PHYSICAL DATA ANALYSIS

A PRIMER

Caltech

Frank Rice

Contents

USEFUL TABLES AND FORMULAS

(See the referred chapters of the text for details.)

Uncertainty propagation

In the following table, "x" and "z" are used where one should really refer to estimates of μ_x and μ_z. The derivative $f'(x)$ is evaluated at the estimated value of μ_x. Symbols "a" and "b" refer to real-valued parameters which do not have an associated uncertainty. More complete information and derivations are found in the section *Propagation of uncertainties* in Chapter 2.

Table I: Naïve uncertainty propagation

Common functions of a single uncertain variable $x \pm \sigma_x$

$z = a + bx$	$\sigma_z = \lvert b \rvert \sigma_x$
$z = ax^2$	$\frac{\sigma_z}{\lvert z \rvert} = 2\frac{\sigma_x}{\lvert x \rvert}$
$z = \sqrt{ax}$	$\frac{\sigma_z}{\lvert z \rvert} = \frac{1}{2}\frac{\sigma_x}{\lvert x \rvert}$
$z = ax^b$	$\frac{\sigma_z}{\lvert z \rvert} = \lvert b \rvert \frac{\sigma_x}{\lvert x \rvert}$
$z = ae^{bx}$	$\frac{\sigma_z}{\lvert z \rvert} = \lvert b \rvert \sigma_x$
$z = a \ln \lvert bx \rvert$	$\sigma_z = \lvert a \rvert \frac{\sigma_x}{\lvert x \rvert}$
$z = f(x)$	$\sigma_z = \lvert f'(x) \rvert \sigma_x$

Uncertainty propagation for a function of several variables

$$z \approx z(x_1, x_2, \ldots, x_n)$$

$$\sigma_z \approx \sqrt{\sum_{i=1}^{n} \left(\frac{\partial z}{\partial x_i}\right)^2 \sigma_{x_i}^2 \quad + \quad 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \frac{\partial z}{\partial x_i}\frac{\partial z}{\partial x_j}\sigma_{x_i x_j}^2}$$

If the arguments are all independent random variables, then the sum over the covariance terms vanishes.

Common functions of two independent, uncertain variables $x \pm \sigma_x$ and $y \pm \sigma_y$

$z = ax + by$	$\sigma_z = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}$
$z = axy$	$\frac{\sigma_z}{\lvert z \rvert} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$

Point estimates

See *Point estimation* in Chapter 2 and *Weighted mean of several measurements* in Chapter 3.

Point estimate from N samples

$$Y \pm \sigma_Y = \bar{y} \pm \sqrt{s^2/N}$$

$$\bar{y} = \frac{1}{N}\sum_i y_i \; ; \quad s^2 = \frac{1}{N-1}\sum_i (y_i - \bar{y})^2$$

Accuracy of the uncertainty

$$\frac{\sigma_{\sigma_Y}}{\sigma_Y} \sim \frac{1}{\sqrt{2(N-1)}}$$

Weighted mean of samples with different uncertainties

$$\bar{x} = \left(\sum_{i=1}^{N} \frac{x_i}{s_i^2}\right) \div \left(\sum_{i=1}^{N} \frac{1}{s_i^2}\right) \; ; \qquad \frac{1}{\sigma_{\bar{x}}} = \sqrt{\sum_{i=1}^{N} \frac{1}{s_i^2}}$$

Normal distribution

See the section *Noise and the normal distribution* in Chapter 1.

The normal (Gaussian) distribution

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

Table II

Properties of the normal distribution

Total probability within ±1σ of the mean:	68.3%
Total probability within ±2σ of the mean:	95.5%
Total probability beyond 3σ from the mean:	0.27%
Range around mean for 50% probability:	±0.67σ
½ maximum PDF locations:	±1.18σ
Probability between ½ maximum points:	76.1%

Table III

Even moments about the mean of the normal distribution
(odd moments all vanish by symmetry)

m	2	4	6	$2n$
$\overline{(x-\mu)^m}$	σ^2	$3\sigma^4$	$15\sigma^6$	$\frac{(2n)!}{2^n\, n!}\,\sigma^{2n}$

χ^2 calculations

See Chapter 1, *The χ2 distribution*, and Chapter 4.

