\log L$ by whatever means available (e.g. numerically). In the limit that the bin size is very small (i.e. N very large) the likelihood function becomes the same as that of the ML method without binning (equation (6.2)). Thus the binned ML technique does not encounter any difficulties if some of the bins have few or no entries. This is in contrast to an alternative technique using the method of least squares discussed in Section 7.5.

As an example consider again the sample of 50 measured particle decay times that we examined in Section 6.2, for which the maximum likelihood result without binning is shown in Fig. 6.2. Figure 6.8 shows the same sample displayed as a histogram with a bin width of $\Delta t = 0.5$. Also shown is the fit result obtained from maximizing the log-likelihood function based on equation (6.36). The result is $\hat{\tau} = 1.067$, in good agreement with the unbinned result of $\hat{\tau} = 1.062$. Estimating the standard deviation from the curvature of the log-likelihood at its maximum (equation (6.23)) results in $\hat{\sigma}_{\hat{\tau}} = 0.171$, slightly larger than that obtained without binning.

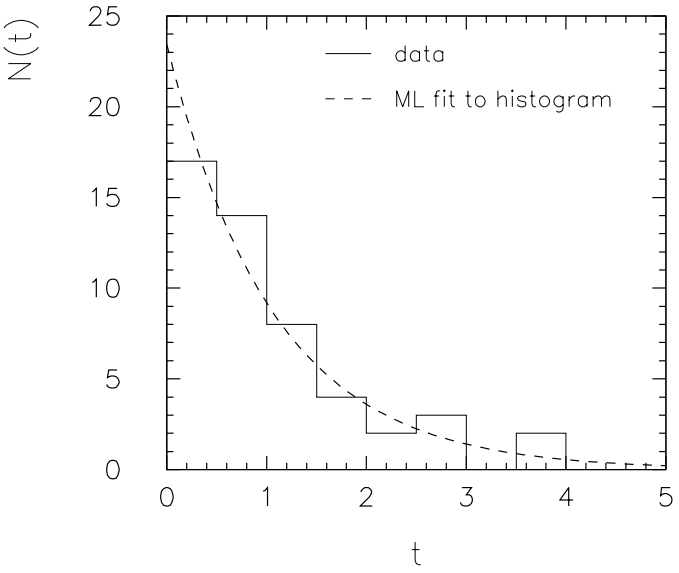

Figure 6.8: Histogram of the data sample of 50 particle decay times from Section 6.2 with the ML fit result.

In many problems one may want to regard the total number of entries n as a random variable from a Poisson distribution with mean ν. The value of ν may itself depend on the other parameters θ, or it may be independent of them.[2] That is, the measurement is

[2]For example, in a particle scattering reaction both the total cross section (i.e. ν) and the angular distribution of the outgoing particles $(\lambda_1, \ldots, \lambda_N)$ depend in general on parameters such as particle masses and coupling constants.

defined to consist of first determining n from a Poisson distribution and then distributing n observations of x in a histogram with N bins, giving $k_1, \ldots, k_N$. The joint p.d.f. for n and $k_1, \ldots, k_N$ is the product of a Poisson distribution and a multinomial distribution,

$$f_{joint}(n, k_1, \ldots, k_N; \nu, \lambda_1, \ldots, \lambda_N) = \frac{\nu^n e^{-\nu}}{n!} \frac{n!}{k_1! \cdots k_N!} \left(\frac{\lambda_1}{\nu}\right)^{k_1} \cdots \left(\frac{\lambda_N}{\nu}\right)^{k_N} , \qquad (6.37)$$

where one has the constraints $\sum_{i=1}^{N} \lambda_i = \nu$ and $\sum_{i=1}^{N} k_i = n$. Using these in equation (6.37) gives

$$f_{joint}(k_1, \ldots, k_N; \lambda_1, \ldots, \lambda_N) = \prod_{i=1}^{N} \frac{\lambda_i^{k_i}}{k_i!} e^{-\lambda_i} , \qquad (6.38)$$

where the mean number of entries in each bin now depends on the parameters θ and ν,

$$\lambda_i(\theta, \nu) = \nu \int_{x_i^{min}}^{x_i^{max}} f(x; \theta) dx . \qquad (6.39)$$

From the joint p.d.f. (6.38) one sees that the problem is equivalent to treating the number of entries in each bin as an *independent* Poisson random variable with mean λ_i. Taking the logarithm of the joint p.d.f. and dropping terms that do not depend on the parameters gives

$$\log L(\theta, \nu) = \sum_{i=1}^{N} k_i \log \lambda_i(\theta, \nu) - \nu . \qquad (6.40)$$

Setting the derivative of $\log L$ with respect to ν equal to zero gives the ML estimator for the total number of entries $\hat{\nu}$,

$$\hat{\nu} = \sum_{i=1}^{N} k_i = n , \qquad (6.41)$$

as one might expect. Since the expectation value of a Poisson variable is equal to its mean, $\hat{\nu} = n$ is an unbiased estimator for ν. The estimators for the parameters θ are clearly the same as those from the case where n was treated as a constant. Any quantity depending on the total number of entries will now have an additional source of fluctuation, however, since $\hat{\nu}$ is a random variable. If ν does not depend on the parameters θ, one has $\partial^2 \log L / \partial\theta\partial\nu = 0$, and thus ν and θ are uncorrelated.

6.10 Testing Goodness-of-Fit with Maximum Likelihood

Although the principle of maximum likelihood defines a technique for estimating the parameters of a hypothesized p.d.f. it does not provide a convenient method of assessing the goodness-of-fit. That is, for a given p.d.f. the ML principle says one should maximize the likelihood function to estimate the parameters. One does not immediately know however, whether L_{max} could have been higher for some other p.d.f., or how high in absolute terms one should expect L_{max} to be if the hypothesis is correct. This is one of the major disadvantages of the maximum likelihood method compared to e.g. the method of least squares discussed in Chapter 7.

One way to investigate the goodness-of-fit is to perform a large number of Monte Carlo experiments with the same number of measurements as the real experiment. For the "true" Monte Carlo parameters, the ML estimates from the real experiment can be used. One then looks at the distribution of the value of the maximized log-likelihood function, $\log L_{max}$. This is shown in Fig. 6.9 for the example of the scattering experiment discussed in Section 6.8. There the data set shown in Fig. 6.5 gave $\hat{\alpha} = 0.508$, $\hat{\beta} = 0.466$ and $\log L_{max} = 2436.4$. Using these parameter values results in the distribution of $\log L_{max}$ values shown in Fig. 6.9. One would have reason to suspect the hypothesis if the real experiment gave a lower $\log L_{max}$ than some large fraction of the Monte Carlo experiments. Given a value of $\log L_{max}$ from a real experiment, one can compute a confidence level (or P-value) as described in Section 4.3, to be used as a measure of the goodness-of-fit.

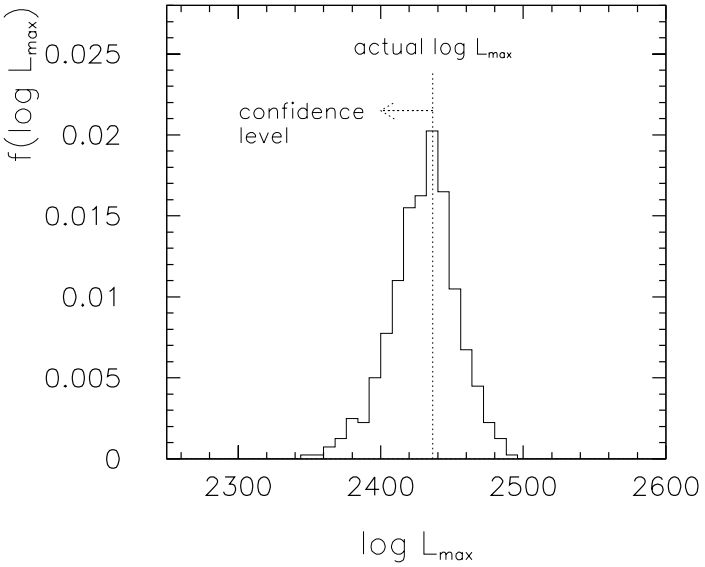

Figure 6.9: Normalized histogram of the values of the maximized log-likelihood function $\log L_{max}$ for 500 Monte Carlo experiments. The vertical line shows the value of $\log L_{max}$ obtained using the data shown in Fig. 6.5. (See text.)

A much simpler qualitative test of the goodness-of-fit is to compare a histogram of the data (normalized to unit area) with the fitted p.d.f. Although the ML fit itself is independent of the binning of the histogram, a visual comparison of the two is a way to quickly check whether the hypothesis is reasonable.

6.11 Combining Measurements with Maximum Likelihood

Consider an experiment in which one has n measured values of a random variable x, for which the p.d.f. $f_x(x;\theta)$ depends on an unknown parameter θ. Suppose in another experiment one has m measured values of a different random variable y, whose p.d.f. $f_y(y;\theta)$ depends on the same parameter θ. For example, x could be the invariant mass of electron-positron pairs produced in proton-antiproton collisions, and y could be the invariant mass of muon pairs. Both distributions have peaks at around the mass M_Z of the Z^0 boson, and so both p.d.f.'s contain M_Z as a parameter. One then wishes to combine the two experiments in order to obtain the best estimate of the parameter.

The two experiments together can be interpreted as a single measurement of a vector containing n values of x and m values of y. The likelihood function is therefore

$$L(\theta) = \prod_{i=1}^{n} f_x(x_i;\theta) \cdot \prod_{j=1}^{m} f_y(y_j;\theta) = L_x(\theta) \cdot L_y(\theta)\,. \tag{6.42}$$

or equivalently its logarithm is given by the sum $\log L(\theta) = \log L_x(\theta) + \log L_y(\theta)$.

Thus as long as the likelihood functions of the experiments are available, the full likelihood function can be constructed and the ML estimator for θ based on both experiments can be determined. This technique includes of course the special case where x and y are the same random variable, and the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ simply represent two different subsamples of the data.

More frequently one does not report the likelihood functions themselves, but rather only estimates of the parameters. Suppose the two experiments based on measurements of x and y give estimators $\hat{\theta}_x$ and $\hat{\theta}_y$ for the parameter θ, which themselves are random variables distributed according to the p.d.f.'s $g_x(\hat{\theta}_x;\theta)$ and $g_y(\hat{\theta}_y;\theta)$. The two estimators can be regarded as the outcome of a single experiment yielding the two-dimensional vector $(\hat{\theta}_x, \hat{\theta}_y)$. As long as $\hat{\theta}_x$ and $\hat{\theta}_y$ are independent, the log-likelihood function is given by the sum

$$\log L(\theta) = \log g_x(\hat{\theta}_x;\theta) + \log g_y(\hat{\theta}_y;\theta)\,. \tag{6.43}$$

For large data samples the p.d.f.'s g_x and g_y can be assumed to be Gaussian, and one reports the estimated standard deviations $\hat{\sigma}_{\hat{\theta}_x}$ and $\hat{\sigma}_{\hat{\theta}_y}$ as the errors on $\hat{\theta}_x$ and $\hat{\theta}_y$. As will be seen in Chapter 7, the problem is then equivalent to the method of least squares, and the combined estimate for θ is given by the weighted average

$$\hat{\theta} = \frac{\hat{\theta}_x/\hat{\sigma}^2_{\hat{\theta}_x} + \hat{\theta}_y/\hat{\sigma}^2_{\hat{\theta}_y}}{1/\hat{\sigma}^2_{\hat{\theta}_x} + 1/\hat{\sigma}^2_{\hat{\theta}_y}}\,, \tag{6.44}$$

with the estimated variance

$$\hat{V}[\hat{\theta}] = \frac{1}{1/\hat{\sigma}^2_{\hat{\theta}_x} + 1/\hat{\sigma}^2_{\hat{\theta}_y}} \,. \tag{6.45}$$

This technique can clearly be generalized to combine any number of measurements.

Chapter 7

The Method of Least Squares

7.1 Connection with Maximum Likelihood

In many situations a measured value y can be regarded as a Gaussian random variable centered about the quantity's true value λ. This follows from the central limit theorem as long as the total error (i.e. deviation from the true value) is the sum of a large number of small contributions. A more detailed discussion of the conditions under which errors can be regarded as Gaussian can be found in references [Bra92], [Ead71], [Fro79].

With this motivation for the importance of Gaussian errors, consider a set of N independent Gaussian random variables y_i, each of which is associated to another variable $x_i, i = 1, \ldots, N$, which is assumed to be known without error. (For example, one has N measurements of a cross section $y_i = \sigma(E_i)$ at different energies $x_i = E_i$.) Assume that each has a different unknown mean, λ_i, and a different but known variance, σ_i^2. The N measurements of y_i can be equivalently regarded as a single measurement of an N-dimensional random vector. The joint p.d.f. for the y_i is the product of N Gaussians,

$$g(y_1, \ldots, y_N; \lambda_1, \ldots, \lambda_N, \sigma_1^2, \ldots, \sigma_N^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-(y_i - y_i^T)^2}{2\sigma_i^2}\right] . \qquad (7.1)$$

Suppose further that we have a hypothesis for the functional dependence of λ on x, $\lambda = f(x; \vec{\theta})$ which depends on unknown parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. The primary aim of the method of least squares is to estimate the parameters $\theta_1, \ldots, \theta_m$. In addition, the method allows for a simple evaluation of the goodness-of-fit of the hypothesized p.d.f.

Taking the logarithm of the joint p.d.f. and dropping additive terms that do not depend on the parameters gives the log-likelihood function,

$$\log L(\vec{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - f(x_i; \vec{\theta}))^2}{\sigma_i^2} . \qquad (7.2)$$

This is maximized by finding the values of the parameters $\vec{\theta}$ that minimize the quantity

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{(y_i - f(x_i;\vec{\theta}))^2}{\sigma_i^2} \ , \tag{7.3}$$

namely, the quadratic sum of the differences between measured and hypothesized values, weighted by the inverse of the variances. This is the basis of the method of least squares (LS), and is used to define the procedure even in cases where the individual measurements y_i are not Gaussian, but as long as they are independent.

If the measurements are not independent but described by an N-dimensional Gaussian p.d.f. with known covariance matrix V but unknown mean values, the corresponding log-likelihood function is obtained from the logarithm of the joint p.d.f. given by equation (2.26)

$$\log L(\vec{\theta}) = -\frac{1}{2} \sum_{i,j=1}^{N} (y_i - f(x_i;\vec{\theta}))(V^{-1})_{ij}(y_j - f(x_j;\vec{\theta})) \ , \tag{7.4}$$

where additive terms not depending on the parameters have been dropped. This is maximized by minimizing the quantity

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^{N} (y_i - f(x_i;\vec{\theta}))(V^{-1})_{ij}(y_j - f(x_j;\vec{\theta})) \ , \tag{7.5}$$

which reduces to equation (7.3) if the covariance matrix (and hence its inverse) are diagonal.

The parameters that minimize the χ^2 are called the LS estimators, $\hat{\theta}_1, \ldots, \hat{\theta}_m$. As will be discussed in Section 7.5, the resulting minimum χ^2 follows under certain circumstances the χ^2 distribution, as defined in Section 2.6. Because of this the quantity defined by equations (7.3) or (7.5) is often called χ^2, even in more general circumstances where its minimum value is not distributed according to the χ^2 p.d.f.

7.2 Linear Least-Squares Fit

Although one can carry out the least-squares procedure for any function $f(x;\vec{\theta})$, the resulting χ^2 value and LS estimators have particularly desirable properties for the case where $f(x;\vec{\theta})$ is a linear function of the parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$:

$$f(x;\theta_1, \ldots, \theta_m) = \sum_{i=1}^{m} \theta_i h_i(x) \ , \tag{7.6}$$

where the $h_i(x)$ are any linearly independent functions of x. (What is required is that f is linear in the parameters θ_i. The $h_i(x)$ are not in general linear in x, they are just linearly independent from each other, i.e. one cannot be expressed as a linear combination

of the others.) For this case, the estimators and their variances can be found analytically, although depending on the tools available one may still prefer to maximize χ^2 numerically with a computer. Furthermore, the estimators have zero bias and minimum variance. This follows from the Gauss-Markov theorem (see [Ken79] Section 19.3) and holds regardless of the number of measurements N, and the p.d.f.'s of the individual measurements.

Using equation (7.6) for the form of $f(x;\vec{\theta})$ in the definition of the χ^2 (equation (7.3)) and differentiating with respect to the parameters θ_i to find the minimum gives

$$\frac{\partial \chi^2}{\partial \theta_k} = \sum_{i=1}^{N} \frac{-2h_k(x_i)}{\sigma_i^2} \left[y_i - \sum_{j=1}^{m} \theta_j h_j(x_i) \right]_{\vec{\theta}=\hat{\vec{\theta}}} = 0, \quad k = 1, \ldots, m \; , \tag{7.7}$$

or m equations which can be solved for the m estimators, $\hat{\theta}_1, \ldots, \hat{\theta}_m$. (These equations and their solutions are often expressed in matrix form; see e.g. references [Fro79], [Bra92].) The estimators can be shown to have zero bias and a covariance matrix whose inverse can be estimated by

$$(V^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k) h_j(x_k)}{\sigma_k^2} \; , \tag{7.8}$$

or equivalently by

$$(V^{-1})_{ij} = \frac{1}{2} \left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\vec{\theta}=\hat{\vec{\theta}}} \; . \tag{7.9}$$

If the variances σ_k^2 are exactly known, then equations (7.8) and (7.9) give the exact inverse covariance matrix $(V^{-1})_{ij}$. Note that equation (7.9) coincides with the RCF bound for the covariance matrix in the situation of Section 7.1, with $\log L = -\chi^2/2$.