N data points y_i, 1-parameter model of a point value: $y = Y(a)$

$$\chi^2_{(N-1)} = \sum_{i=1}^{N} \frac{\left(y_i - Y(a)\right)^2}{\sigma_i^2}$$

degrees of freedom $\nu = N - 1$

N data points (x_i, y_i), *M*-parameter model of a function: $y = f(x; a_1, a_2, \ldots, a_m)$

$$\chi^2_{(N-M)} = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_M)\right)^2}{\sigma_i^2}$$

degrees of freedom $\nu = N - M$

Data point variances for χ^2

$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2 - 2\left(\frac{df}{dx}\right)\sigma_{x_i y_i}^2$$

Reduced χ^2 with ($N - M$) degrees of freedom

$$\tilde{\chi}^2_{(N-M)} \equiv \frac{\chi^2_{(N-M)}}{N-M}$$

$$\sigma_{\tilde{\chi}^2_{(N-M)}} = \sqrt{\frac{2}{N-M}}$$

Table IV

Properties of the reduced χ^2 distribution

Total probability for $\tilde{\chi}^2 > 1 + 1\sigma_{\tilde{\chi}^2}$:	< 16%
Total probability for $\tilde{\chi}^2 > 1 + 2\sigma_{\tilde{\chi}^2}$:	< 5%
Total probability for $\tilde{\chi}^2 < 1 - 2\sigma_{\tilde{\chi}^2}$:	< 2.25%

Covariances of the fit parameter estimates from χ^2 minimization

$$\boldsymbol{\Sigma} = 2\left(\mathbf{H}^{-1}\right)$$

Covariance matrix $\boldsymbol{\Sigma}$: $\Sigma_{jk} = \sigma^2_{a_j a_k}$

Hessian matrix $\mathbf{H}$: $H_{jk} = \left.\frac{\partial^2(\chi^2)}{\partial a_j \, \partial a_k}\right|_{\min(\chi^2)}$

Pearson's χ^2 test of a distribution

See *Testing for normally-distributed data scatter* in Chapter 3.

$$\chi^2_{(Q-3)} = \frac{1}{(Q-1)N}\sum_{i=1}^{Q}(QN_i - N)^2$$ **Quantile χ^2 test**

Q: number of quantiles, N: number of data points, N_i: the number falling in the ith quantile.

The lower boundary of the first quantile is $-\infty$; the Qth quantile has an upper boundary of $+\infty$. For a Gaussian with mean 0 and variance 1 the upper boundary of the ith quantile is:

$$y_i = -\sqrt{2}\,\mathrm{erfc}^{-1}(2i/Q)$$ **i^{th} quantile upper boundary**

$\mathrm{erfc}^{-1}()$ is the inverse of the *complementary error function*:

$$\mathrm{erfc}(y) = (2/\sqrt{\pi})\int_y^\infty \exp(-t^2)\,dt$$

Table V

Quantile boundaries (only values above the mean shown)

Q	y_i					$limit\ \chi^2$
4	0.6745					3.84
6	0.4307	0.9674				7.81
8	0.3186	0.6745	1.1503			11.1
10	0.2533	0.5244	0.8416	1.2816		14.1
12	0.2104	0.4307	0.6745	0.9674	1.3830	16.9

The boundaries are symmetric about the mean and include the mean (for even Q). For a general Gaussian, multiply the boundary values by the standard deviation and add the mean. The limit χ^2 is for a p-value of 5%.

PREFACE

This text is intended to be a short introduction and ready reference to basic techniques and issues of data analysis important for the beginning physical scientist to understand. Its presentation has developed from the author's more than twenty years teaching this subject to sophomore-level physics majors at Caltech, most of whom intended to go on to doctoral study of experimental physics or applied physics. The physics laboratory course sequence at Caltech has evolved through the years to become an important hurdle for our undergraduates, introducing them to the rigors and rewards of experimental research as a professional activity. As modern research equipment has grown in precision and speed, so has grown the need for physics students to have some familiarity with an ever more sophisticated repertoire of data analysis techniques to call on. This text is intended to provide key elements of the foundation upon which understanding of these techniques may be built.

Theory and experiment

One of the most prevalent characteristics of human thought is our almost irresistible tendency to create generalizations and abstractions about the world and our experiences of it.[*] Even though any particular physical object or specific sequence of actual events is unique, we continually and often unconsciously sift through our experiences and create mental models of those seemingly ideal, underlying characteristics various objects or sequences of actions appear to share. These apparent regularities in our day-to-day experiences form the basis of what we conclude are the underlying rules, or "laws," governing the natures and behaviors of the myriad things making up the physical world, and each of us consciously or unconsciously calls on these abstract mental models to decide how to act and how to react as we live our daily lives. As our experiences accumulate, we (albeit often reluctantly) refine our many mental models to improve our ability to anticipate future events and to better plan our own actions.

Through our gift of language we can share our findings with others. Joint efforts can then further refine and expand as well as record representations of these abstract mental models. Perhaps the ultimate expression of humanity's creativity in this regard (besides language itself) is the development of the abstract studies of *quantity* and *mathematics*. The concept of the number *two* abstracted from observations of pairs of objects or events or the concept of *addition* abstracted from aggregations of originally separate collections of things provide fundamental examples. Our ability to create mental models extends further, however, in a way possibly unique to humans: we can ponder these abstract concepts and invent deeper abstractions which are even more removed from direct experience. Confining ourselves to the realm of mathematics, the ideas of *real and complex numbers*, of a *function*, of spaces with more than three dimensions, and of *infinity* provide a few examples.

[*] This statement itself serves as a typical example.

Theories of physics are expressed mathematically and thus relate two or more quantities abstracted from observed behaviors of things. As with all abstract mental models, physical theories are idealizations of what we observe to happen during any particular actual sequence of events in the world. The abstract entities representing the primitive atoms of a theory are, naturally, mental constructs of the theorist. How these abstractions and their mathematical relationships correspond to actually experienced sequences of events may be more or less ambiguous or idealized. Consequently, theories provide approximations of what behavior may be observed in a specific experiment or situation encountered in the real world (although they can be very precise).

Thus, physical theories express abstract mathematical relationships among numerical quantities with physical meaning (the quantities correspond to observable properties and behaviors of things in the world, such as lengths, weights, rates, etc.). Included in most fundamental physical theories are mathematical expressions which contain *free parameters*: universal numerical quantities with no a priori, purely theoretical method of determining their values. Examples of such *fundamental constants of nature* include the electron's charge and rest mass or the values of Planck's constant and the gravitational constant. The only way to determine these quantities is to conduct experiments and to observe what happens in the real world, and then choose values for them which best model the behaviors of things observed during actual specific events.[*]

These facts demand that as experimenters *we must make measurements* when attempting to evaluate a physical theory. But imperfect calibration and the finite numerical precision of the instruments we use, along with the lack of exact repeatability of our most precise observations, indicate that each of our observations always has, naturally, some error — our measurements are wrong. On the other hand, because most theories involve idealizations and simplifications of actual objects and specific situations ("Consider a spherical cow in a vacuum…"), their descriptions and predictions are only approximate — the theory is also wrong. So how can we make any progress? How can we have confidence in the accuracy of anything we conclude about the world? The following pages introduce elementary methods to begin to answer these questions.

Acknowledgements

The author was introduced to the Caltech approach to early undergraduate physics laboratory instruction in the mid-1990's by Don Skelton, the long-time instructor and manager of the freshman and sophomore physics laboratory courses. As a graduate student TA under Don's patient tutelage, I learned how to select from the many available techniques and formulas found in statistics textbooks and apply them to the problems of actual physics data analysis encountered by my students. Don was an enthusiastic adopter of the use of

[*] Assuming, of course, that a particular theory does provide a useful abstract mental model of the world, and thus its free parameters really do have useful meanings.

personal computers for student analysis of experimental data, and his program *FFIT*, written entirely in Basic, was a surprisingly powerful tool for data analysis and curve fitting on computers containing only a fraction of a megabyte of ram and with clock speeds of only a few megahertz. Don's text *A Primer on Data Analysis* was read by hundreds of Caltech undergraduates and was the inspiration for this effort.

Don's *FFIT* software was converted to *Mathematica*® code in 1997 by Anastasios (Tasos) Vayonakis, who renamed the program *CurveFit* and spent many weeks optimizing the program's algorithms. Neither the computers nor the *Mathematica* numerical routines of the time were very efficient, but Tasos's efforts were generally quite successful. A Caltech graduate student, Tasos was a meticulous researcher who conducted several exacting experiments, and I learned a tremendous amount about experimental technique and analysis from him. Since 2001 the author has been the sole maintainer of *CurveFit* code, and, along with the growth in computer power and *Mathematica*'s sophistication, the current version has expanded so much that it would appear nearly unrecognizable to those early pioneers. Even so, it currently still retains core data structures and algorithms inspired by the code Don and Tasos developed decades ago.

Finally, I must gratefully acknowledge the contributions of my friend and colleague David Politzer, who has partnered with me to teach undergraduate experimental physics these last few years. A brilliant, dedicated teacher as well as physicist, David has provided his invaluable perspective as a theorist to me and our students. Hopefully together we have shown them that theory and experiment form an inseparable collaboration in the quest for knowledge about our world.

Frank Rice
Pasadena, CA

Chapter 1
Random variables and their statistical description

RANDOM VARIABLES, PROBABILITY DENSITY, EXPECTED VALUE

Noise and the problem of repeatability

Consider a common situation encountered when attempting to accurately measure the value of an observed quantity such as a voltage or a length: attempts to repeat the measurement, if they are precise enough, will result in a set of various different values for it: the measurement exhibits seemingly random variations from trial (observation) to trial. For example, the observations may consist of the readings of a voltmeter repeated at one-second intervals or may be the readings recorded during repeated attempts to measure the length of a resonant cavity using a precision caliper. In the first case, the voltage reading variations may be caused by many tiny, independent (not causally related) sources of electrical fluctuations within the voltmeter itself, the experimental apparatus, and the connecting wiring. In the second case, the length measurements may vary because the experimenter places the calipers in slightly different locations each time, or looks at the scale from slightly varying angles, or uses slightly varying pressures when fitting the caliper jaws to the cavity. We lump the resulting *scatter* in our measured values into the general category of *noise* in our measurement process. Figure 1-1 illustrates the problem.