For the case of $f(x;\vec{\theta})$ linear in the parameters $\vec{\theta}$ the χ^2 is parabolic in $\vec{\theta}$:

$$\chi^2(\theta_1, \ldots, \theta_m) = \chi^2(\hat{\theta}_1, \ldots, \hat{\theta}_m) + \frac{1}{2} \sum_{i,j=1}^{m} \left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\vec{\theta}=\hat{\vec{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j) \; . \tag{7.10}$$

Combining this with the expression for the variance given by equation (7.9) yields the contours in parameter space whose tangents are at $\hat{\theta}_i \pm \hat{\sigma}_i$, corresponding to a one standard deviation departure from the LS estimates:

$$\chi^2(\theta_1, \ldots, \theta_m) = \chi^2(\hat{\theta}_1, \ldots, \hat{\theta}_m) + 1 = \chi^2_{min} + 1 \; . \tag{7.11}$$

This contour corresponds directly to the covariance ellipse seen in connection with the maximum-likelihood problem of Section 6.10. If the function $f(x;\vec{\theta})$ is not linear in the parameters, then the contour defined by equation (7.11) is not, in general, elliptical, and

one can no longer obtain the standard deviations from the tangent planes. It defines a region in parameter space, however, which can be interpreted as a *confidence region*, the size of which reflects the statistical uncertainty of the fitted parameters. The concept of confidence regions will be defined more precisely in Chapter 9.

7.3 Least-Squares Fit of a Polynomial

As an example of the least-squares method consider the data shown in Fig. 7.1, consisting of five values of a quantity y measured with errors Δy at different values of x. Assume the measured values y_i each come from a Gaussian distribution centered around y_i^T (which is unknown) with a standard deviation $\sigma_i = \Delta y_i$ (assumed known). As a hypothesis for $\lambda = f(x)$ one might try a polynomial of order m (i.e. $m+1$ parameters),

$$f(x;\theta_0,\ldots,\theta_m) = \sum_{i=0}^{m} \theta_i x^i \ . \tag{7.12}$$

This is a special case of the linear least squares fit described in Section 7.2 with the coefficient functions $h_i(x)$ equal to powers of x. Figure 7.1 shows the LS fit result for polynomials of order zero, one and four. The zero-order polynomial is simply the average of the measured values, with each point weighted inversely by the square of its error. This hypothesis gives $\hat{\theta}_0 = 2.665 \pm 0.127$ and $\chi^2 = 45.5$ for four degrees of freedom (five points minus one free parameter). The data are better described by a straight-line fit (first order polynomial) giving $\hat{\theta}_0 = 0.932 \pm 0.297$, $\hat{\theta}_1 = 0.675 \pm 0.105$ and $\chi^2 = 3.99$ for three degrees of freedom. Since there are only five data points, the fourth order polynomial (with five free parameters) goes exactly through every point yielding a χ^2 of zero. The use of the χ^2 value to evaluate the goodness-of-fit will be discussed in Section 7.5.

As in the case of the maximum likelihood method, the statistical errors and covariances of the estimators can be estimated in several ways. All are related to the *change* in the χ^2 as the parameters are moved away from the values for which χ^2 is a minimum. Fig. 7.2(a) shows the χ^2 as a function of θ_0 for the case of the zero-order polynomial. The χ^2 curve is a parabola, since the hypothesized fit function is linear in the parameter θ_0 (see equation (7.10)). The variance of the LS estimator $\hat{\theta}_0$ can be evaluated by any of the methods discussed in Section 7.2: from the change in the parameter necessary to increase the minimum χ^2 by one, from the curvature (second derivative) of the parabola at its minimum, or from the quadratic sum of the inverse errors (equation (7.8)).

Figure 7.2(b) shows a contour of $\chi^2 = \chi^2_{min} + 1$ (the covariance ellipse) for the first-order polynomial (two-parameter) fit. From the inclination of the ellipse one can see that the estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ are negatively correlated. Equation 7.9 gives

$$\begin{aligned} \hat{\sigma}_{\hat{\theta}_0} &= (\widehat{V}[\hat{\theta}_0])^{1/2} = 0.297 \\ \hat{\sigma}_{\hat{\theta}_1} &= (\widehat{V}[\hat{\theta}_1])^{1/2} = 0.105 \end{aligned} \tag{7.13}$$

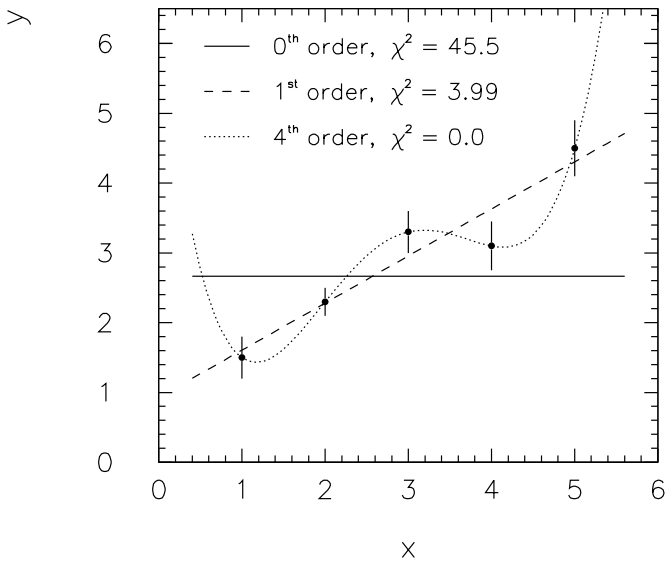

Figure 7.1: Least squares fits of polynomials of order 0, 1 and 4 to five measured values.

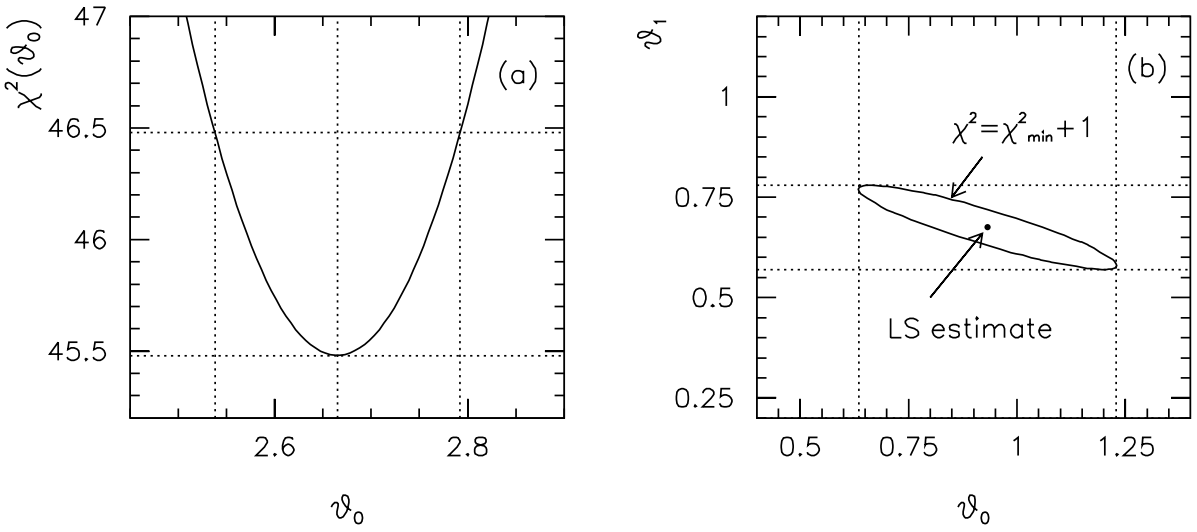

Figure 7.2: (a) The χ^2 as a function of θ_0 for the zero order polynomial fit shown in Fig. 7.1. The horizontal lines indicate χ^2_{min} and $\chi^2_{min}+1$. The corresponding θ_0 values (vertical lines) are the LS estimate $\hat{\theta}_0$ and $\hat{\theta}_0 \pm \hat{\sigma}_{\hat{\theta}_0}$. (b) The LS estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ for the first order polynomial fit in Fig. 7.1. The tangents to the contour $\chi^2(\hat{\theta}_0, \hat{\theta}_1) = \chi^2_{min} + 1$ correspond to $\hat{\theta}_0 \pm \hat{\sigma}_{\hat{\theta}_0}$ and $\hat{\theta}_1 \pm \hat{\sigma}_{\hat{\theta}_1}$.

$$\widehat{V}_{01} \;=\; \widehat{\mathrm{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.0282 \;,$$

corresponding to a correlation coefficient of $r = -0.904$. As in the case of maximum likelihood, the standard deviations correspond to the tangents of the covariance ellipse, and the correlation coefficient to its angle of inclination (see equations (6.32) and (6.33)).

7.4 Least Squares with Binned Data

In the previous examples, the function relating the "true" values λ to the variable x was not necessarily a p.d.f. for x, but an arbitrary function. It *can* be a p.d.f., however, or it can be proportional to one. Suppose, for example, one has n observations of a random variable x from which one makes a histogram with N bins. Let y_i be the number of entries in bin i and $f(x;\theta)$ be a hypothesized p.d.f. for which one would like to estimate the parameter θ (or parameters $\vec{\theta} = \theta_1, \ldots, \theta_m$). The number of entries predicted in bin i, $\lambda_i = E[y_i]$, is then

$$\lambda_i(\theta) = n \int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx = np_i(\theta) \;, \tag{7.14}$$

where x_i^{min} and x_i^{max} are the bin limits and $p_i(\theta)$ is the probability to have an entry in bin i. The parameter θ is found by minimizing the quantity

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{\sigma_i^2} \;, \tag{7.15}$$

where σ_i^2 is the variance of the number of entries in bin i. Note that here the function $f(x;\theta)$ is normalized to one, since it is a p.d.f., and the function that is fitted to the histogram is $\lambda_i(\theta)$.

If the mean number of entries in each bin is small compared to the total number of entries, the contents of each bin is approximately Poisson distributed. The variance is therefore equal to the mean (see equation (2.13)) so that equation (7.15) becomes

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{\lambda_i(\theta)} = \sum_{i=1}^{N} \frac{(y_i - np_i(\theta))^2}{np_i(\theta)} \;. \tag{7.16}$$

An alternative method often used to simplify matters is to approximate the variance of the number of entries in bin i by the number of entries actually observed y_i, rather than by the predicted number $\lambda_i(\theta)$. This is the so-called *modified least-squares method* (MLS) for which one minimizes

$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\theta))^2}{y_i} = \sum_{i=1}^{N} \frac{(y_i - np_i(\theta))^2}{y_i} \; . \tag{7.17}$$

This may be easier to deal with computationally, but has the disadvantage that the errors may be poorly estimated (or χ^2 may even be undefined) if any of the bins contain few or no entries.

When using the LS method for fitting to a histogram one should be aware of the following potential problem. Often instead of using the observed total number of entries n to obtain λ_i from equation (7.14), an additional adjustable parameter ν is introduced as a normalization factor. The predicted number of entries in bin i, $\lambda_i(\theta, \nu) = E[y_i]$, then becomes

$$\lambda_i(\theta, \nu) = \nu \int_{x_i^{min}}^{x_i^{max}} f(x;\theta)dx = \nu p_i(\theta) \; . \tag{7.18}$$

This step would presumably be taken in order to eliminate the need to count the number of entries n. In principle it is simple to determine n but in practice it may require a few extra lines of programming. One can easily show, however, that introducing an adjustable normalization parameter leads to an incorrect estimate of the total number of entries. Consider the LS case where the variances are taken from the predicted number of entries ($\sigma_i^2 = \lambda_i$). Using equation (7.18) for λ_i and differentiating the resulting χ^2 with respect to ν gives the estimator

$$\hat{\nu}_{LS} = n + \frac{\chi^2}{2} \; . \tag{7.19}$$

For the MLS case ($\sigma_i^2 = y_i$) one obtains

$$\hat{\nu}_{MLS} = n - \chi^2 \; . \tag{7.20}$$

Since one expects a contribution to χ^2 on the order of one per bin, the relative error in the number of entries is typically $N/2n$ too high (LS) or N/n too low (MLS). If one takes as a rule of thumb that each bin should have at least five entries one could have an (unnecessary) error in the normalization of $10 - 20\%$.

Although the bias introduced may be smaller than the corresponding statistical error, a result based on the average of such fits could easily be wrong by an amount larger than the statistical error of the average. Therefore, one should determine the normalization directly from the number of entries. If this is not practical (e.g. because of software constraints) one should at least be aware that a potential problem exists, and the bin size should be chosen such that the bias introduced is acceptably small.

The least squares method with binned data can be compared to the maximum likelihood technique of Section 6.9. In that case the joint p.d.f. for the bin contents y_i was taken to be a multinomial distribution, or alternatively each y_i was regarded as

a Poisson random variable. Recall that in the latter case, where the expected total number of entries ν was treated as an adjustable parameter, the correct value $\hat{\nu} = n$ was automatically found (equation (6.41)). Furthermore it has been pointed out in [Ead71] (Section 8.4.5 and references therein) that the variances of ML estimators converge faster to the minimum variance bound than LS or MLS estimators, giving an additional reason to prefer the maximum likelihood method for histogram fitting.

7.5 Testing Goodness-of-Fit with χ^2

If the measured values y_i are Gaussian, the resulting estimators coincide with the ML estimators, as seen in Section 7.1. Furthermore, the χ^2 value can be used as a test of how likely it is that the hypothesis, if true, would yield the observed data.

The quantity $(y_i - f(x_i;\theta))/\sigma_i$ is a measure of the deviation between the ith measurement y_i and the function $f(x_i)$, so χ^2 is a measure of total agreement between observed data and hypothesis. It can be shown (see e.g. [Fro79, Bra92]) that for the case where

(1) the $y_i, i = 1, N$ are independent Gaussian random variables with known variances, σ_i^2 (or are distributed according to an N-dimensional Gaussian with known covariance matrix V);

(2) the hypothesis $f(x;\theta_1,\ldots,\theta_m)$ is linear in the parameters $\theta_i, i = 1, m$, and;

(3) the functional form of the hypothesis is correct,

then the value of χ^2 defined by equation (7.3) (or for correlated y_i by equation (7.5)) is distributed according to the χ^2-distribution with $N - m$ degrees of freedom as defined in Section 2.6, equation (2.29).

As seen in Section 2.6, the expectation value of a random variable z from the χ^2-distribution is equal to the number of degrees of freedom. One often quotes therefore the χ^2 divided by the number of degrees of freedom n_D (the number of data points minus the number of independent parameters) as a measure of goodness-of-fit. If it is near one, then all is as expected. If it is much less than one, then the fit is better than expected given the size of the measurement errors. This is not bad in the sense of providing evidence against the hypothesis, but it is usually grounds to check that the errors σ_i have not been overestimated or are not correlated.

If χ^2/n_D is much larger than one, then there is some reason to doubt the hypothesis. As discussed in Section 4.3, one often quotes a confidence level (CL) for a given χ^2, which is the probability that the hypothesis would lead to a χ^2 value worse (i.e. greater) than the one actually obtained. That is,

$$CL = \int_{\chi^2}^{\infty} f(z;n_D)dz \ , \tag{7.21}$$

where $f(z;n_D)$ is the χ^2-distribution for n_D degrees of freedom. Values can be computed numerically (with e.g. the CERN routine PROB, number G100 [CER96]) or looked up in standard graphs or tables (e.g. references [PDG94, Bra92]). The CL at which one decides to reject a hypothesis is subjective, but note that underestimated errors, σ_i, can cause a correct hypothesis to give a bad χ^2.

For the polynomial fit considered in Section 7.3, one obtained for the straight-line fit $\chi^2 = 3.99$ for three degrees of freedom (five data points minus two free parameters). Computing the confidence level using equation (7.21) gives $CL = 0.263$. That is, if the true function $\lambda = f(x)$ were a straight line and if the experiment were repeated many times, each time yielding values for $\hat{\theta}_0$, $\hat{\theta}_1$ and χ^2, then one would expect the χ^2 values to be worse (i.e. higher) than the one actually obtained ($\chi^2 = 3.99$) in 26.3% of the cases. This can be checked by performing a large number of Monte Carlo experiments where the "true" parameters θ_0 and θ_1 are taken from the results of the real experiment, and a "measured" value for each data point is generated from a Gaussian of width σ given by the corresponding errors. Figure 7.3 shows a normalized histogram of the χ^2 values from 1000 simulated experiments along with the predicted χ^2 distribution for three degrees of freedom.

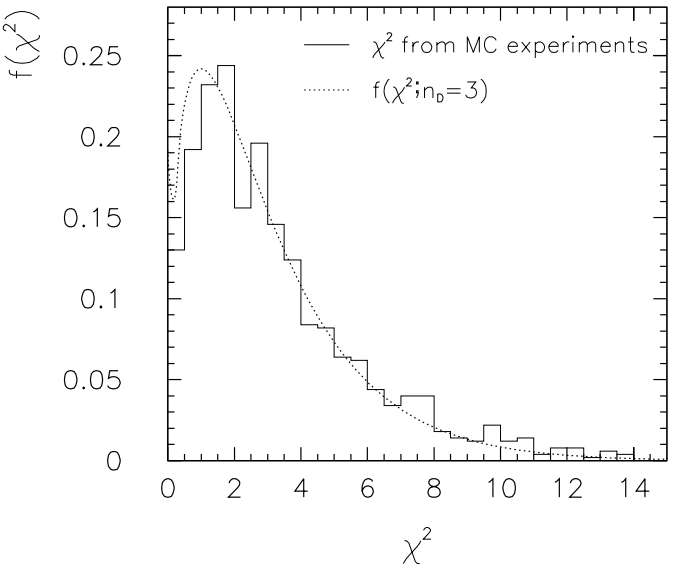

Figure 7.3: Normalized histogram of χ^2 values from 1000 Monte Carlo experiments along with the predicted χ^2-distribution for three degrees of freedom.

For the fit to the horizontal line one had $\chi^2 = 45.5$ for four degrees of freedom. The corresponding confidence level is $CL = 3.1 \cdot 10^{-9}$. If the horizontal-line hypothesis were true, one would expect a χ^2 as high as the one obtained in only three out of a billion experiments, so this hypothesis can safely be ruled out. The advantage of the χ^2 is that it is not necessary to simulate a billion experiments to make a judgement about the goodness-of-fit, since we know that as long as the data points are measurements of Gaussian random variables, the χ^2 value will be distributed according to the χ^2 distribution. This is one of the main advantages of the method of least squares over maximum likelihood, where the value of the maximized likelihood function cannot be interpreted so directly.

One should keep in mind the distinction between having small statistical errors and having a good (i.e. small) χ^2. The statistical errors are related to the *change* in χ^2 when the parameters are varied away from their fitted values, and not to the absolute value of χ^2 itself. From equation (7.8) one can see that the covariance matrix depends only on the coefficient functions $h_i(x)$ (i.e. on the composite hypothesis $f(x;\vec{\theta})$) and on the errors of the individual measurements σ_k, but is independent of the measured values y_k. To demonstrate this point, consider the fit to the horizontal line done in Section 7.3, which yielded the estimate $\hat{\theta}_0 = 2.665 \pm 0.127$ and $\chi^2 = 45.5$ for four degrees of freedom. Figure 7.4 shows a set of five data points with the same x values and the same errors, Δy, but with different y values. A fit to a horizontal line gives $\hat{\theta}_0 = 2.839 \pm 0.127$ and $\chi^2 = 4.48$. The error on $\hat{\theta}_0$ stays the same, but the χ^2 value is now such that the horizontal-line hypothesis provides a good description of the data. The χ^2 vs. θ_0 curves for the two cases have the same curvature, but one is simply shifted vertically with respect to the other by a constant offset.

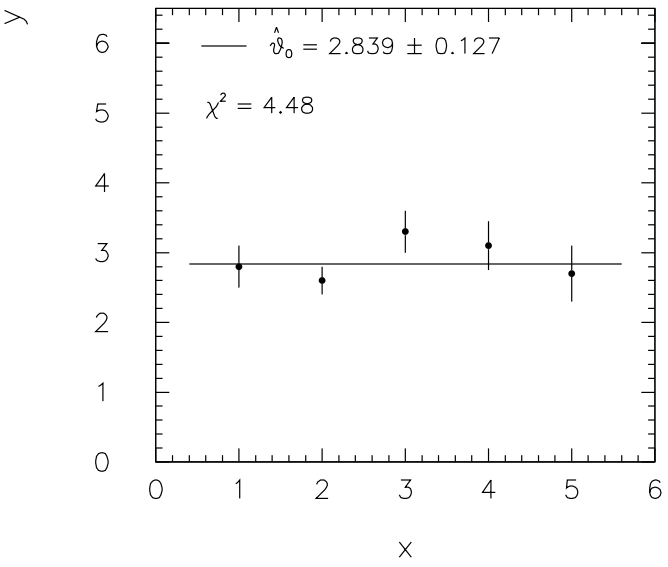

Figure 7.4: Least-squares fit of a zero order polynomial to data with the same x values and errors as shown in Fig. 7.1, but with different y values. Although the χ^2 value is much smaller than in the previous example, the error of $\hat{\theta}_0$ remains the same.

7.6 Combining Measurements with Least Squares

A special case of the method of least squares is often used to combine a number of measurements of the same quantity. Suppose that a quantity y has been measured N times (e.g. by N different experimental groups) yielding independent values y_i and estimated errors (standard deviations) σ_i for $i = 1, \ldots, N$. Since one assumes that the true value is the same for all the measurements, the value λ is a constant (i.e. the function $\lambda = f(x;\theta)$ is a constant, and thus the variable x does not actually appear in the problem). Equation (7.3) becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \lambda)^2}{\sigma_i^2} \ , \tag{7.22}$$

where λ plays the role of the parameter θ. Setting the derivative of χ^2 with respect to λ equal to zero and solving for λ gives the LS estimator $\hat{\lambda}$,

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2} \ , \tag{7.23}$$

which is the well-known formula for a weighted average. From the second derivative of χ^2 one obtains the variance of $\hat{\lambda}$ (see equation (7.8)),

$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2} \ . \tag{7.24}$$

From equation (7.24) one sees that the variance of the weighted average is smaller than any of the variances of the individual measurements. Furthermore, if one of the measured y_i has a much smaller variance than the rest, then this measurement will tend to dominate both in the value and variance of the weighted average.

Chapter 8

The Method of Moments

Although the methods of maximum likelihood and least squares lead to estimators with optimal or nearly optimal properties, they are sometimes difficult to implement. A simpler technique for parameter estimation is the so-called *method of moments* (MM).

Suppose one has a set of n observations of a random variable x, $x_1, \ldots, x_n$, and a hypothesis for the form of the underlying p.d.f. $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ represent m unknown parameters. The idea is to first construct m linearly independent functions $a_i(x), i = 1, \ldots, m$. The $a_i(x)$ are themselves random variables whose expectation values $e_i = E[a_i(x)]$ are functions of the true parameters,

$$E[a_i(x)] = \int a_i(x) f(x; \theta_1, \ldots, \theta_m) dx = e_i(\theta_1, \ldots, \theta_m) \, . \tag{8.1}$$

The functions $a_i(x)$ must be chosen such that the expectation values (8.1) can be computed, so that the functions $e_i(\theta_1, \ldots, \theta_m)$ can be determined.

Since we have seen in Section 5.2 that the sample mean is an unbiased estimator for the expectation value of a random variable, we can estimate the expectation value $e_i = E[a_i(x)]$ by the arithmetic mean of the function $a_i(x)$ evaluated with the observed values of x,

$$\hat{e}_i = \overline{a}_i = \frac{1}{n} \sum_{j=1}^{n} a_i(x_j) \, . \tag{8.2}$$

The MM estimators for the parameters $\theta_1, \ldots, \theta_m$ are defined by setting the expectation values $e_i(\theta_1, \ldots, \theta_m)$ equal to the corresponding estimators $\hat{e}_i$ and solving for the parameters. That is, one solves the following system of m equations for $\hat{\theta}_1, \ldots, \hat{\theta}_m$:

$$\begin{aligned} e_1(\hat{\theta}_1, \ldots, \hat{\theta}_m) &= \frac{1}{n} \sum_{i=1}^{n} a_1(x_i) \\ &\vdots \end{aligned} \tag{8.3}$$

$$e_m(\hat{\theta}_1, \ldots, \hat{\theta}_m) \;=\; \frac{1}{n}\sum_{i=1}^{n} a_m(x_i)\;.$$

Possible choices for the functions $a_i(x)$ are integer powers of the variable x: $x^1, \ldots, x^m$, so that the expectation values $E[a_i(x)] = E[x^i]$ are the ith algebraic moments of x (hence the name "method of moments"). Other sets of m linearly independent functions are possible, however, as long as one can compute their expectation values and obtain m independent functions of the parameters.

We would also like to estimate the covariance matrix for the estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$. In order to obtain this we first estimate the covariance $\mathrm{cov}[a_i(x), a_j(x)]$ using equation (5.9),

$$\mathrm{cov}[a_i(x), a_j(x)] = \frac{1}{n-1}\sum_{k=1}^{n}(a_i(x_k) - \overline{a}_i)(a_j(x_k) - \overline{a}_j)\;. \tag{8.4}$$

From this it follows that the covariance $\mathrm{cov}[\overline{a}_i, \overline{a}_j]$ of the arithmetic means of the functions is

$$\begin{aligned}
\mathrm{cov}[\overline{a}_i, \overline{a}_j] &= \mathrm{cov}\left[\frac{1}{n}\sum_{k=1}^{n} a_i(x_k),\; \frac{1}{n}\sum_{l=1}^{n} a_j(x_l)\right] \\
&= \frac{1}{n^2}\sum_{k,l=1}^{n} \mathrm{cov}[a_i(x_k), a_j(x_l)] \\
&= \frac{1}{n}\,\mathrm{cov}[a_i, a_j]\;. \qquad (8.5)
\end{aligned}$$

The last line follows from the fact there are n terms in the sum over k and l with $k = l$, which each give $\mathrm{cov}[a_i, a_j]$. The other $n^2 - n$ terms have $k \neq l$, for which the covariance $\mathrm{cov}[a_i(x_k), a_j(x_l)]$ vanishes, since the individual x values are independent. The covariance matrix $\mathrm{cov}[\hat{e}_i, \hat{e}_j]$ for the estimators of the expectation values $\hat{e}_i = \overline{a}_i$ can thus be estimated by

$$\mathrm{cov}[\hat{e}_i, \hat{e}_j] = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(a_i(x_k) - \overline{a}_i\right)\left(a_j(x_k) - \overline{a}_j\right)\;. \tag{8.6}$$

In order to obtain the covariance matrix $\mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ for the estimators of the parameters themselves, one can then use equation (8.6) with the error propagation formula (1.52),

$$\mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j] = \sum_{k,l}\frac{\partial\hat{\theta}_i}{\partial\hat{e}_k}\frac{\partial\hat{\theta}_j}{\partial\hat{e}_l}\,\mathrm{cov}[\hat{e}_k, \hat{e}_l] \tag{8.7}$$

Note that even though the value of each measurement x_i is used (i.e. there is no binning of the data) one does not in general exhaust all of the information about the

form of the p.d.f. For example with $a_i(x) = x^i, i = 1, \ldots, m$, only information about the first m moments of x is used, but some of the parameters may be more sensitive to higher moments. For this reason the MM estimators have in general larger variances than those obtained from the principles of maximum likelihood or least squares, discussed in Chapters 6 and 7. (See e.g. [Ead71] Section 8.2.2, [Fro79] Chapters 11 and 12.) Because of its simplicity, however, the method of moments is particularly useful if the estimation procedure must be repeated a large number of times.

As an example consider the p.d.f. for the continuous random variable x given by

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{d_1+\alpha d_2+\beta d_3}\,, \tag{8.8}$$

with $x_{min} \le x \le x_{max}$ and where

$$d_n = \frac{1}{n}\left(x_{max}^n - x_{min}^n\right). \tag{8.9}$$

We have already encountered this p.d.f. in Section 6.8, where the parameters α and β were estimated using the method of maximum likelihood; here for comparison we will use the method of moments. For this we need two linearly independent functions of x, which should be chosen such that their expectation values can easily be computed. A rather obvious choice is

$$\begin{aligned} a_1 &= x \\ a_2 &= x^2\,. \end{aligned} \tag{8.10}$$

The expectation values $e_1 = E[a_1]$ and $e_2 = E[a_2]$ are found to be

$$\begin{aligned} e_1 &= \frac{d_2+\alpha d_3+\beta d_4}{d_1+\alpha d_2+\beta d_3} \\ e_2 &= \frac{d_3+\alpha d_4+\beta d_5}{d_1+\alpha d_2+\beta d_3}\,, \end{aligned} \tag{8.11}$$

with d_n again given by equation (8.9). Solving these two equations for α and β and replacing e_1 and e_2 by $\hat{e}_1$ and $\hat{e}_2$ gives the MM estimators,

$$\begin{aligned} \hat{\alpha} &= \frac{(\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_1 - d_3) - (\hat{e}_1 d_1 - d_2)(\hat{e}_2 d_3 - d_5)}{(\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_3 - d_5) - (\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_2 - d_4)} \\ \hat{\beta} &= \frac{(\hat{e}_1 d_1 - d_2)(\hat{e}_2 d_2 - d_4) - (\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_1 - d_3)}{(\hat{e}_1 d_2 - d_3)(\hat{e}_2 d_3 - d_5) - (\hat{e}_1 d_3 - d_4)(\hat{e}_2 d_2 - d_4)}\,. \end{aligned} \tag{8.12}$$

From the example of Section 6.8 we had a data sample of 2000 x values generated with $\alpha = 0.5$, $\beta = 0.5$, $x_{min} = -0.95$, $x_{max} = 0.95$. Using the same data here gives

$$\begin{aligned} \hat{\alpha} &= 0.493 \pm 0.051 \\ \hat{\beta} &= 0.410 \pm 0.106 \, . \end{aligned}$$

The statistical errors are obtained by means of error propagation from the covariance matrix for $\hat{e}_1$ and $\hat{e}_2$, which is estimated using equation (8.6). Similarly one obtains the correlation coefficient $r = 0.417$.

These results are similar to those obtained using maximum likelihood, and the error estimates are actually slightly smaller. The latter fact is the result, however, of a statistical fluctuation in estimating the variances. In fact the variances of MM estimators are in general greater than or equal to those of the ML estimators; a Monte Carlo calculation gives for the MM case here $\hat{\sigma}_{\hat{\alpha}} = 0.053$, $\hat{\sigma}_{\hat{\beta}} = 0.111$. This is to be compared with $\hat{\sigma}_{\hat{\alpha}} = 0.051$, $\hat{\sigma}_{\hat{\beta}} = 0.112$ as obtained in Section 6.8 using maximum likelihood. Thus for this particular example the statistical errors are almost the same using either method. The method of moments has the advantage, however, that the estimates can be obtained without having to maximize the likelihood function, which in this example (and most others) would require a more complicated numerical calculation.

Chapter 9

Statistical Errors, Confidence Intervals and Limits

9.1 The Standard Deviation as Statistical Error

In Chapters 5 – 8 several methods for estimating properties of p.d.f.'s (e.g. moments, parameters) have been discussed along with techniques for obtaining the variance of the estimators. The variance (or equivalently its square root, the standard deviation) of an estimator is a measure of how widely its value would be distributed if the experiment were to be repeated many times with the same number of observations per experiment. As such, the standard deviation σ is often reported as the statistical uncertainty of a measurement, and is referred to as the *standard error*.

For example, suppose one has n observations of a random variable x and a hypothesis for the p.d.f. $f(x;\theta)$ which contains an unknown parameter θ. From the sample $x_1, \ldots, x_n$ a function $\hat{\theta}(x_1, \ldots, x_n)$ is constructed (using e.g. maximum likelihood) as an estimator for θ. Using one of the techniques discussed in Chapters 5 - 8 (e.g. analytic method, RCF bound, Monte Carlo, graphical) the standard deviation of $\hat{\theta}$ can be estimated. Let $\hat{\theta}_{exp}$ be the value of the estimator actually obtained, and $\hat{\sigma}_{\hat{\theta}}$ the estimate of its standard deviation. In reporting the measurement of θ as $\hat{\theta}_{exp} \pm \hat{\sigma}_{\hat{\theta}}$ one means that repeated estimates all based on n observations of x would be distributed according to a p.d.f. $g(\hat{\theta})$ centered around some true value θ and true standard deviation $\sigma_{\hat{\theta}}$, which are estimated to be $\hat{\theta}_{exp}$ and $\hat{\sigma}_{\hat{\theta}}$.

Although this definition of statistical error bars could in principle be used regardless of the form of the estimator's p.d.f. $g(\hat{\theta})$, it is not, in fact, the conventional definition if $g(\hat{\theta})$ is not Gaussian. In such cases one uses confidence intervals as described in the next section, which can in general lead to asymmetric error bars. In Section 9.3 it is shown that if $g(\hat{\theta})$ is Gaussian, then the so-called 68.3% confidence interval is the same as the interval covered by $\hat{\theta}_{exp} \pm \hat{\sigma}_{\hat{\theta}}$.

9.2 Classical Confidence Intervals (Exact Method)

An alternative (and often equivalent) method of reporting the statistical error of a measurement is with a so-called confidence interval. Suppose as above that one has n observations of a random variable x which can be used to evaluate an estimator $\hat{\theta}(x_1, \ldots, x_n)$ for a parameter θ, and that the value obtained is $\hat{\theta}_{exp}$. Furthermore, suppose that based on e.g. an analytical calculation or a Monte Carlo study, one knows the p.d.f of $\hat{\theta}$, $g(\hat{\theta};\theta)$, which contains the true value θ as a parameter. That is, the real value of θ is not known, but for a given θ, one knows what the p.d.f. of $\hat{\theta}$ would be.

Figure 9.1 shows a probability density for an estimator $\hat{\theta}$ for a particular value of the true parameter θ. From $g(\hat{\theta};\theta)$ one can determine the value u_α such that there is a fixed probability α to observe $\hat{\theta} \geq u_\alpha$, and similarly the value v_β such that there is a probability β to observe $\hat{\theta} \leq v_\beta$. The values u_α and v_β depend on the true value of θ, and are thus defined by requiring

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta)\,. \tag{9.1}$$

and

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta} = G(v_\beta(\theta);\theta)\,, \tag{9.2}$$

where G is the cumulative distribution corresponding to the p.d.f. $g(\hat{\theta};\theta)$.

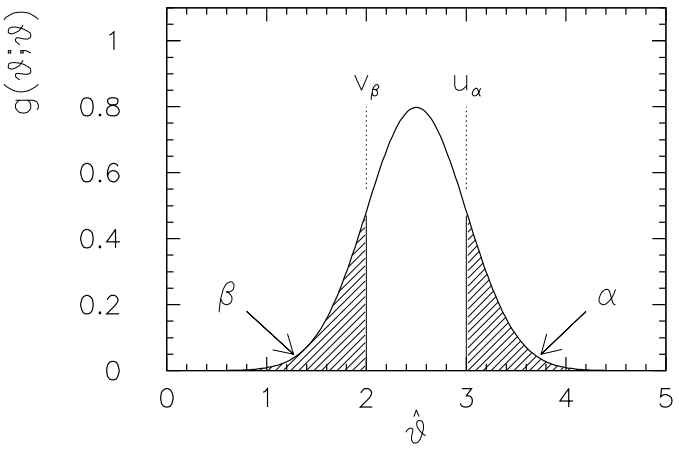

Figure 9.1: A p.d.f. $g(\hat{\theta};\theta)$ for an estimator $\hat{\theta}$ for a given value of the true parameter θ. The two shaded regions show the values of $\hat{\theta} \leq v_\beta$, which has a probability β, and $\hat{\theta} \geq u_\alpha$, which has a probability α.