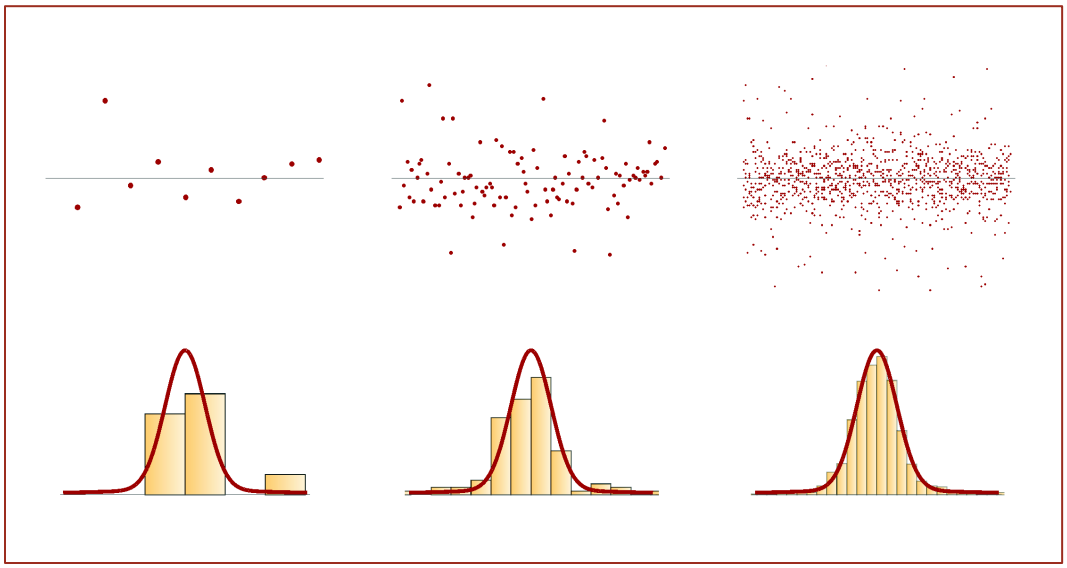


Figure 1-1: an illustration of scatter in the repeated measurement of an experimental quantity. As the number of measurements increases, the relative frequencies of observed values, indicated by the histogram bars, evidently converge to a smooth curve, centered on what could be interpreted as the underlying, "true" value to be measured (indicated by the horizontal line in each scatter plot).

In cases such as these we may believe that there is really a single, "true" voltage or length that could be measured if only our instrumentation and experimental techniques were absolutely precise, accurate, and free of noise. Clearly, because of the scatter in the measured data, each individual measurement is then in error to some degree. We further may have reason to believe that, because we are careful and unbiased, the errors in the individual measurements are completely unrelated to each other in direction and exact magnitude. We may then expect that the various measurements generally fall within some more or less well-defined range centered near the underlying, "true" value, as illustrated in Figure 1-1. What we want is some reasonable way to estimate the underlying value and to estimate how uncertain our estimate of that value may be. This chapter provides an overview of a basic mathematical framework which will allow us to develop such a method.

The next section describes the properties of an abstract mathematical model of a generator of a set of noisy measurements: a *random variable*. Some of these properties may seem somewhat abstruse, but they actually provide the axiomatic foundation for the mathematics we will use throughout this text. We may not spend a lot of words explaining why we need these properties, and few theorems will be explicitly derived from them in this text, but their justifications will be found in any thorough statistics reference.

Random variables and processes, statistical ensembles, probability density

Assume that we make N measurements of an experimental quantity x, and noise in the measurement process introduces random errors in the measurements. We obtain the set of various measurement values $\{x_i\}$, with the integer $i \in \{1 \ldots N\}$. To proceed with the mathematical analysis of our data, we now assume that we can reasonably model these values as *samples* of a *continuous random variable* x. A "random variable" is not really a variable at all. It is a generalized type of function (also called a *distribution*). For our present purpose, think of the random variable x as a function taking a single integer i as its argument and returning the ith sample value $x_i = x(i)$.* Conceptually, the infinite sequence of values $\{x_i\}$ for $i \in \{1 \ldots \infty\}$ is called a *random process* associated with x.† Sometimes the random variable might be considered to be a function of a continuous argument such as time t. In this case, $x_i = x(t_i)$, where t_i is the time at which the ith sample is acquired. In this case $x(t)$ for all t becomes a random process associated with the random variable x.

This physically-nonexistent, abstract, infinite sequence of measurements $\{x_i\}$ for $i \in \{1 \ldots \infty\}$ is meant to characterize the random variable x in the following sense: we assume that if we were to continue taking measurements forever, holding all experimental

* x is a *continuous* random variable because these return values are real numbers. A *discrete* random variable would return values from some countable set such as the set of nonnegative integers.