Figure 9.2 shows an example of how the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ might appear as a function of the true value of θ. The region between the two curves is called the *confidence belt*. The probability for the estimator to be inside the belt, regardless of the value of θ, is given by

$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta\,. \tag{9.3}$$

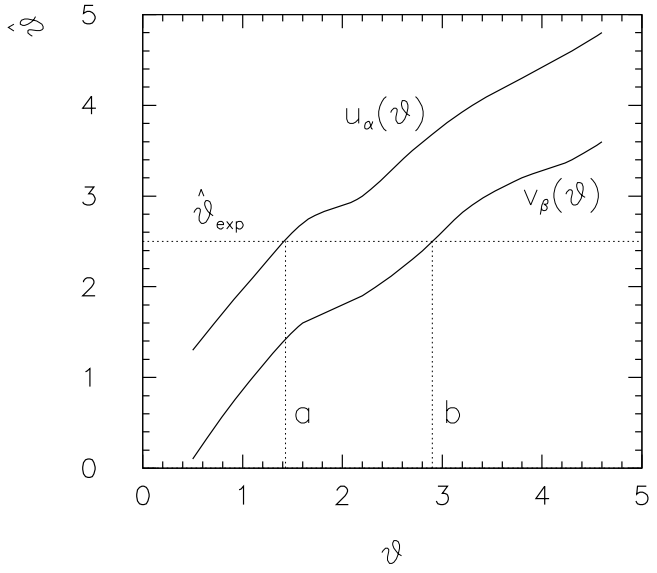

Figure 9.2: Construction of the confidence interval $[a, b]$ given an observed value $\hat{\theta}_{exp}$ of the estimator $\hat{\theta}$ for the parameter θ. (See text.)

As long as $u_\alpha(\theta)$ and $v_\beta(\theta)$ are monotonically increasing functions of θ, which in general should be the case if $\hat{\theta}$ is to be a good estimator for θ, one can determine the inverse functions

$$\begin{aligned} a(\hat{\theta}) &\equiv u_\alpha^{-1}(\hat{\theta}) \,, \\ b(\hat{\theta}) &\equiv v_\beta^{-1}(\hat{\theta}) \,. \end{aligned} \tag{9.4}$$

The inequalities

$$\begin{aligned} \hat{\theta} &\geq u_\alpha(\theta) \,, \\ \hat{\theta} &\leq v_\beta(\theta) \,, \end{aligned} \tag{9.5}$$

then imply respectively

$$\begin{aligned} a(\hat{\theta}) &\geq \theta \,, \\ b(\hat{\theta}) &\leq \theta \,. \end{aligned} \tag{9.6}$$

Equations (9.1) and (9.2) thus become

$$\begin{aligned} P(a(\hat{\theta}) \geq \theta) &= \alpha \,, \\ P(b(\hat{\theta}) \leq \theta) &= \beta \,, \end{aligned} \tag{9.7}$$

or taken together,

$$P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta \,. \tag{9.8}$$

If the functions $a(\hat{\theta})$ and $b(\hat{\theta})$ are evaluated with the value of the estimator actually obtained in the experiment, $\hat{\theta}_{exp}$, then this determines two values, a and b, as illustrated in Fig. 9.2. The interval $[a, b]$ is called a *confidence interval* at a confidence level[1] of $1 - \alpha - \beta$. The idea behind its construction is that equations (9.7), and hence also (9.8), hold regardless of the true value of θ, which, of course, is unknown. It should be emphasized that a and b are random values, since they depend on the estimator $\hat{\theta}$, which is itself a function of the data. If the experiment were repeated many times, the interval $[a, b]$ would include the true value of the parameter θ in a fraction $1 - \alpha - \beta$ of the experiments.

In some situations one may only be interested in a *one-sided confidence interval* or *limit*. That is, the value a represents a lower limit on the parameter θ such that $a \leq \theta$ with the probability $1-\alpha$. Similarly, b represents an upper limit on θ such that $P(\theta \leq b) = 1-\beta$.

Two-sided intervals (i.e. both a and b specified), are not uniquely determined by the confidence level $1 - \alpha - \beta$. One often chooses, for example, $\alpha = \beta = \gamma/2$ giving a so-called *central* confidence interval with probability $1 - \gamma$. Note that a central confidence interval does not necessarily mean that a and b are equidistant from the estimated value $\hat{\theta}$, but only that the probabilities α and β are equal.

By construction the value a gives the (hypothetical) value of the true parameter θ for which a fraction α of repeated estimates $\hat{\theta}$ would be higher than the one actually obtained, $\hat{\theta}_{exp}$, as is illustrated in Fig. 9.3. Similarly, b is the value of θ for which a fraction β of the estimates would be lower than $\hat{\theta}_{exp}$. That is, taking $\hat{\theta}_{exp} = u_\alpha(a) = v_\beta(b)$, equations (9.1) and (9.2) become

$$\begin{aligned} \alpha &= \int_{\hat{\theta}_{exp}}^{\infty} g(\hat{\theta}; a)\, d\hat{\theta} = 1 - G(\hat{\theta}_{exp}; a)\,, \\ \beta &= \int_{-\infty}^{\hat{\theta}_{exp}} g(\hat{\theta}; b)\, d\hat{\theta} = G(\hat{\theta}_{exp}; b)\,. \end{aligned} \tag{9.9}$$

The previously described procedure to determine the confidence interval is thus equivalent to solving (9.9) for a and b, e.g. numerically.

The confidence interval $[a, b]$ is often expressed by reporting the result of a measurement as $\hat{\theta}^{+d}_{-c}$, where $\hat{\theta}$ is the estimated value, and $c = \hat{\theta} - a$ and $d = b - \hat{\theta}$ are usually displayed as *error bars*. In many cases the p.d.f. $g(\hat{\theta}; \theta)$ is approximately Gaussian, so that an interval of plus or minus one standard deviation around the measured value corresponds to a central confidence interval with $1 - \gamma = 0.683$ (see Section 9.3). The 68.3% central confidence interval is usually adopted as the conventional definition for error bars even when the p.d.f. of the estimator is not Gaussian.

If, for example, the result of an experiment is reported as $\hat{\theta}^{+d}_{-c} = 5.79^{+0.32}_{-0.25}$, it is meant that if one were to construct the interval $[\hat{\theta}-c, \hat{\theta}+d]$ according to the prescription described

[1] This should not be confused with the confidence level of a goodness-of-fit test (see Section 4.3).

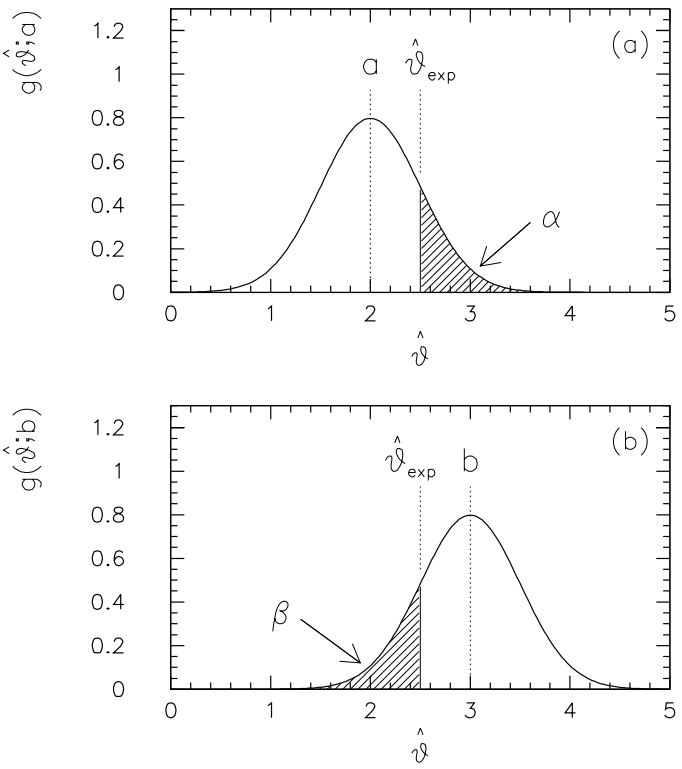

Figure 9.3: (a) The p.d.f. $g(\hat{\theta};a)$, where a is the lower limit of the confidence interval. If the true parameter θ were equal to a, the estimates $\hat{\theta}$ would be greater than the one actually observed $\hat{\theta}_{exp}$ with a probability α. (b) The p.d.f. $g(\hat{\theta};b)$, where b is the upper limit of the confidence interval. If θ were equal to b, $\hat{\theta}$ would be observed less than $\hat{\theta}_{exp}$ with probability β.

above in a large number of similar experiments with the same number of measurements per experiment, then the interval would include the true value θ in $1-\alpha-\beta$ of the cases. It does not mean that the probability (defined in the sense of limiting relative frequency) that the true value of θ is in the fixed interval $[5.54, 6.11]$ is $1-\alpha-\beta$. In the limiting frequency interpretation, the true parameter θ is not a random variable and is assumed to not fluctuate from experiment to experiment. In this sense the probability that θ is in $[5.54, 6.11]$ is either 0 or 1, but we do not know which. The interval itself, however, is subject to fluctuations since it is constructed from the data.

A difficulty in constructing confidence intervals is that the p.d.f. of the estimator $g(\hat{\theta};\theta)$, or equivalently the cumulative distribution $G(\hat{\theta};\theta)$, must be known. An example is given in Section 10.2, where the p.d.f. for the estimator of the mean ξ of an exponential distribution is derived, and from this a confidence interval for ξ is determined. In many practical applications, estimators are Gaussian distributed (at least approximately). In this case the confidence interval can be determined easily; this is treated in detail in the next section. Even in the case of a non-Gaussian estimator, however, a simple approximate technique can be applied using the likelihood function; this is described in Section 9.6.

9.3 Confidence Interval for Gaussian Distributed Estimator

A simple and very important application of a confidence interval is when the distribution of $\hat{\theta}$ is Gaussian with mean θ and standard deviation $\sigma_{\hat{\theta}}$. That is, the cumulative distribution of $\hat{\theta}$ is

$$G(\hat{\theta};\theta,\sigma_{\hat{\theta}}) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta}'-\theta)^2}{2\sigma_{\hat{\theta}}^2}\right) d\hat{\theta}' \,. \tag{9.10}$$

This is a commonly occurring situation since, according to the Central Limit theorem, any estimator that is a linear function of a sum of random variables becomes Gaussian in the large sample limit. We will see that for this case, the somewhat complicated procedure explained in the previous section results in a particularly simple prescription for determining the confidence interval.

Suppose that the standard deviation $\sigma_{\hat{\theta}}$ is known, and that the experiment has resulted in an estimate $\hat{\theta}_{exp}$ for θ. According to equations (9.9), the confidence interval $[a,b]$ is determined by solving the equations

$$\begin{aligned} \alpha &= 1 - G(\hat{\theta}_{exp};a,\sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{exp}-a}{\sigma_{\hat{\theta}}}\right) \,, \\ \beta &= G(\hat{\theta}_{exp};b,\sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{exp}-b}{\sigma_{\hat{\theta}}}\right) \,, \end{aligned} \tag{9.11}$$

for a and b, where G has been expressed using the cumulative distribution of the standard Gaussian Φ (2.24) (see also (2.25)). This gives

$$\begin{aligned} a &= \hat{\theta}_{exp} - \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\alpha) \,, \\ b &= \hat{\theta}_{exp} + \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\beta) \,. \end{aligned} \tag{9.12}$$

Here Φ^{-1} is the inverse function of Φ, i.e. the quantile of the standard Gaussian, and in order to make the two equations symmetric we have used $\Phi^{-1}(\beta) = -\Phi^{-1}(1-\beta)$.

The quantiles $\Phi^{-1}(1-\alpha)$ and $\Phi^{-1}(1-\beta)$ represent how far away the interval limits a and b are located with respect to the estimate $\hat{\theta}_{exp}$ in units of the standard deviation $\sigma_{\hat{\theta}}$. The relationship between the quantiles of the standard Gaussian distribution and the confidence level is illustrated in Fig. 9.4(a) for central and Fig. 9.4(b) for one-sided confidence intervals.

Consider a central confidence interval with $\alpha = \beta = \gamma/2$. The confidence level $1-\gamma$ is often chosen such that the quantile is a small integer, e.g. $\Phi^{-1}(1-\gamma/2) = 1, 2, 3, \ldots$.

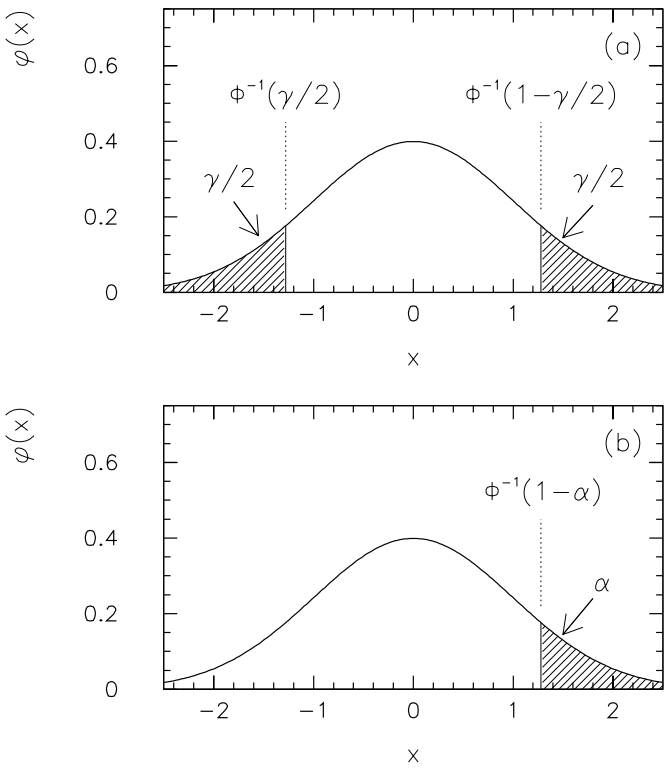

Figure 9.4: The standard Gaussian p.d.f. $\varphi(x)$ showing the relationship between the quantiles Φ^{-1} and the confidence level for (a) a central confidence interval and (b) a one-sided confidence interval.

Similarly, for one-sided intervals (i.e. limits) one often chooses a small integer for $\Phi^{-1}(1-\alpha)$. Commonly used values for both central and one-sided intervals are shown in Table 9.1. Alternatively one can choose a round number for the confidence level instead of for the quantile. Commonly used values are shown in Table 9.2. Other possible values can be obtained from [Bra92, Fro79, Dud88] or from computer routines (e.g. [CER96], routine G105).

Quantile of standard Gaussian $\Phi^{-1}(1-\gamma/2)$	Confidence level for central interval $1-\gamma$	Quantile of standard Gaussian $\Phi^{-1}(1-\alpha)$	Confidence level for one-sided interval $1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9974	3	0.9987

Table 9.1: The values of the confidence level for different values of the quantile Φ^{-1} for central and one-sided confidence intervals. The relationship between the quantile and confidence level is illustrated in Fig. 9.4.

For the 68.3% central confidence interval one has $\alpha = \beta = \gamma/2$, with $\Phi^{-1}(1-\gamma/2) = 1$, i.e. a "1 σ error bar". This results in the simple prescription,

$$[a, b] = [\hat{\theta}_{exp} - \sigma_{\hat{\theta}}, \hat{\theta}_{exp} + \sigma_{\hat{\theta}}] \, . \tag{9.13}$$

Confidence level for for central interval $1-\gamma$	Quantile of standard Gaussian $\Phi^{-1}(1-\gamma/2)$	Confidence level for one-sided interval $1-\alpha$	Quantile of standard Gaussian $\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

Table 9.2: The values of the quantile Φ^{-1} for different values of the confidence level for central and one-sided confidence intervals. The relationship between the quantile and confidence level is illustrated in Fig. 9.4.

Thus for the case of a Gaussian distributed estimator, the 68.3% central confidence interval is given by the estimated value plus or minus one standard deviation. The final result of the measurement of θ is then simply reported as $\hat{\theta}_{exp} \pm \sigma_{\hat{\theta}}$.

If the standard deviation $\sigma_{\hat{\theta}}$ is not known *a priori* but rather is estimated from the data, then the situation is in principle somewhat more complicated. If, for example, the estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$ had been used instead of $\sigma_{\hat{\theta}}$, then it would not have been so simple to relate the cumulative distribution $G(\hat{\theta};\theta,\hat{\sigma}_{\hat{\theta}})$ to Φ, the cumulative distribution of the standard Gaussian, since $\hat{\sigma}_{\hat{\theta}}$ depends in general on $\hat{\theta}$. In practice, however, the recipe given above can still be applied using the estimate $\hat{\sigma}_{\hat{\theta}}$ instead of $\sigma_{\hat{\theta}}$, as long as $\hat{\sigma}_{\hat{\theta}}$ is a sufficiently good approximation of the true standard deviation, e.g. in the large sample limit.[2]

Exact determination of confidence intervals becomes more difficult if the p.d.f. of the estimator $g(\hat{\theta};\theta)$ is not Gaussian, or worse, if it is not known analytically. For a non-Gaussian p.d.f. it is sometimes possible to transform the parameter $\theta \to \eta(\theta)$ such that p.d.f. for the estimator $\hat{\eta}$ is approximately Gaussian. The confidence interval for the transformed parameter η can then be converted back into an interval for θ. An example of this technique is given in Section 9.5.

9.4 Confidence Interval for the Mean of the Poisson Distribution

Along with the Gaussian distributed estimator, another commonly occurring case is where the outcome of a measurement is a Poisson variable k, with $k = 0, 1, 2, \ldots$. Recall from (2.11) that the probability to observe k is

$$f(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \,, \tag{9.14}$$

[2]For the small sample case where $\hat{\theta}$ represents the mean of n Gaussian random variables of unknown standard deviation, the confidence interval can be determined by relating the cumulative distribution $G(\hat{\theta};\theta,\hat{\sigma}_{\hat{\theta}})$ to Student's t-distribution; see e.g. [Fro79], [Dud88] Section 10.2.

and that the parameter λ is equal to the expectation value $E[k]$. The maximum-likelihood estimator for λ can easily be found to be $\hat{\lambda} = k$. Suppose a single measurement has resulted in the value $\hat{\lambda}_{exp} = k_{exp}$, and based on this one would like to construct a confidence interval for the mean λ.