† Or we could consider our data set to have come from some "eternal" random process which extends infinitely into the past as well, in which case the index i could be any integer, $i \in \{-\infty \ldots \infty\}$. In either case, such infinite sequences are abstract concepts, because in reality we have, of course, only our finite set of N measurements.

conditions constant, and that we and our equipment don't age, and that we are just as careful with every measurement, etc., etc., etc., then we could, in principle, generate all of the $\{x_i\}$. *Our first big assumption* (or axiom) is that making ever finer histograms of the relative frequencies of the values of the elements of this sequence would in this infinite limit yield a smooth curve as shown in Figure 1-1.

Our experience tells us that the noise in our measurements makes each result unpredictable (varying randomly), so we may expect that this particular infinite sequence of values $\{x_i\}$ is not the only one which could have been generated by the random variable x. A vast multitude of other infinite sequences $\{x'_i\}$, $\{x''_i\}$, etc., might have been generated, corresponding to different sets of experimental results consistent with our actual measurements. Such an assemblage of all of the possible sequences (processes) which could have been generated by a random variable is called a *statistical ensemble* of sequences. *Our second big assumption* is that the histogram of relative frequencies of values generated from each of these different possible sequences converges to the same result, and therefore *the limiting histogram of any one particular sequence of measurements is a defining characteristic of the random variable*.

Now comes our final, most subtle assumption about the nature of the random variable x characterizing the noise in our measurements. Consider the set obtained by selecting the mth element of each of the various possible sequences: $\{x_m, x'_m, x''_m, \ldots\}$ (for example, $m = 1$, corresponding to the first measurement value in each sequence).[*] *Our final and biggest assumption* is that the histogram of relative frequencies of values in this ensemble of the possible mth measurements converges to the same histogram as do the set of values in any particular sequence of multiple measurements. Thus this single, limiting histogram is truly a fundamental characteristic of the random number x.[†]

The characteristic limiting histogram of the continuous random variable x (Figure 1-1) takes the form of a curve $p_x(x)$, called the *probability density function* (or PDF) of the random variable x. By *probability density*, we mean that $dP = p_x(\xi)\, d\xi$ is the infinitesimal fraction of the members of either a process $\{x_i\}$ or an ensemble of measurements $\{x_m, x'_m, x''_m, \ldots\}$ whose values fall within the interval $d\xi$ around some particular value ξ. The total fraction of an infinite set's members whose values fall within some interval $a < \xi < b$ would be obtained by integrating $p_x(\xi)\, d\xi$ over that interval; the PDF is normalized so that $\int_{-\infty}^{\infty} p_x(\xi)\, d\xi = 1$.

We define the *probability* $P(a < x_j < b)$ of obtaining a value for the jth sample of the random variable x in the range $a < x_j < b$ in terms of the PDF as:

1.1 $$P(a < x_j < b) = \int_a^b p_x(\xi)\, d\xi$$ **Probability and the PDF**

[*] Another example of a *statistical ensemble*.

[†] This final assumption is a form of the *ergodic hypothesis*. It is a critically important foundation for the mathematical analysis of many random processes and systems.

with PDF normalization: $\int_{-\infty}^{\infty} p_x(\xi)\,d\xi = 1.$

We therefore use what may be called a "fractional" definition of probability: $P(a < x_j < b)$ is the *fraction of the total number of members* of the x_j ensemble that have values between a and b. Although the presence of noise makes the actual value obtained in a particular experiment unpredictable, we nevertheless expect that should we carefully repeat the experiment many times, the observed fraction of the results for which $a < x_j < b$ will approach $P(a < x_j < b)$ given in (1.1). This conclusion may also be considered to be a "relative frequency" definition of probability, since it also describes how frequently (on average) measurements are expected to fall in the specified range in the limit that the total number of samples becomes large. Finally, the probability (1.1) might be interpreted as describing the "odds" or "degree of rational belief" that a single experimental measurement of x_j would yield $a < x_j < b$, although one ought to consider whether such interpretations are really in any way substantially different from the fractional definition.

Ensure that you thoroughly understand the motivation behind and possible limitations of this mathematical model we will use for the underlying "generator" of an experimental data set. The necessary inference is illustrated in Figure 1-1 on page 1: as shown by the histograms, the relative frequencies of the various data values seem to converge to a smooth curve as the number of measurements becomes large. The curve is then identified with the PDF of some random variable. *This inference is not logically justifiable, because we will never acquire the infinite number of samples needed to prove that such a limiting PDF exists.*

Our experience, however, with the results of careful observations leads us to believe that this mental picture is nevertheless appropriate for modeling many experimental situations. As with any abstraction we may employ to characterize the physical world's behavior, it must be remembered that our model is only an approximation based on an inference that in some cases may prove to be inadequate.

Functions of a random variable; expected values; mean and variance

With the above motivation and caveats always in the backs of our minds, let us then press on with our mathematical model of a noisy measurement process as that of taking samples of a random variable x with a well-defined PDF p_x. A function $f(x)$ will take on random values when given a sequence of samples of the random variable x as arguments. The *expected value* of f is *defined* in terms of p_x as:

1.2 $$\mathrm{E}[f(x)] \equiv \int_{-\infty}^{\infty} f(x)\,p_x(x)\,dx$$ **Expected Value**

Note that the x in the integrand is a variable of integration and not the random variable x. We'll often make double use of a symbol in this manner, *so stay alert*. E[$f(x)$] is a *weighted*

mean of $f(x)$ over the possible values of the random variable x. Other names for the expected value are *average value* and *mean value*, and $\mathrm{E}[f(x)]$ may also be denoted by symbols such as $\langle f(x) \rangle$, $\overline{f(x)}$, or μ_f. Note that the expected value may not exist for some combinations of functions f and p_x, because the resulting integral (1.2) may not converge. Expected values are just numbers, not random distributions, so $\mathrm{E}[\mathrm{E}[f(x)]] = \mathrm{E}[f(x)]$.* The integral in (1.2) is linear in its integrand, so for numerical constants a and b:

1.3
$$\mathrm{E}[a f(x) + b g(x)] \;=\; a\,\mathrm{E}[f(x)] \;+\; b\,\mathrm{E}[g(x)]$$

Given a random variable x with distribution described by the PDF p_x, the expected value of x itself, $\mathrm{E}[x]$, is called the *statistical* or *distribution mean of* x and is symbolized by μ_x:

1.4
$$\mu_x \;\equiv\; \int_{-\infty}^{\infty} x\, p_x(x)\, dx$$ **Distribution Mean**

This quantity is also called the *first moment* of the distribution described by p_x. Higher moments are given by the expectation values $\mathrm{E}[x^2]$, $\mathrm{E}[x^3]$, etc. *Moments about the mean* are also useful: $\mathrm{E}[x - \mu_x]$, $\mathrm{E}[(x - \mu_x)^2]$, etc. The first moment about the mean identically vanishes by (1.2) and (1.3): $\mathrm{E}[x - \mu_x] \equiv 0$.

The second moment about the mean of a random variable x is called the distribution *variance* and is quite important. It is conventionally given the symbol σ_x^2:

1.5
$$\sigma_x^2 \;\equiv\; \mathrm{E}[(x-\mu_x)^2] \;=\; \int_{-\infty}^{\infty} (x-\mu_x)^2\, p_x(x)\, dx \;=\; \mathrm{E}[x^2] - \mu_x^2$$ **Variance**

The final expression in (1.5) comes from multiplying out $(x - \mu_x)^2$ and using the linearity relation (1.3). The positive square root of the variance is called the *standard deviation*, σ_x. The standard deviation, from (1.5), is the "root of the mean squared deviation," or *RMS deviation* of a distribution about its mean. It is a common indicator of the size of the scatter in a measured quantity, as illustrated in Figure 1-2.

PDF of a derived random variable

We have considered how to determine the expected value of a function $f(x)$ of some random variable x whose PDF p_x is known, and we have used this method to define the *mean* and *variance* of x. Given a function $f(x)$, we now wish to determine not only its expected value, but also the distribution (PDF) of its returned values. In other words, we define a new random variable y derived from x using the function $y = f(x)$, and ask how its PDF $p_y(y)$ may be determined from $f(x)$ and $p_x(x)$.

By the definition of the PDF, the infinitesimal probability that the original random variable returns a value within dx of some chosen x is $dP = p_x(x)\, dx$. If $f(x)$ has a unique inverse, so that $y = f(x) \rightarrow x = g(y)$, then dP may also be found in terms of $y = f(x)$ using the chain rule:

* $\mathrm{E}[f(x)]$ is an *idempotent* function, as are, for example, the absolute value of a real number or a projection operator acting on a quantum mechanical state vector.

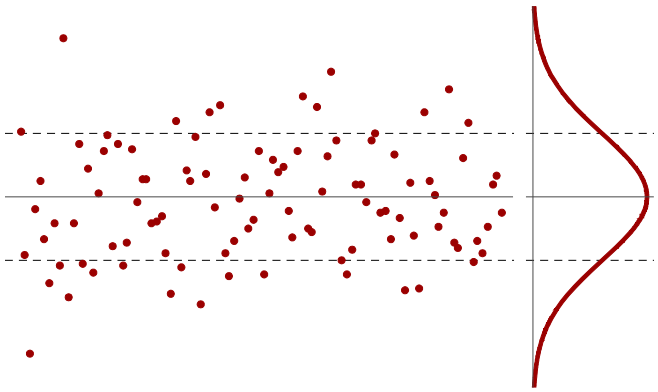


Figure 1-2: A typical sequence of independent samples (time horizontally, value vertically) of a random variable whose probability density is plotted at right. The mean of the distribution is shown by the solid horizontal line; the dashed lines are at one distribution standard deviation above and below the mean. The particular probability density shown is that of the *normal distribution*.