For the case of a discrete variable, the procedure for determining the confidence interval described in Section 9.2 cannot be directly applied. This is because the quantities that determine the confidence belt, $u_\alpha(\theta)$ and $v_\beta(\theta)$, do not exist for all values of the parameter θ. For the Poisson case, for example, we would need to find $u_\alpha(\lambda)$ and $v_\beta(\lambda)$ such that $P(\hat{\lambda} \geq u_\alpha(\lambda)) = \alpha$ and $P(\hat{\lambda} \leq v_\beta(\lambda)) = \beta$ for arbitrary α and β and for all values of the parameter λ. But if α and β are fixed, then because $\hat{\lambda}$ only takes on discrete values, these equations hold in general only for particular values of λ.

A confidence interval $[a, b]$ can still be determined, however, by using equations (9.9). For the case of a discrete random variable and a parameter λ these become

$$\begin{aligned} \alpha &= P(\hat{\lambda} \geq \hat{\lambda}_{exp}; a)\,, \\ \beta &= P(\hat{\lambda} \leq \hat{\lambda}_{exp}; b)\,, \end{aligned} \tag{9.15}$$

and in particular for a Poisson variable one has

$$\begin{aligned} \alpha &= \sum_{k=k_{exp}}^{\infty} f(k;a) = 1 - \sum_{k=0}^{k_{exp}-1} f(k;a) = 1 - e^{-a} \sum_{k=0}^{k_{exp}-1} \frac{a^k}{k!}\,, \\ \beta &= \sum_{k=0}^{k_{exp}} f(k;b) = e^{-b} \sum_{k=0}^{k_{exp}} \frac{b^k}{k!}\,. \end{aligned} \tag{9.16}$$

For any estimate $\hat{\lambda} = k_{exp}$ and given probabilities α and β these equations can be solved numerically for a and b. Note that the lower limit a cannot be determined if $k_{exp} = 0$. Equations (9.15) say that if $\lambda = a$ ($\lambda = b$), then the probability is α (β) to observe a value greater (less) than *or equal to* the one actually observed. The fact that k_{exp} is included in the inequalities leads to a conservatively large confidence interval, i.e.

$$\begin{aligned} P(\lambda \geq a) &\geq 1 - \alpha \\ P(\lambda \leq b) &\geq 1 - \beta \\ P(a \leq \lambda \leq b) &\geq 1 - \alpha - \beta\,. \end{aligned} \tag{9.17}$$

An important special case is when the observed number k_{exp} is zero, and one is interested in establishing an upper limit b. Equation (9.15) becomes

$$\beta = \sum_{k=0}^{0} \frac{b^k \, e^{-b}}{k!} = e^{-b} \, , \tag{9.18}$$

or $b = -\log\beta$. For the upper limit at a confidence level of $1 - \beta = 95\%$ one has $b = -\log(0.05) = 2.996 \approx 3$. Thus if the number of occurrences of some rare event is treated as a Poisson variable with mean λ, and one looks for events of this type and finds none, then the 95% upper limit on the mean is 3. That is, if the mean were in fact $\lambda = 3$, then the probability to observe zero would be 5%.

9.5 Confidence Interval for Correlation Coefficient, Transformation of Parameters

In many situations one can assume that the p.d.f. for an estimator is Gaussian, and thus use the results of the previous section to obtain a confidence interval. As an example where this is often not the case, consider the correlation coefficient ρ of two continuous random variables x and y distributed according to a two-dimensional Gaussian p.d.f. $f(x, y)$ (equation (2.28)). Suppose we have a sample of n independent observations of x and y, and we would like to determine a confidence interval for ρ based on the estimator r (5.10)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2 \cdot \sum_{k=1}^{n}(y_k - \overline{y})^2\right)^{1/2}} \, . \tag{9.19}$$

The p.d.f. $g(r; \rho, n)$ has a rather complicated form; it is given e.g. in [Mui82] p. 151. A graph is shown in Fig. 9.5 for a sample of size $n = 20$ for several values of the true correlation coefficient ρ. One can see that $g(r; \rho, n)$ is asymmetric and that the degree of asymmetry depends on ρ. It can be shown that $g(r; \rho, n)$ approaches a Gaussian in the large sample limit, but for this approximation to be valid, one requires fairly large sample. (At least $n \geq 500$ is recommended [Bra92].) For smaller samples such as in Fig. 9.5, one cannot rely on the Gaussian approximation for $g(r; \rho, n)$, and thus one cannot use (9.12) to determine the confidence interval.

In principle one is then forced to return to the procedure of Section 9.2, which in this case would be quite difficult computationally. There exists, however, a much simpler method to determine an approximate confidence interval for ρ. It has been shown by Fisher that the p.d.f. of the statistic

$$z = \tanh^{-1} r = \frac{1}{2} \log \frac{1 + r}{1 - r} \tag{9.20}$$

approaches the Gaussian limit much more quickly as a function of the sample size n than that of r (see [Fis90] and references therein). This can be used as an estimator for ζ, defined as

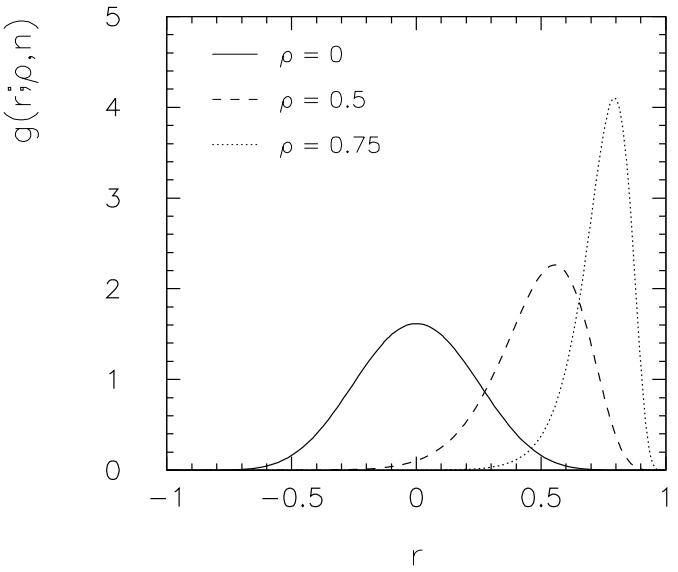

Figure 9.5: Probability density $f(r;\rho,n)$ for the estimator r of the correlation coefficient ρ for a sample of size $n = 20$, and various values of ρ.

$$\zeta = \tanh^{-1}\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho}\,. \tag{9.21}$$

One can show that the expectation value of z is approximately given by

$$E[z] = \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \tag{9.22}$$

and its variance by

$$V[z] = \frac{1}{n-3}\,. \tag{9.23}$$

We will assume that the sample is large enough that z has a Gaussian p.d.f. and that the bias term $\rho/2(n-1)$ in (9.22) can be neglected. Given a sample of n measurements of x and y, z can be determined according to equation (9.20) and its standard deviation $\hat{\sigma}_z$ can be estimated by using the variance from equation (9.23). One can use these to determine the interval $[z - \hat{\sigma}_z, z + \hat{\sigma}_z]$, or in general the interval $[a, b]$ given by (9.12). These give the lower limit a for ζ with confidence level $1 - \alpha$ and an upper limit b with confidence level $1 - \beta$. The confidence interval $[a, b]$ for $\zeta = \tanh^{-1}\rho$ can then be converted back to an interval $[A, B]$ for ρ simply by using the inverse of the transformation (9.20), i.e. $A = \tanh a$ and $B = \tanh b$.

Consider for example a sample of size $n = 20$ for which one has obtained the estimate $r = 0.5$. From equation (5.13) the standard deviation of r can be estimated as $\hat{\sigma}_r = (1 - r^2)/\sqrt{n} = 0.168$. If one were to make the incorrect approximation that r is Gaussian distributed for such a small sample, this would lead to a 68.3% central confidence interval for ρ of $[0.332, 0.668]$, or $[0.067, 0.933]$ at a confidence level of 99%.

Thus since the sample correlation coefficient r is almost three times the standard error $\hat{\sigma}_r$, one might be led to the incorrect conclusion that there is significant evidence for a non-zero value of ρ, i.e. a "$3\,\sigma$ effect". By using the Fisher z-transformation, however, one obtains $z = 0.549$ and $\hat{\sigma}_z = 0.243$. This corresponds to a 99% central confidence interval of $[-.075, 1.174]$ for ζ, and $[-.075, 0.826]$ for ρ. Thus the 99% central confidence interval includes zero.

Recall that the lower limit of the confidence interval is equal to the hypothetical value of the true parameter such that r would be observed higher than the one actually observed with the probability α. One can ask, for example, what the confidence level would be for a lower limit of zero. If we had assumed that $g(r; \rho, n)$ was Gaussian, the corresponding probability would be 0.14%. By using the z-transformation, however, the confidence level for a limit of zero is 2.3%, i.e. if ρ were zero one would obtain r greater than or equal to the one observed, $r = 0.5$, with a probability of 2.3%. The actual evidence for a non-zero correlation is therefore not nearly as strong as one would have concluded by simply using the standard error $\hat{\sigma}_r$ with the assumption that r is Gaussian.

9.6 Confidence Intervals Using the Likelihood Function or χ^2

Even in the case of a non-Gaussian estimator, the confidence interval can be determined with a simple approximate technique which makes use of the likelihood function (or equivalently the χ^2 function where one has $L = \exp(-\chi^2/2)$). Consider first a maximum-likelihood estimator $\hat{\theta}$ for a parameter θ in the large sample limit. In this limit it can be shown that the p.d.f. $g(\hat{\theta}; \theta)$ becomes Gaussian,

$$g(\hat{\theta}; \theta) = \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta} - \theta)^2}{2\sigma_{\hat{\theta}}^2}\right) , \tag{9.24}$$

centered about the true value of the parameter θ and with a standard deviation $\sigma_{\hat{\theta}}$.

Also in the large sample limit, one can show that the likelihood function itself becomes Gaussian in form centered about the ML estimate $\hat{\theta}$,

$$L(\theta) = L_{max} \exp\left(\frac{-(\theta - \hat{\theta})^2}{2\sigma_{\hat{\theta}}^2}\right) . \tag{9.25}$$

From the RCF inequality (6.17), which for an ML estimator in the large sample limit becomes an equality, one obtains that $\sigma_{\hat{\theta}}$ in the likelihood function (9.25) is the same as in the p.d.f. (9.24). This has already been encountered in Section 6.7, equation (6.25), where the likelihood function was used to estimate the variance of an estimator $\hat{\theta}$. This led to a simple prescription for estimating $\sigma_{\hat{\theta}}$, since by changing the parameter θ by N

standard deviations, the log-likelihood function decreases by $N^2/2$ from its maximum value,

$$\log L(\hat{\theta} \pm N\,\sigma_{\hat{\theta}}) = \log L_{max} - \frac{N^2}{2} \; . \tag{9.26}$$

From the results of the previous section, however, we know that for a Gaussian distributed estimator $\hat{\theta}$ the 68.3% central confidence interval can be constructed from the estimator and its estimated standard deviation $\hat{\sigma}_{\hat{\theta}}$ as $[a,b] = [\hat{\theta} - \hat{\sigma}_{\hat{\theta}}, \hat{\theta} + \hat{\sigma}_{\hat{\theta}}]$, (or more generally according to (9.12) for a confidence level of $1-\gamma$). The 68.3% central confidence interval is thus given by the values of θ at which the log-likelihood function decreases by 1/2 from its maximum value. (This is assuming, of course, that $\hat{\theta}$ is the ML estimator and thus corresponds to the maximum of the likelihood function.)

In fact, it can be shown that even if the likelihood function is not a Gaussian function of the parameters, the central confidence interval $[a,b] = [\hat{\theta} - c, \hat{\theta} + d]$ can still be approximated by using

$$\log L(\hat{\theta}^{+d}_{-c}) = \log L_{max} - \frac{N^2}{2} \; , \tag{9.27}$$

where $N = \Phi^{-1}(1-\gamma/2)$ is the quantile of the standard Gaussian corresponding to the desired confidence level $1-\gamma$. (For example, $N = 1$ for a 68.3% central confidence interval; see Table 9.1.) In the case of a least-squares fit with Gaussian errors, i.e. with $\log L = -\chi^2/2$, the prescription becomes

$$\chi^2(\hat{\theta}^{+d}_{-c}) = \chi^2_{min} + N^2 \; , \tag{9.28}$$

A heuristic proof that the intervals defined by equations (9.27) and (9.28) approximate the classical confidence intervals of Section 9.2 is given in [Ead71, Fro79]. Equations (9.27) and (9.28) represent one of the most commonly used methods for estimating statistical uncertainties. One should keep in mind, however, that the correspondence with the method of Section 9.2 is only exact in the large sample limit. Several authors (e.g. [Fro79, Hud64]) have recommended using the term "likelihood interval" for an interval obtained from the likelihood function. Regardless of the name, it should be kept in mind that it is interpreted here as an approximation to the classical confidence interval, i.e. a random interval constructed so as to include the true parameter value with a given probability.

As an example consider the estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$ for the parameter τ of an exponential distribution, as in the example of Section 6.2 (see also Section 6.7). There, the maximum-likelihood method was used to estimate τ given a sample of $n = 50$ measurements of an exponentially distributed random variable t. This sample was sufficiently large that the standard deviation $\sigma_{\hat{\tau}}$ could be approximated by the values of τ where the log-likelihood function decreased by 1/2 from its maximum (see Fig. 6.4). This gave $\hat{\tau} = 1.06$ and $\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$.

Figure 9.6 shows the log-likelihood function $\log L(\tau)$ as a function of τ for a sample of only $n = 5$ measurements of an exponentially distributed random variable, generated using the Monte Carlo method with the true parameter $\tau = 1$. Because of the smaller sample size the log-likelihood function is less parabolic than before.

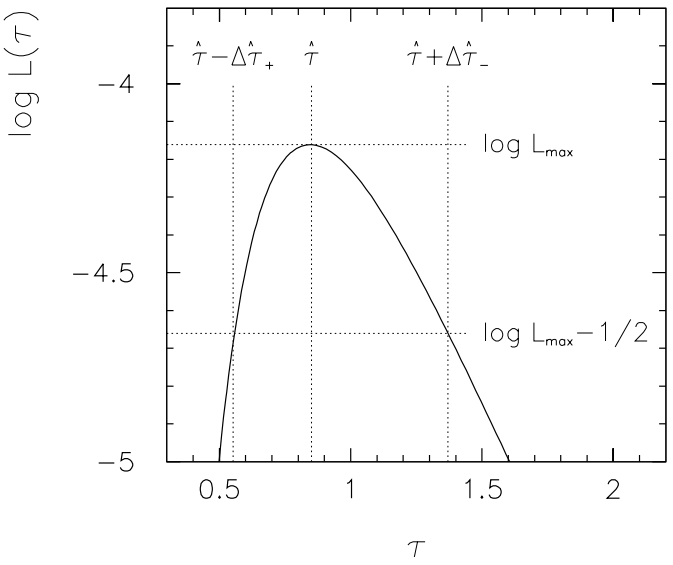

Figure 9.6: Log-likelihood function $\log L(\tau)$ as a function of τ for a sample of $n = 5$ measurements. The interval $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+]$ determined by $\log L(\tau) = \log L_{max} - 1/2$ can be used to approximate the 68.3% central confidence interval.

One could still use the half-width of the interval determined by $\log L_{max} - 1/2$ to approximate the standard deviation $\sigma_{\hat{\tau}}$, but this is not really what we want. The statistical uncertainty is better communicated by giving the confidence interval, since one then knows the probability that the interval covers the true parameter value. Furthermore, by giving a central confidence interval (and hence asymmetric errors, $\Delta\hat{\tau}_- \neq \Delta\hat{\tau}_+$), one has equal probabilities for the true parameter to be higher or lower than the interval limits. As illustrated in Fig. 9.6, the central confidence interval can be approximated by the values of τ where $\log L(\tau) = \log L_{max} - 1/2$, which gives $[\hat{\tau} - \Delta\hat{\tau}_-, \hat{\tau} + \Delta\hat{\tau}_+] = [0.55, 1.37]$ or $\hat{\tau} = 0.85^{+0.52}_{-0.30}$.

In fact, the same could have been done in Section 6.7 by giving the result there as $\hat{\tau} = 1.062^{+0.165}_{-0.137}$. Whether one chooses this method or simply reports an averaged symmetric error (i.e. $\hat{\tau} = 1.06 \pm 0.15$) will depend on how accurately the statistical error needs to be given. For the case of $n = 5$ shown in Fig. 9.6, the error bars are sufficiently asymmetric that one would probably want to use the 68.3% central confidence interval and give the result as $\hat{\tau} = 0.85^{+0.52}_{-0.30}$.

9.7 Multidimensional Confidence Regions

In Section 9.2, a confidence interval $[a, b]$ was constructed so as to have a certain probability $1 - \gamma$ of containing a parameter θ. In order to generalize this to the case of n parameters, $\vec{\theta} = (\theta_1, \ldots, \theta_n)$, one might attempt to find an n-dimensional confidence

interval $[\vec{a}, \vec{b}]$ constructed so as to have a given probability that $a_i < \theta_i < b_i$, simultaneously for all i. This turns out to be computationally difficult, and is rarely done.

It is nevertheless quite simple to construct a *confidence region* in the parameter space such that the true parameter $\vec{\theta}$ is contained within the region with a given probability (at least approximately). This region will not have the form $a_i < \theta_i < b_i$, $i = 1, \ldots, n$, but will be more complicated, approaching an n-dimensional hyperellipsoid in the large sample limit.