$$dP = p_x(x)\,dx = p_x(g(y))\left|g'(y)\right|dy = p_y(y)\,dy$$

The absolute value of the derivative $dx/dy = g'(y)$ is used so that we can always have $dy > 0$ when $dx > 0$, and therefore $p_y(y) > 0$. Thus we have derived an expression for $p_y(y)$ in terms of $p_x(x)$, using $g'(y) = 1/f'(x)$:

$$p_y(y) = p_x(g(y))\left|g'(y)\right| = p_x(x(y))/\left|f'(x)\right|$$

$p_y(y)$ is singular (blows up) wherever $f'(x)$ vanishes. This is fine as long as the resulting probability $\int p_y(y)\,dy$ remains finite when integrating over the singularity. Generally, these singularities imply another difficulty: $y = f(x)$ may not have a unique inverse, because several values of x may result in the same value for y (a common example is $y = x^2$). Thus there may exist several *branches* of the inverse function: $x_1 = g_1(y)$, $x_2 = g_2(y)$, etc. In this case the derived PDF must be calculated from $p_x(x)$ using a sum over the various branches of the inverse function $x(y)$:

1.6 $$p_y(y) = \sum_i p_x(x_i(y))/\left|f'(x_i)\right|$$ **Derived PDF of *y* = *f*(*x*)**

STATISTICS OF SEVERAL RANDOM VARIABLES

Joint distributions of more than one random variable, statistical independence

Given two random variables x and y, say, we may ask for the probabilities of obtaining various combinations of values when they are jointly sampled. We thus obtain the idea of a *joint probability distribution*, and in particular, their *joint probability density* p_{xy}. The differential probability of obtaining a value within dx of a particular x value along with obtaining a value within dy of a particular y value is then $p_{xy}(x, y)\,dx\,dy$. As an extension of (1.1), we get equation (1.7) on page 7.

1.7 $$P(a_x < x < b_x \text{ and } a_y < y < b_y) = \int_{a_y}^{b_y}\int_{a_x}^{b_x} p_{xy}(x,y)\,dx\,dy$$ **Joint probability density**

The individual PDFs of the two random variables, p_x and p_y, may be obtained from p_{xy} by integrating over the other variable:

$$p_x(x) = \int_{-\infty}^{\infty} p_{xy}(x,y)\,dy; \quad p_y(y) = \int_{-\infty}^{\infty} p_{xy}(x,y)\,dx$$

The individual PDFs obtained using these expressions are properly normalized if p_{xy} is.

Two random variables are *statistically independent* if and only if their joint PDF factors into their individual PDFs:

1.8 $$p_{xy}(x,y) \equiv p_x(x)p_y(y)$$ **Statistical independence**

In this case the joint PDF integral (1.7) factors into two independent integrals, one over each of the two variables' PDFs. Extending these results to more than two variables is straightforward.

Statistical independence of random variables reflects the mathematical modelling of causally independent physical sources of variation in a measurement process.

Expected values of functions of multiple random variables

Extending the definition (1.2) of the expected value or mean value to a function of multiple random numbers is straightforward:

1.9 $$\mathrm{E}[f(x,y)] \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,p_{xy}(x,y)\,dx\,dy$$

The joint distribution means and variances of x and y are thus:

$$\mu_x \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,p_{xy}(x,y)\,dx\,dy; \quad \mu_y \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\,p_{xy}(x,y)\,dx\,dy$$

$$\sigma_x^2 \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)^2\,p_{xy}(x,y)\,dx\,dy; \quad \sigma_y^2 \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-\mu_y)^2\,p_{xy}(x,y)\,dx\,dy$$

If the random variables x and y are statistically independent, then each of these expressions simplifies to that for a single random variable, (1.4) and (1.5).

Another second moment exists for the joint distribution, the *covariance* of random variables x and y:

1.10 $$\sigma_{xy}^2 \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)\,p_{xy}(x,y)\,dx\,dy = \mathrm{E}[xy]-\mu_x\mu_y$$ **Covariance**

The covariance may be either positive or negative, but will vanish identically if the two random variables are independent. The covariance is positive if joint samples of the two variables tend to be above their respective means or below their means at the same time and is negative if the joint samples tend to vary oppositely away from their means. Examples are shown in Figure 1-3. Scaling the covariance by the product of the individual variables' standard deviations results in the *correlation coefficient* $\rho_{xy} = \sigma_{xy}^2/(\sigma_x\sigma_y)$ (more properly called Pearson's correlation coefficient*). Since $-1 \le \rho_{xy} \le 1$, this quantity is a measure of the relative strength of the correlation between the variations of x and y away from their means.