As in the single parameter case, one makes use of the fact that both the joint p.d.f. for the estimator $\vec{\hat{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ as well as the likelihood function become Gaussian in the large sample limit. That is, the joint p.d.f. of $\vec{\hat{\theta}}$ becomes

$$g(\vec{\hat{\theta}}|\vec{\theta}) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2} Q(\vec{\hat{\theta}}, \vec{\theta})\right] , \tag{9.29}$$

where Q is defined as

$$Q(\vec{\hat{\theta}}, \vec{\theta}) = (\vec{\hat{\theta}} - \vec{\theta})^T V^{-1} (\vec{\hat{\theta}} - \vec{\theta}) . \tag{9.30}$$

Here V^{-1} is the inverse covariance matrix and the superscript T indicates a transposed (i.e. row) vector. Contours of constant $g(\vec{\hat{\theta}}|\vec{\theta})$ correspond to constant $Q(\vec{\hat{\theta}}, \vec{\theta})$. These are ellipses (or for more than two dimensions, hyperellipsoids) in $\vec{\hat{\theta}}$-space centered about the true parameters $\vec{\theta}$. Figure 9.7(a) shows a contour of constant $Q(\vec{\hat{\theta}})$, where $\vec{\theta}_{true}$ represents a particular value of $\vec{\theta}$.

Also as in the one-dimensional case, one can show that the likelihood function $L(\vec{\theta})$ takes on a Gaussian form centered about the ML estimators $\vec{\hat{\theta}}$,

$$L(\vec{\theta}) = L_{max} \exp\left[-\frac{1}{2}(\vec{\theta} - \vec{\hat{\theta}})^T V^{-1} (\vec{\theta} - \vec{\hat{\theta}})\right] = L_{max} \exp\left[-\frac{1}{2} Q(\vec{\theta}, \vec{\hat{\theta}})\right] . \tag{9.31}$$

The inverse covariance matrix V^{-1} is the same here as in (9.29); this can be seen from the RCF inequality (6.20) and using the fact that the ML estimators attain the RCF bound in the large sample limit. The quantity Q here is regarded as a function of the parameters $\vec{\theta}$ which has its maximum at the estimates $\vec{\hat{\theta}}$. This is shown in Fig. 9.7(b) for $\vec{\hat{\theta}}$ equal to a particular value $\vec{\hat{\theta}}_{exp}$. Because of the symmetry between $\vec{\theta}$ and $\vec{\hat{\theta}}$ in the definition (9.30), the quantities Q have the same value in both the p.d.f. (9.29) and in the likelihood function (9.31), i.e. $Q(\vec{\hat{\theta}}, \vec{\theta}) = Q(\vec{\theta}, \vec{\hat{\theta}})$.

As discussed in Section 7.5, it can be shown that if $\vec{\hat{\theta}}$ is described by an n-dimensional Gaussian p.d.f. $g(\vec{\hat{\theta}}, \vec{\theta})$, then the quantity $Q(\vec{\hat{\theta}}, \vec{\theta})$ is distributed according to a χ^2-distribution for n degrees of freedom. The statement that $Q(\vec{\hat{\theta}}, \vec{\theta})$ is less than some

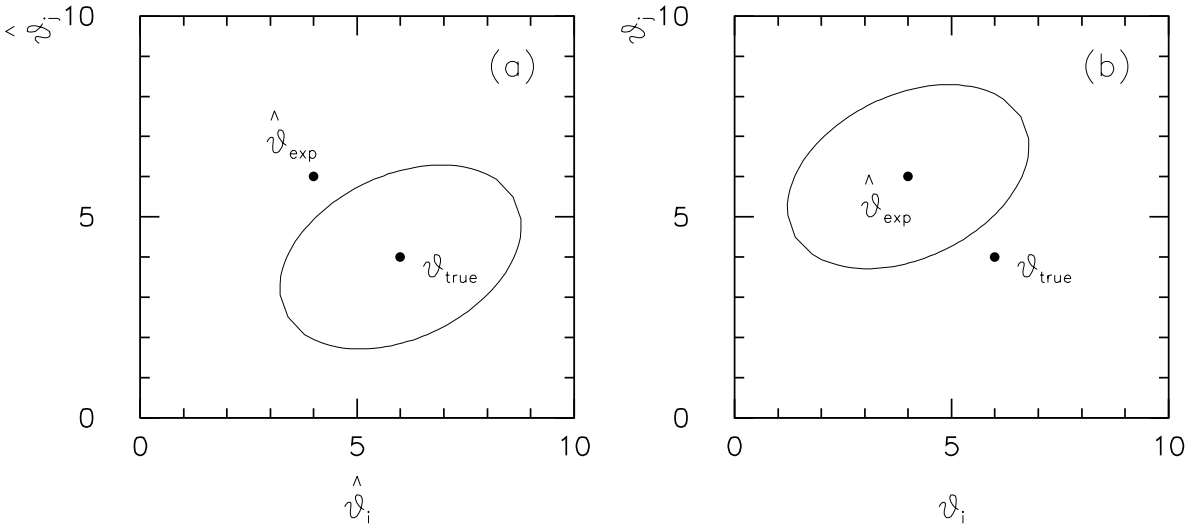

Figure 9.7: (a) A contour of constant $g(\vec{\hat{\theta}};\vec{\theta}_{true})$ (i.e. constant $Q(\vec{\hat{\theta}},\vec{\theta}_{true})$) in $\vec{\hat{\theta}}$-space. (b) A contour of constant $L(\vec{\theta})$ corresponding to constant $Q(\vec{\hat{\theta}}_{exp},\vec{\theta})$ in $\vec{\theta}$-space. The values $\vec{\theta}_{true}$ and $\vec{\hat{\theta}}_{exp}$ represent particular (i.e. constant) values of $\vec{\theta}$ and $\vec{\hat{\theta}}$, respectively

value Q_γ, i.e. that the estimate is within a certain distance of the true value $\vec{\theta}$, implies $Q(\vec{\theta},\vec{\hat{\theta}}) < Q_\gamma$, i.e. that the true value θ is within the same distance of the estimate. The two events therefore have the same probability,

$$P(Q(\vec{\theta},\vec{\hat{\theta}}) \leq Q_\gamma) = \int_0^{Q_\gamma} f(z;n)dz \ , \tag{9.32}$$

where $f(z;n)$ is the χ^2 distribution for n degrees of freedom (equation (2.29)). The value Q_γ is chosen to correspond to a given probability content,

$$\int_0^{Q_\gamma} f(z;n)dz = 1-\gamma \ . \tag{9.33}$$

That is,

$$Q_\gamma = F^{-1}(1-\gamma;n) \tag{9.34}$$

is the quantile of order $1-\gamma$ of the χ^2-distribution. The region of $\vec{\theta}$-space defined by $Q(\vec{\theta},\vec{\hat{\theta}}) \leq Q_\gamma$ is called a *confidence region* with the confidence level $1-\gamma$. For a likelihood function of Gaussian form (9.31) it can be constructed by finding the values of $\vec{\theta}$ at which the log-likelihood function decreases by $Q_\gamma/2$ from its maximum value,

$$\log L(\vec{\theta}) = \log L_{max} - \frac{Q_\gamma}{2} \ . \tag{9.35}$$

As in the single parameter case, one can still use the prescription given by (9.35) even if the likelihood function is not Gaussian, in which case the probability statement (9.32) is only approximate. For an increasing number of parameters, the approach to the Gaussian limit becomes slower as a function of the sample size, and furthermore it is difficult to quantify when a sample is large enough for (9.32) to apply. If needed, one can determine the probability that a region constructed according to (9.35) includes the true parameter by means of a Monte Carlo calculation.

Quantiles of the χ^2-distribution $Q_\gamma = f^{-1}(1-\gamma; n)$ for several confidence levels $1-\gamma$ and $n = 1, 2, 3, 4, 5$ parameters are given in Table 9.3. Values of the confidence level are given for various values of the quantile Q_γ are given in Table 9.4.

Q_γ	$1-\gamma$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.394	0.199	0.090	0.037
2.0	0.843	0.632	0.427	0.264	0.151
4.0	0.955	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Table 9.3: The values of the confidence level $1-\gamma$ for different values of Q_γ and for $n = 1, 2, 3, 4, 5$ fitted parameters.

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.64	9.21	11.3	13.3	15.1

Table 9.4: The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1, 2, 3, 4, 5$ fitted parameters.

For $n = 1$ the expression (9.34) for Q_γ can be shown to imply

$$\sqrt{Q_\gamma} = \Phi^{-1}(1-\gamma/2)\,, \tag{9.36}$$

where Φ^{-1} is the inverse function of the standard normal distribution. The procedure here thus reduces to that for a single parameter given in Section 9.6, where $N = \sqrt{Q_\gamma}$ is the half width of the interval in standard deviations (see equations (9.26), (9.27)). The values for $n = 1$ in Tables 9.3 and 9.4 are thus related to those in Tables 9.1 and 9.2 by equation (9.36).

For increasing n, the confidence level for a given Q_γ decreases. For example, in the single parameter case, $Q_\gamma = 1$ corresponds to $1-\gamma = 0.683$. For $n = 2$, $Q_\gamma = 1$ gives a confidence level of only 0.394, and in order to obtain $1-\gamma = 0.683$ one needs $Q_\gamma = 2.30$.

We should emphasize that, as in the single parameter case, the confidence region $Q(\vec{\theta}, \hat{\vec{\theta}}) \leq Q_\gamma$ is a random region in $\vec{\theta}$-space. The confidence region varies upon repetition of the experiment, since $\hat{\vec{\theta}}$ is a random variable. The true parameters, on the other hand, are unknown constants.

9.8 Bayesian Intervals

An alternative approach to quantifying statistical uncertainty is by use of subjective probability as introduced in Section 1.2. Here both the result of a measurement x and a parameter θ are treated as random variables. One's knowledge about θ is summarized by its probability density $p(\theta)$ which gives the degree of belief that θ has a given value.

Consider again the situation above with n observations of a random variable x, $x_1, \ldots, x_n$, assumed to be distributed according to some p.d.f. $f(x;\theta)$ which depends on an unknown parameter θ. (The Bayesian approach can easily be generalized to several parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. For simplicity we will consider here only a single parameter.) Recall that the likelihood function is the joint p.d.f. for the data $\vec{x} = (x_1, \ldots, x_n)$ for a given value of θ, and thus can be written

$$L(\vec{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta) \; . \tag{9.37}$$

What we would like is the conditional p.d.f. for θ given the data $p(\theta|\vec{x})$. This is obtained from the likelihood via Bayes' theorem (equation (1.25))

$$p(\theta|\vec{x}) = \frac{L(\vec{x}|\theta)\,\pi(\theta)}{\int L(\vec{x}|\theta')\,\pi(\theta')d\theta'} \; , \tag{9.38}$$

where $\pi(\theta)$ is the *prior* probability density for θ, reflecting the state of knowledge of θ before consideration of the data. $p(\theta|\vec{x})$ is called the *posterior* probability density for θ given the data $\vec{x}$.

In Bayesian statistics all information about θ is contained in the posterior p.d.f. $p(\theta|\vec{x})$. Since it is rarely practical to report the entire p.d.f., especially when θ is multidimensional, an appropriate way of summarizing it must be found. The first step in this direction is an estimator, which clearly should be the value of θ at which $p(\theta|\vec{x})$ is a maximum. In practice Bayesian estimators are not used much in the physical sciences, with the classical methods of maximum likelihood and least squares being more widely accepted. Note, however, that if the prior p.d.f. $\pi(\theta)$ is taken to be a constant, then $p(\theta|\vec{x})$ is proportional to the likelihood function $L(\vec{x}|\theta)$ and the Bayesian and ML estimators coincide. As long as $\pi(\theta)$ is relatively flat compared to $L(\vec{x}|\theta)$, this statement still holds approximately.

In addition to giving an estimator of the single most probable value of θ, the posterior density[3] $p(\theta)$ can be summarized by giving an interval $[a, b]$ such that for given probabilities α and β one has

$$\begin{aligned} \alpha &= \int_{-\infty}^{a} p(\theta|\vec{x})\, d\theta \\ \beta &= \int_{b}^{\infty} p(\theta|\vec{x})\, d\theta \,. \end{aligned} \tag{9.39}$$

Choosing $\alpha = \beta$ then gives a central interval, with e.g. $1 - \alpha - \beta = 68.3\%$. Another possibility is to choose α and β such that all values of $p(\theta)$ inside the interval $[a, b]$ are higher than any values outside, which implies $p(a) = p(b)$. One can easily show that this gives the shortest possible interval.

The Bayesian approach expressed by equation (9.38) gives a method for updating one's state of knowledge in light of newly acquired data. To do this, however, one must specify what the state of knowledge was before the measurement via the prior density $\pi(\theta)$. If nothing is known previously, one may assume that all values of θ are equally likely. This assumption is sometimes called *Bayes' postulate*, expressed here by $\pi(\theta) = constant$. If the range of θ is infinite then a constant $\pi(\theta)$ cannot be normalized, and is called an *improper* prior. This is usually not, in fact, a problem since $\pi(\theta)$ always appears multiplied by the likelihood function, resulting in a normalizable posterior p.d.f. For some improper prior densities this may not always be the case; see e.g. equation (9.44) in the next section.

In cases where θ can only take on discrete values, the use of Bayes' postulate is unambiguously defined. If θ is continuous, however, the situation is more difficult. Suppose one has a continuous parameter θ defined in the interval $[0, 10]$. One would then take the prior p.d.f. $\pi_\theta(\theta) = 0.1$ in equation (9.38) to get the posterior density $p_\theta(\theta)$. Another experimenter, however, could decide that some nonlinear function $a(\theta)$ was more appropriate as the parameter. Using the techniques for transformation of variables, one could find the corresponding density $p_a(a) = p_\theta(\theta)|d\theta/da|$. On the other hand, one could express the likelihood function directly in terms of a, and assume that the prior density $\pi_a(a)$ is constant. For example, if $a = \theta^2$, then $\pi_a(a) = 0.01$ in the interval $[0, 100]$. Using this in equation (9.38), however, would lead to a posterior density in general different from the $p_a(a)$ obtained by transformation of variables. That is, complete ignorance about θ ($\pi_\theta(\theta) = constant$) implies a nonuniform prior density for a nonlinear function of θ ($\pi_a(a) \neq constant$).

An important case where Bayesian intervals have proven useful is when one has objective prior information about the value of a parameter, such as a physical boundary. With classical confidence intervals there is no easy way of incorporating such information, whereas this is straightforward when using the Bayesian approach. This situation is treated in the next section.

[3] In some cases we will suppress reference to the data $\vec{x}$ in the posterior p.d.f. and simply write $p(\theta)$. The conditional probability for θ given $\vec{x}$ is implied.

9.9 Limits Near a Physical Boundary

Often the purpose of an experiment is to search for a new effect, the existence of which would imply that a certain parameter is not equal to zero. For example, one could attempt to measure the mass of the neutrino, which in the standard theory is massless. If the data yield a value of the parameter significantly different from zero, then the new effect has been discovered, and the parameter's value and a confidence interval to reflect its error are given as the result. If, on the other hand, the data result in a fitted value of the parameter that is consistent with zero, then the result of the experiment is reported by giving an upper limit on the parameter. (A similar situation occurs when absence of the new effect corresponds to a parameter being large or infinite; one then places a lower limit. For simplicity we will consider here only upper limits.)

If there are no restrictions on the possible values of the parameter, then the classical and Bayesian techniques described in the previous sections will lead to similar (or identical) results, albeit with differences in their interpretation. A significant difference in the two approaches becomes evident, however, if the parameter is only allowed to take on values in a restricted range. In particle physics, for example, this is the case with the neutrino mass mentioned above and with quantities such as cross sections and particle lifetimes, the true values of which must be positive (or zero) by definition.

The difficulty arises when an estimator can take on values in the excluded region. This can occur if the estimator $\hat{\theta}$ for a parameter θ is of the form $\hat{\theta} = x - y$, where both x and y are random variables, i.e. they have random measurement errors. The mass squared of a particle, for example, can be estimated by measuring independently its energy E and momentum p, and using $\widehat{m^2} = E^2 - p^2$. Although the mass squared should come out positive, measurement errors in E^2 and p^2 could result in a negative value for $\widehat{m^2}$. Then the question is how to place a limit on m^2, or more generally on a parameter θ when the estimate is in or near an excluded region.

Consider further the example of an estimator $\hat{\theta} = x - y$ where x and y are Gaussian variables with means μ_x, μ_y and variances σ_x^2, σ_y^2. One can show that the difference $\hat{\theta} = x - y$ is also a Gaussian variable with $\theta = \mu_x - \mu_y$ and $\sigma_{\hat{\theta}}^2 = \sigma_x^2 + \sigma_y^2$. (This can easily be shown using characteristic functions as described in Chapter 11.)

Assume that θ is known *a priori* to be non-negative (e.g. like the mass squared), and suppose the experiment has resulted in a value $\hat{\theta}_{exp}$ for the estimator $\hat{\theta}$. According to (9.12), the upper limit θ_{up} at a confidence level $1 - \beta$ is

$$\theta_{up} = \hat{\theta}_{exp} + \sigma_{\hat{\theta}} \, \Phi^{-1}(1 - \beta) \, . \tag{9.40}$$

For the commonly used 95% confidence level one has from Table 9.2 $\Phi^{-1}(0.95) = 1.645$.

The interval $(-\infty, \theta_{up}]$ is constructed to include the true value θ with a probability of 95%, regardless what θ actually is. Suppose now that the standard deviation is $\sigma_{\hat{\theta}} = 1$, and one obtains $\hat{\theta}_{exp} = -2.0$. From equation (9.40) one obtains $\theta_{up} = -0.355$ at a confidence level of 95%. Not only is $\hat{\theta}_{exp}$ in the forbidden region (as half of the estimates should be

if θ is really zero) but the upper limit, i.e. the entire confidence interval, is below zero as well. This is not particularly unusual, and in fact is expected to happen in 5% of the experiments if the true value of θ is zero.

As far as the definition of the confidence interval is concerned, nothing fundamental has gone wrong. The interval was designed to cover the true value of θ in a certain fraction of repeated experiments, and we have obviously obtained one of those experiments where θ is not in the interval. But this is not a very satisfying result, since it was already known *a priori* that θ is greater than zero (and certainly greater than $\theta_{up} = -0.355$) without having to perform the experiment.

Regardless of the upper limit, it is important to report the actual value of the estimate obtained and its standard deviation, i.e. $\hat{\theta}_{exp} \pm \sigma_{\hat{\theta}}$, even if the estimate is in the physically excluded region. In this way, the average of many experiments (e.g. as in Section 7.6) will converge to the correct value (as long as the estimator is unbiased). In cases where the p.d.f. of $\hat{\theta}$ is significantly non-Gaussian, the entire likelihood function $L(\theta)$ should be given, which can be combined with that of other experiments as discussed in Section 6.11.

Nevertheless, most experimenters want to report some sort of upper limit, and in situations such as the one described above a number of techniques have been proposed (see e.g. [Hig83, Jam91]). There is unfortunately no established convention on how this should be done, and one should therefore state what procedure was used.

As a solution to the difficulties posed by an upper limit in an unphysical region, one might be tempted to simply increase the confidence level until the limit enters the allowed region. In the previous example, if we had taken a confidence level $1 - \beta = 0.99$, then from Table 9.2 one has $\Phi^{-1}(0.99) = 2.326$, giving $\theta_{up} = 0.326$. This would lead one to quote an upper limit that is smaller than the intrinsic resolution of the experiment ($\sigma_{\hat{\theta}} = 1$) at a very high confidence level of 99%, which is clearly misleading. Worse, of course, would be to adjust the confidence level to give an arbitrarily small limit, e.g. $\Phi^{-1}(0.97725) = 2.00001$, or $\theta_{up} = 10^{-5}$ at a confidence level of 97.725%!

In order to avoid this type of difficulty, a commonly used technique is to simply shift a negative estimate to zero before applying equation (9.40), i.e.

$$\theta_{up} = \max(\hat{\theta}_{exp}, 0) + \sigma_{\hat{\theta}} \, \Phi^{-1}(1 - \beta) \, . \tag{9.41}$$

In this way the upper limit is always at least the same order of magnitude as resolution of the experiment. If $\hat{\theta}_{exp}$ is positive, the limit coincides with that of the classical procedure. This technique has a certain intuitive appeal and is often used, but the interpretation as an interval that will cover the true parameter value with probability $1 - \beta$ no longer applies. The coverage probability is clearly greater than $1 - \beta$, since the shifted upper limit (9.41) is in all cases greater than or equal to the classical one (9.40).

Another alternative is to report a Bayesian upper limit as discussed in Section 9.8. Here one has the advantage that prior knowledge, e.g. $\theta \geq 0$, can easily be incorporated by setting the prior p.d.f. $\pi(\theta)$ to zero in the excluded region. Bayes' theorem then gives

a posterior probability $p(\theta)$ with $p(\theta) = 0$ for $\theta < 0$. The upper limit is thus determined by

$$1 - \beta = \int_0^{\theta_{up}} p(\theta) d\theta = \frac{\int_0^{\theta_{up}} L(\theta)\, \pi(\theta)\, d\theta}{\int_0^{\infty} L(\theta)\, \pi(\theta)\, d\theta} \,. \tag{9.42}$$

The difficulties here have already been mentioned in Section 9.8, namely, that there is no unique way to specify the prior density $\pi(\theta)$. A common choice is

$$\pi(\theta) = \begin{cases} 0 & \theta < 0 \\ 1 & \theta \geq 0 \end{cases} \,. \tag{9.43}$$

The prescription says in effect to normalize the likelihood function to unit area in the physical region, and then integrate it out to θ_{up} such that the fraction of area covered is $1 - \beta$. This procedure has been recommended by, among others, the Particle Data Group [PDG94]. Although the method is simple, it has some conceptual drawbacks. For the case where one knows $\theta \geq 0$ (e.g. the neutrino mass) one does not really believe that $0 < \theta < 1$ has the same prior probability as $10^{40} < \theta < 10^{40} + 1$. Furthermore, the upper limit derived from $\pi(\theta) = constant$ is not invariant with respect to a nonlinear transformation of the parameter.

It has been argued [Jef48] that in cases where $\theta \geq 0$ but with no other prior information, one should use

$$\pi(\theta) = \begin{cases} 0 & \theta \leq 0 \\ \frac{1}{\theta} & \theta > 0 \end{cases} \,. \tag{9.44}$$

This has the advantage that upper limits are invariant with respect to a transformation of the parameter by raising to an arbitrary power. This is equivalent to a uniform (improper) prior of the form (9.43) for $\log \theta$. It is unusable, however, for the case discussed here, since the integrals in (9.42) diverge. Therefore, despite its conceptual difficulties, the uniform prior density is the most commonly used choice for setting limits on parameters.

Figure 9.8 shows the upper limits at 95% confidence level derived according to the classical, shifted, and Bayesian techniques as a function of $\hat{\theta}_{exp} = x - y$ for $\sigma_{\hat{\theta}} = 1$. For the Bayesian limit, a prior density $\pi(\theta) = constant$ was used. The shifted and classical techniques are equal for $\hat{\theta}_{exp} \geq 0$. The Bayesian limit is always positive, and always greater than or equal to the classical limit. As $\hat{\theta}_{exp}$ becomes larger than the experimental resolution $\sigma_{\hat{\theta}}$, the Bayesian and classical limits rapidly approach each other.

9.10 Upper Limit on the Mean of Poisson Variable with Background

As a final example recall Section 9.4 where an upper limit was placed on the mean λ of a Poisson variable k. Often one is faced with a somewhat more complicated situation where

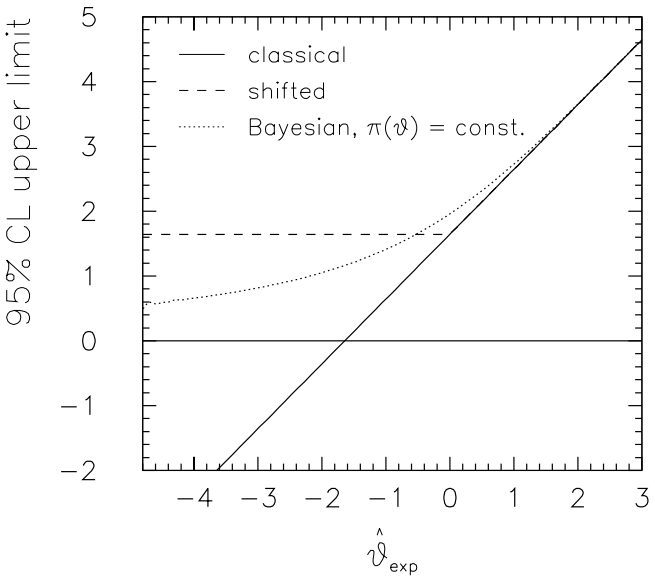

Figure 9.8: Upper limits at 95% confidence level for the example of Section 9.9 using the classical, shifted and Bayesian techniques. The shifted and classical techniques are equal for $\hat{\theta}_{exp} \geq 0$.

the observed value of k is the sum of the desired signal events k_s as well as background events k_b,

$$k = k_s + k_b \; , \tag{9.45}$$

where both k_s and k_b can be regarded as Poisson variables with means λ_s and λ_b, respectively. Suppose for the moment that the mean for the background λ_b is known without any uncertainty. For λ_s one only knows *a priori* that $\lambda_s \geq 0$. The goal is to construct an upper limit for the signal parameter λ_s given a measured value of k.

Since k is the sum of two Poisson variables, one can show that it is itself a Poisson variable, with the probability function,

$$f(k; \lambda_s, \lambda_b) = \frac{(\lambda_s + \lambda_b)^k}{k!} \, e^{-(\lambda_s + \lambda_b)} \; . \tag{9.46}$$

The maximum likelihood estimator for λ_s is

$$\hat{\lambda}_s = k - \lambda_b \; , \tag{9.47}$$

which clearly has zero bias since $E[k] = \lambda_s + \lambda_b$. Equations (9.15) used to determine the confidence interval become

$$\begin{aligned} \alpha &= P(\hat{\lambda}_s \geq \hat{\lambda}_s^{exp}; \lambda_s^{lo}) = \sum_{k \geq k_{exp}} \frac{(\lambda_s^{lo} + \lambda_b)^k \, e^{-(\lambda_s^{lo} + \lambda_b)}}{k!} \; , \\ \beta &= P(\hat{\lambda}_s \leq \hat{\lambda}_s^{exp}; \lambda_s^{up}) = \sum_{k \leq k_{exp}} \frac{(\lambda_s^{up} + \lambda_b)^k \, e^{-(\lambda_s^{up} + \lambda_b)}}{k!} \; , \end{aligned} \tag{9.48}$$

which can be solved numerically for the lower and upper limits λ_s^{lo} and λ_s^{up}. Comparing with the case $\lambda_b = 0$, one sees that the limits from (9.48) are related to what would be obtained without background simply by

$$\begin{aligned} \lambda_s^{lo} &= \lambda_s^{lo}(\text{no background}) - \lambda_b \ , \\ \lambda_s^{up} &= \lambda_s^{up}(\text{no background}) - \lambda_b \ . \end{aligned} \tag{9.49}$$

The difficulties here are similar to those encountered in the previous example. The problem occurs when the total number of events observed k_{exp} is not large compared to the expected number of background events λ_b. Values of λ_s^{up} for $1 - \beta = 0.95$ are shown in Fig. 9.9(a) as a function of the expected number of background events λ_b. For small enough k_{exp} and a high enough background level λ_b, a non-negative solution for λ_s^{up} does not exist. This situation can occur, of course, because of fluctuations in k_s and k_b.

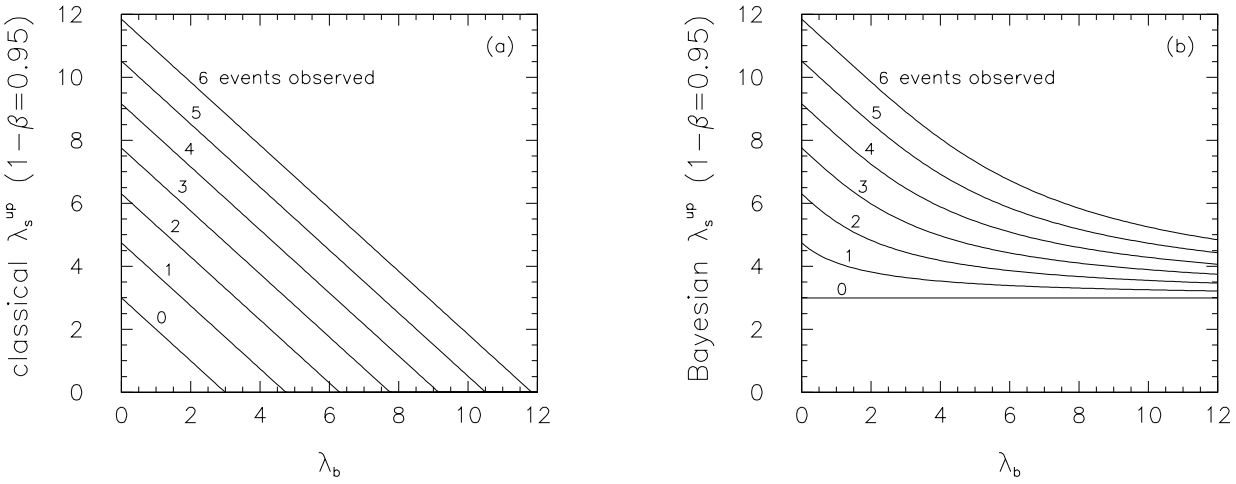

Figure 9.9: Upper limits λ_s^{up} at a confidence level of $1 - \beta = 0.95$ for different numbers of events observed k_{exp} and as a function of the expected number of background events λ_b. (a) The classical limit. (b) The Bayesian limit based on a uniform prior density for λ_s.

Because of these difficulties, the classical limit is not recommended in this case. As previously mentioned, one should always report $\hat{\lambda}_s$ and an estimate of its variance even if it comes out negative. In this way the average of many experiments will converge to the correct value. If, in addition, one wishes to report an upper limit on λ_s, the Bayesian method can be used with e.g. a uniform prior density [Hel83]. The likelihood function is given by the probability (9.46), now regarded as a function of λ_s,

$$L(\lambda_s) = \frac{(\lambda_s + \lambda_b)^k}{k!} \, e^{-(\lambda_s + \lambda_b)} \ . \tag{9.50}$$

The posterior probability density for λ_s is obtained as usual from Bayes' theorem,

$$p(\lambda_s) = \frac{L(\lambda_s)\,\pi(\lambda_s)\,d\lambda_s}{\int_0^\infty L(\lambda_s')\,\pi(\lambda_s')\,d\lambda_s'}\,. \tag{9.51}$$

Taking $\pi(\lambda_s)$ to be constant for $\lambda_s \geq 0$ and zero for $\lambda_s < 0$, the upper limit λ_s^{up} is given by

$$\begin{aligned} 1-\beta &= \frac{\int_0^{\lambda_s^{up}} L(\lambda_s)\,\pi(\lambda_s)\,d\lambda_s}{\int_0^\infty L(\lambda_s')\,\pi(\lambda_s')\,d\lambda_s'} \\ &= \frac{\int_0^{\lambda_s^{up}} (\lambda_s+\lambda_b)^{k_{exp}}\,e^{-(\lambda_s+\lambda_b)}\,d\lambda_s}{\int_0^\infty (\lambda_s+\lambda_b)^{k_{exp}}\,e^{-(\lambda_s+\lambda_b)}\,d\lambda_s}\,. \end{aligned} \tag{9.52}$$

The integrals can be related to incomplete gamma functions (see e.g. [Arf70]) allowing equation (9.52) to be expressed as

$$\beta = \frac{e^{-(\lambda_s+\lambda_b)}\,\sum_{k=0}^{k_{exp}} \frac{(\lambda_s^{up}+\lambda_b)^k}{k!}}{e^{-\lambda_b}\,\sum_{k=0}^{k_{exp}} \frac{\lambda_b^k}{k!}}\,. \tag{9.53}$$

This can be solved numerically for the upper limit λ_s^{up}. The upper limit as a function of λ_b is shown in Fig. 9.9(b) for various values of k_{exp}. For the case without background, setting $\lambda_b = 0$ gives

$$\beta = e^{-\lambda_s^{up}} \sum_{k=0}^{k_{exp}} \frac{(\lambda_s^{up})^k}{k!}\,, \tag{9.54}$$

which is identical to the equation for the classical upper limit (9.16). This can be seen by comparing Figs. 9.9(a) and (b). The Bayesian limit is always greater than or equal to the corresponding classical one, with the two agreeing only for $\lambda_b = 0$.

The agreement for the case without background must be considered accidental, however, since the Bayesian limit depends on the particular choice of a constant prior density $\pi(\lambda_s)$. Nevertheless, the coincidence spares one the trouble of having to defend either the classical or Bayesian viewpoint, which may account for the general acceptance of the uniform prior density in this case.

Chapter 10

Characteristic Functions and Related Examples

10.1 Definition and Properties of the Characteristic Function

The *characteristic function* $\phi_x(k)$ for a random variable x with p.d.f. $f(x)$ is defined as the expectation value of e^{ikx},

$$\phi_x(k) = E[e^{ikx}] = \int_{-\infty}^{\infty} e^{ikx} f(x) dx \ . \tag{10.1}$$

This is essentially the Fourier transform of the probability density function. It is useful in proving a number of important theorems, in particular those involving sums of random variables. Some characteristic functions of important p.d.f.'s are given in Table 10.1. Further examples can be found in [Ead71] Chapter 4.

Suppose one has n independent random variables $x_1, \ldots, x_n$, with p.d.f.'s $f_1(x_1), \ldots, f_n(x_n)$, and corresponding characteristic functions $\phi_1(k), \ldots, \phi_n(k)$, and consider the sum $z = \sum_i x_i$. The characteristic function $\phi_z(k)$ for z is related to those of the x_i by

$$\begin{aligned} \phi_z(k) &= \int \cdots \int \exp\left(ik \sum_{i=1}^{n} x_i \right) f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n \\ &= \int e^{ikx_1} f_1(x_1) dx_1 \cdots \int e^{ikx_n} f_n(x_n) dx_n \\ &= \phi_1(k) \cdots \phi_n(k) \ . \end{aligned} \tag{10.2}$$

That is, the characteristic function for a sum of random variables is given by the product of the individual characteristic functions.

Distribution	p.d.f.	characteristic function $\phi(k)$
Binomial	$f(n;N,p) = \frac{N!}{n!(N-n)!}\, p^n\, (1-p)^{N-n}$	$(p(e^{ik}-1)+1)^N$
Poisson	$f(n;\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$	$\exp(\lambda(e^{ik}-1))$
Uniform	$f(x;a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{ibk}-e^{iak}}{(b-a)ik}$
Exponential	$f(x;\xi) = \frac{1}{\xi} e^{-x/\xi}$	$\frac{1}{1-ik\xi}$
Gaussian	$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	$\exp(i\mu k - \frac{1}{2}\sigma^2 k^2)$
Chi-Square	$f(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$	$(1-2ik)^{-n/2}$
Cauchy	$f(x) = \frac{1}{\pi}\, \frac{1}{1+x^2}$	$e^{-\lvert k\rvert}$

Table 10.1: Characteristic functions for several commonly used probability functions.

To find the p.d.f. $f(z)$ one must compute the inverse Fourier transform,

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(k)\, e^{-ikz}\, dk\,. \tag{10.3}$$

Even if one is unable to invert the transform to find $f(z)$, one can easily determine its moments. Differentiating the characteristic function m times gives

$$\begin{aligned} \frac{d^m}{dk^m}\phi_z(k)\bigg|_{k=0} &= \frac{d^m}{dk^m} \int e^{ikz} f(z) dz \bigg|_{k=0} \\ &= i^m \int z^m\, f(z) dz \\ &= i^m\, \mu'_m \end{aligned} \tag{10.4}$$

where $\mu'_m = E[z^m]$ is the mth algebraic moment of z. One can use this, for example, to show that the mean and variance of the Gaussian distribution are

$$
\begin{aligned}
E[x] &= -i\frac{d}{dk}(\exp(i\mu k - \tfrac{1}{2}\sigma^2 k^2))\Big|_{k=0} = \mu \\
V[x] &= E[x^2] - (E[x])^2 \\
&= -\frac{d^2}{dk^2}(\exp(i\mu k - \tfrac{1}{2}\sigma^2 k^2))\Big|_{k=0} - \mu^2 = \sigma^2 \, . \qquad (10.5)
\end{aligned}
$$

The property (10.2) allows us to prove a number of results that have been used already in previous chapters. For example, consider the sum z of two Gaussian random variables x and y with means μ_x, μ_y and variances σ_x^2, σ_y^2. According to (10.2) the characteristic function for z is related to those of x and y by

$$
\begin{aligned}
\phi_z(k) &= \phi_x(k)\,\phi_y(k) \\
&= \exp(i\mu_x k - \tfrac{1}{2}\sigma_x^2 k^2) \cdot \exp(i\mu_y k - \tfrac{1}{2}\sigma_y^2 k^2) \\
&= \exp(i(\mu_x + \mu_y)k \; - \; \tfrac{1}{2}(\sigma_x^2 + \sigma_y^2)) \, . \qquad (10.6)
\end{aligned}
$$

This shows that z is itself a Gaussian random variable with mean $\mu_z = \mu_x + \mu_y$ and variance $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$. The corresponding property for the difference of two Gaussian variables was used in the example of Section 9.9.

In a similar way one can show that the sum of Poisson variables with means λ_i is itself a Poisson variable with mean $\sum_i \lambda_i$. Also using (10.2) one can show that for n independent Gaussian random variables x_i with means μ_i and variances σ_i^2, the sum of squares

$$
z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \qquad (10.7)
$$

follows a χ^2-distribution for n degrees of freedom. A proof of the Central Limit Theorem based on similar arguments is given in [Bra92] Chapter 5.

10.2 Use of Characteristic Function to Find p.d.f. of an Estimator

Consider n independent observations of a random variable x from an exponential distribution $f(x;\xi) = (1/\xi)\exp(-x/\xi)$. In Section 6.2 it was seen that the maximum likelihood estimator $\hat{\xi}$ for ξ was the sample mean of the observed x_i:

$$
\hat{\xi} = \frac{1}{n}\sum_{i=1}^{n} x_i \, . \qquad (10.8)
$$

If the experiment were repeated many times one would obtain values of $\hat{\xi}$ distributed according to a p.d.f. $g(\hat{\xi}; n, \xi)$ which depends on the number of observations per experiment n and the true value of the parameter ξ.

Suppose one wants to find $g(\hat{\xi}; n, \xi)$. The characteristic function for x is

$$\begin{aligned} \phi_x(k) &= \int e^{ikx} f(x) dx \\ &= \int_0^\infty e^{ikx} \frac{1}{\xi} e^{-x/\xi} dx \\ &= \frac{1}{1 - ik\xi} \, . \end{aligned} \tag{10.9}$$

Applying equation (10.2) for the sum $z = \sum_{i=1}^{n} x_i = n\hat{\xi}$ gives

$$\phi_z(k) = \frac{1}{(1 - ik\xi)^n} \, . \tag{10.10}$$

The p.d.f. $g_z(z)$ for z is found by computing the inverse Fourier transform of $\phi_z(k)$,

$$g_z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ikz}}{(1 - k\xi)^n} \, dk \, . \tag{10.11}$$

The integrand has a pole of order n at $-i/\xi$ in the complex k plane. Closing the contour in the lower half plane and using the residue theorem gives

$$g_z(z) = \frac{1}{(n-1)!} \frac{z^{n-1}}{\xi^n} e^{-z/\xi} \, . \tag{10.12}$$

Transforming to find p.d.f. for the estimator $\hat{\xi} = z/n$ gives

$$\begin{aligned} g(\hat{\xi}; n, \xi) &= g_z(z) \left| dz/d\hat{\xi} \right| \\ &= n g_z(n\hat{\xi}) \\ &= \frac{n^n}{(n-1)!} \frac{\hat{\xi}^{n-1}}{\xi^n} e^{-n\hat{\xi}/\xi} \, , \end{aligned} \tag{10.13}$$

which is a special case of the gamma distribution (see e.g. [Ead71] Chapter 4). Figure 10.1 shows the distribution $g(\hat{\xi}; n, \xi)$ for several values of the parameters. For $n = 5$ measurements one sees that the p.d.f. is roughly centered about the true value ξ, but has a long tail extending to higher values of $\hat{\xi}$. In Fig. 10.1(b) one sees that the p.d.f. becomes approximately Gaussian as the number of measurements n increases, as required by the Central Limit Theorem.

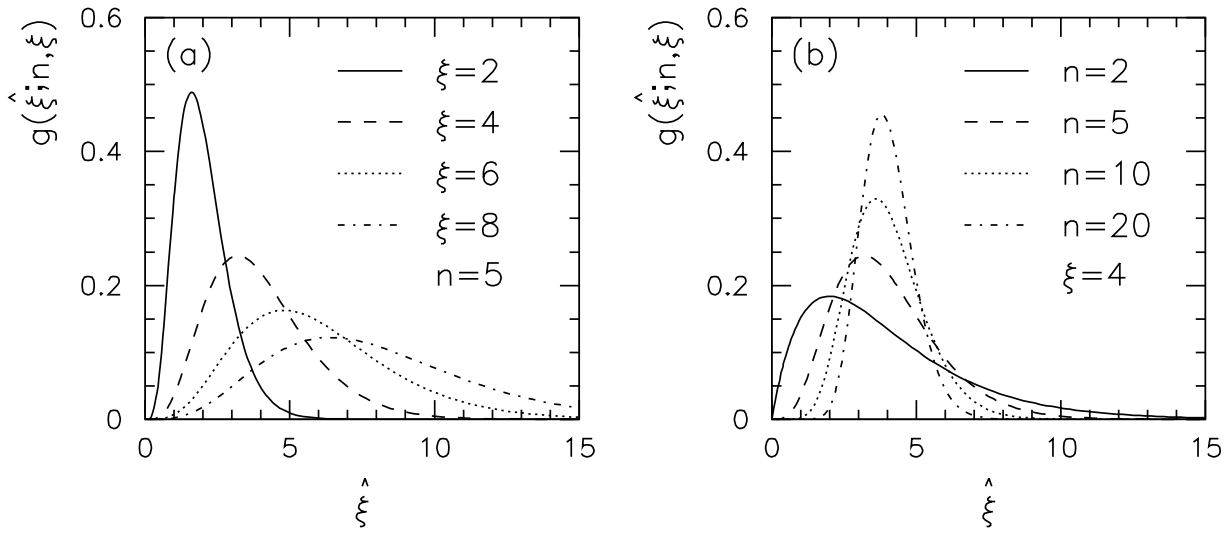

Figure 10.1: The sampling p.d.f. $g(\hat{\xi};n,\xi)$ for the estimator $\hat{\xi}$ for various values of n and ξ. (a) $n = 5$ measurements and various values of the true parameter ξ. (b) $\xi = 4$ and various numbers of measurements n.

Expectation Value for Mean Lifetime and Decay Constant

Using now the conventional notation for particle lifetimes, equation (10.13) gives the p.d.f. of $\hat{\tau} = (1/n)\sum_{i=1}^{n} t_i$ used to estimate the mean lifetime τ of a particle given n decay-time measurements $t_1, \ldots t_n$. Recall that the expectation value of $\hat{\tau}$ was computed in Section 6.2 by using the formula

$$E[\hat{\tau}(t_1,\ldots t_n)] = \int_0^\infty \cdots \int_0^\infty \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \frac{1}{\tau}e^{-t_1/\tau} \cdots \frac{1}{\tau}e^{-t_n/\tau} dt_1 \cdots dt_n = \tau \ . \qquad (10.14)$$

This result could have also been obtained directly from the p.d.f. of $\hat{\tau}$ (see equation (10.13)),

$$\begin{aligned} E[\hat{\tau}] &= \int_0^\infty \hat{\tau}\, g(\hat{\tau};n,\tau)\, d\hat{\tau} \qquad (10.15) \\ &= \int_0^\infty \hat{\tau}\, \frac{n^n}{(n-1)!}\, \frac{\hat{\tau}^{n-1}}{\tau^n}\, e^{-n\hat{\tau}/\tau}\, d\hat{\tau} \\ &= \tau \ . \end{aligned}$$

It was also shown in Section 6.2 that the maximum likelihood estimator for a function of a parameter is given by the same function of the ML estimator for the original parameter. For example, the ML estimator for the decay constant $\lambda = 1/\tau$ is $\hat{\lambda} = 1/\hat{\tau}$. From $g(\hat{\tau};n,\tau)$ one can compute the p.d.f. $h(\hat{\lambda})$,

$$
\begin{aligned}
h(\hat{\lambda}; n, \lambda) &= g(\hat{\tau}; n, \tau) \left| d\hat{\tau}/d\hat{\lambda} \right| \\
&= \frac{n^n}{(n-1)!} \frac{\lambda^n}{\hat{\lambda}^{n+1}} e^{-n\lambda/\hat{\lambda}} \, .
\end{aligned} \tag{10.16}
$$

The expectation value of $\hat{\lambda}$ is

$$
\begin{aligned}
E[\hat{\lambda}] &= \int_0^\infty \hat{\lambda} \, h(\hat{\lambda}; n, \lambda) \, d\hat{\lambda} \\
&= \int_0^\infty \frac{n^n}{(n-1)!} \frac{\lambda^n}{\hat{\lambda}^n} e^{-n\lambda/\hat{\lambda}} \, d\hat{\lambda} \\
&= \frac{n}{n-1} \lambda \, .
\end{aligned} \tag{10.17}
$$

One sees that even though the maximum likelihood estimator $\hat{\tau} = (1/n) \sum_{i=1}^n t_i$ is an unbiased estimator for τ, the estimator $\hat{\lambda} = 1/\hat{\tau}$ is not an unbiased estimator for $\lambda = 1/\tau$. The bias, however, goes to zero in the limit that n goes to infinity.

Confidence Intervals for Mean of Exponential Random Variable

The p.d.f. $g(\hat{\xi}; n, \xi)$ from equation (10.13) can be used to determine a confidence interval according to the procedure given in Section 9.2. Suppose n observations of the exponential random variable x have been used to evaluate the estimator $\hat{\xi}$ for the parameter ξ, and the value obtained is $\hat{\xi}_{exp}$. The goal is to determine an interval $[a, b]$ given the data $x_1, \ldots x_n$ such that the probabilities $P[a < \xi] = \alpha$ and $P[\xi < b] = \beta$ hold for fixed α and β regardless of the true value ξ.

The confidence interval is found by solving equations (9.9) for a and b,

$$
\begin{aligned}
\alpha &= \int_{\hat{\xi}_{exp}}^\infty g(\hat{\xi}; a) \, d\hat{\xi} \, , \\
\beta &= \int_{-\infty}^{\hat{\xi}_{exp}} g(\hat{\xi}; b) \, d\hat{\xi} \, .
\end{aligned} \tag{10.18}
$$

Figure 10.2 shows the 68.3% confidence intervals for various values of n assuming a measured value $\hat{\xi}_{exp} = 1$. Also shown are the intervals one would obtain from the measured value plus or minus the estimated standard deviation. As n becomes larger the p.d.f. $g(\hat{\xi}; n, \xi)$ becomes Gaussian (as it must by the Central Limit Theorem) and the 68.3% central confidence interval approaches $[\hat{\xi}_{exp} - \hat{\sigma}_{\hat{\xi}}, \hat{\xi}_{exp} + \hat{\sigma}_{\hat{\xi}}]$. An example similar to the one given here can be found in [Bra92] page 207, where the confidence intervals are estimated using the likelihood function.

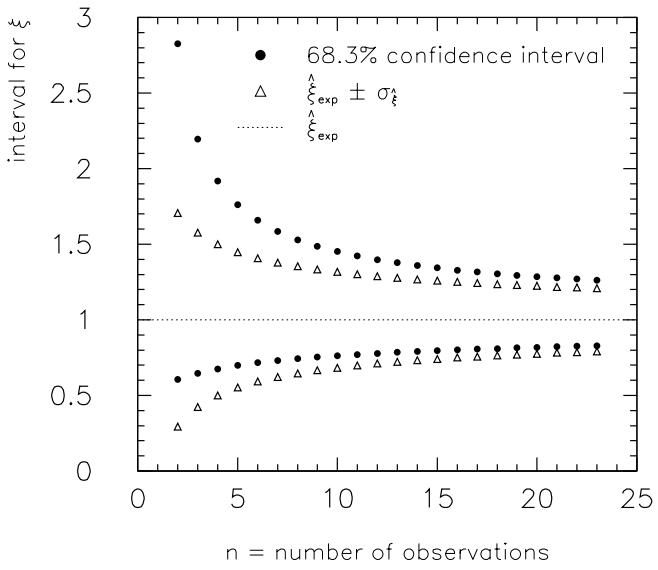

Figure 10.2: Classical confidence intervals for the parameter of the exponential distribution ξ (between solid points) and the interval $[\hat{\xi}_{exp} - \hat{\sigma}_{\hat{\xi}}, \hat{\xi}_{exp} + \hat{\sigma}_{\hat{\xi}}]$ (between open triangles) for different values of the number of measurements n, assuming an observed value $\hat{\xi}_{exp} = 1$.

Chapter 11

Applications and Examples

In preparation.

Bibliography

Among the following references, of special use for data analysis are the books by Barlow [Bar89], Brandt [Bra92], Eadie et al. [Ead71], Frodeson et al. [Fro79], and Lyons [Lyo86]. A collection of important statistical methods is included in the *Review of Particle Properties* by the Particle Data Group [PDG94], published every two years.

[All80] W. Allison and J. Cobb, *Relativistic Charged Particle Identification by Energy Loss*, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

[Arf70] G. Arfken, *Mathematical Methods for Physicists*, Academic Press, New York, 1970.

[Bar89] R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, John Wiley & Sons, Chichester, 1989.

[Bay63] T. Bayes, *An essay towards solving a problem in the doctrine of chances*, Philosophical Transactions of the Royal Society, **53** (1763) 370. Reprinted in Biometrika, **45** (1958) 293.

[Ber88] J.O. Berger and D.A. Berry, *Statistical Analysis and the Illusion of Objectivity*, American Scientist, **76**, No. 2, (1988) 159.

[Bra92] S. Brandt, *Datenanalyse*, 3. Auflage, BI-Wissenschaftsverlag, Mannheim, 1992.

[CER96] CERN Program Library, CERN, Geneva, 1996.

[Dud88] Edward J. Dudewicz and Satya N. Mishra, *Modern Mathematical Statistics*, John Wiley & Sons, New York, 1988.

[Ead71] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics*, North Holland, 1971.

[Fis90] R.A. Fisher, *Statistical Methods, Experimental Design and Scientific Inference*, a re-issue of *Statistical Methods for Research Workers*, *The Design of Experiments*, and *Statistical Methods and Scientific Inference*, Oxford University Press, Oxford, 1990.

[Fro79] A.G. Frodesen, O. Skjeggestad and H. Tøfte, *Probability and Statistics in Particle Physics*, Universitetsforlaget, Bergen-Oslo-Tromsø, 1979.

[Hel83] O. Helene, *Upper Limit of Peak Area*, Nucl. Instr. and Meth. **212** (1983) 319.

[Hig83] V.L. Highland, *Estimation of Upper Limits from Experimental Data*, Temple University Note COO-3539-38, 1983.

[Hud63] D.J. Hudson, *Lectures on Elementary Statistics and Probability*, CERN 63-29, 1963.

[Hud64] D.J. Hudson, *Statistics Lectures II: Maximum Likelihood and Least Squares Theory*, CERN 64-18, 1964.

[Jam89] F. James, M. Roos, CERN Program Library routine D506 (long write-up), 1989; F. James, *Interpretation of the Errors on Parameters as given by MINUIT*, supplement to long write-up of routine D506, 1978.

[Jam90] F. James, *A review of pseudorandom number generators*, Comp. Phys. Comm. **60** (1990) 329.

[Jam91] F. James and M. Roos, *Statistical notes on the problem of experimental observations near an unphysical region*, Phys. Rev. **D44** (1991) 299.

[Jef48] Harold Jeffreys, *Theory of Probability*, 2nd edition, Oxford University Press, London, 1948.

[Joh68] J. Johnston, G.B. Price and F.S. Van Vleck, *Sets, Functions and Probability*, Addison-Wesley, Reading, Massachusetts, 1968.

[Ken79] M.G.Kendall and A. Stuart, *The Advanced Theory of Statistics*, Hafner, New York, Volume I, 1977; Volume II, 1979; Volume III, 1968.

[Kol33] A.N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea Publishing Co., New York, 1956.

[Lan44] L. Landau, *On the Energy Loss of Fast Particles by Ionisation*, J. Phys. USSR, **8** (1944) 201.

[Lec88] P.L. L'Ecuyer, *Comm. ACM*, **31** (1988) 742.

[Lyo86] L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, Cambridge, 1986.

[Mac69] H.D. Maccabee and D.G. Papworth, *Correction to Landau's Energy Loss Formula*, Phys. Lett. **30A** (1969) 241.

[Mui82] R.J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.

[Oha94] A. O'hagan, *Kendall's Advanced Theory of Statistics*, Vol. 2B, *Bayesian Inference*, Edward Arnold, London, 1994.

[PDG94] The Particle Data Group, *Review of Particle Properties*, Phys. Rev. **D50** (1994) 1271.

[Per87] Donald H. Perkins, *Introduction to High Energy Physics*, Addison-Wesley, Menlo Park, California, 1987.

[Qui83] Chris Quigg, *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions*, Benjamin/Cummings, Reading, Massachusetts, 1983.

[Shi92] J. Shiers and M. Goossens, *HBOOK Reference Manual version 4.15*, CERN Program Library Long Write-up Y250 (1992).

[Spr79] M.D. Springer, *The Algebra of Random Variables*, John Wiley & Sons, New York, 1979.

[Yos85] G.P. Yost, *Lectures on Probability and Statistics*, Lawrence Berkeley Laboratory report LBL-16993, 1985.

Index