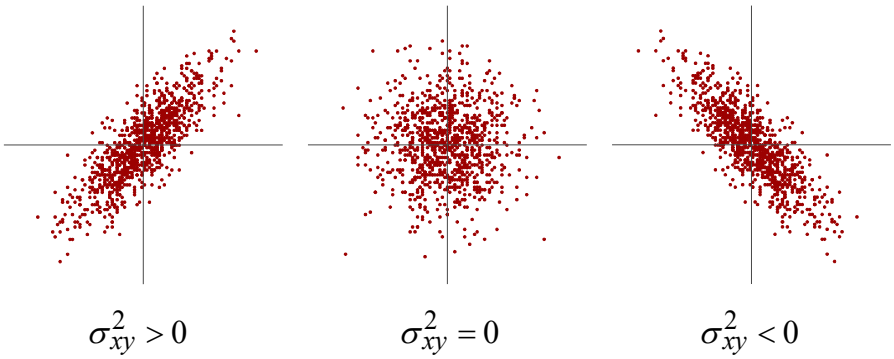
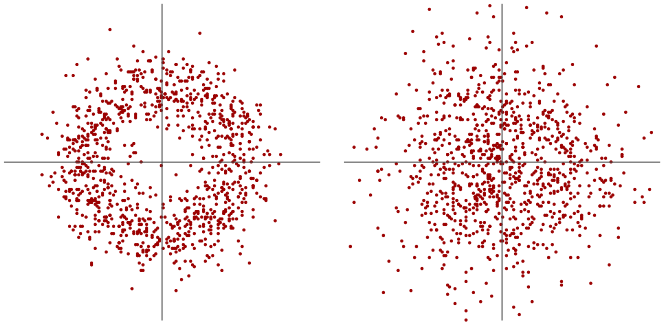


Figure 1-3: Samples of joint distributions showing the effects of *covariance* on their behaviors.

The vanishing of their covariance does not guarantee that two random variables are independent, as illustrated in Figure 1-4.

Figure 1-4: Samples of two different joint distributions of random variables *x* and *y*, both of which have vanishing covariance. The variables are statistically independent in the right distribution, but are clearly not in the joint distribution depicted at left.



* English mathematician Karl Pearson (1857–1936).

The sum and arithmetic mean of a set of random variables

An important special case of a function of multiple random variables is their sum. We wish to determine its distribution mean and variance. Let

$$f(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} x_i$$

where the x_i are elements of a finite set of N random variables.* Using the linearity of the expected value (1.3), the mean (expected value) of the random distribution of the sum f will simply be the sum of the individual random variable means, and the variance of f may be found using (1.3) and (1.10):

1.11 **Statistical mean of a sum**

$$\mu_f = \mathrm{E}[f(x_1, x_2, \ldots, x_N)] = \sum_{i=1}^{N} \mu_{x_i}$$

$$\sigma_f^2 = \mathrm{E}[f^2] - (\mu_f)^2 = \mathrm{E}\left[\sum_{i,j=1}^{N} x_i x_j\right] - \sum_{i,j=1}^{N} \mu_{x_i}\mu_{x_j} = \sum_{i,j=1}^{N} \left(\mathrm{E}[x_i x_j] - \mu_{x_i}\mu_{x_j}\right)$$

$$= \sum_{i,j=1}^{N} \sigma_{x_i x_j}^2$$

So the sum's variance is the sum of the terms' covariances. Whenever $i = j$ in a covariance, that term is just the variance $\sigma_{x_i}^2$; there will be exactly N such terms. Each covariance with $i \neq j$ is counted twice, since $\sigma_{x_i y_i}^2 = \sigma_{y_i x_i}^2$. Thus we can rewrite the above sum as:

1.12 **Variance of a sum**

$$\sigma_f^2 = \sum_{i=1}^{N} \sigma_{x_i}^2 + 2\sum_{i=2}^{N}\sum_{j=1}^{i-1} \sigma_{x_i x_j}^2$$

If the individual random variables x_i are all independent of each other (or just uncorrelated, which is a weaker criterion), then the sum over the covariance terms in (1.12) vanishes identically, and the standard deviation of the sum is given by a *Pythagorean sum* (square root of the sums of the squares), or *quadrature*, of the component random variables. This will turn out to be a most important result (1.13).

Standard deviation of a sum of independent random variables

1.13

$$\boxed{\sigma_{\Sigma x_i} = \sqrt{\Sigma \sigma_{x_i}^2}}$$

* Note that in this section the various x_i represent different random variables, not samples of a single random variable x.

The arithmetic mean (average) of the N random variables is their sum divided by N. Therefore, using (1.11), the arithmetic mean of a group of random variables will have a distribution with an expected value (statistical mean) which is the arithmetic mean of the members' expected values. Dividing the sum (1.11) by N will clearly reduce the sum distribution's width by the same factor; this implies that the arithmetic mean's standard deviation will also be N times smaller than that of the sum, (1.13).

> If N independent random variables all have the same standard deviation σ, then their sum's standard deviation will be $\sqrt{N}\,\sigma$, and the standard deviation of their arithmetic mean (average) will be $\sigma/\sqrt{N}$.

PDF of the sum or average of several random variables

Given two random variables x and y with joint PDF $p_{xy}(x, y)$, we can calculate the PDF of the derived random variable $z = x + y$ by considering all combinations of x and y whose sum is z. The resulting PDF must therefore be:

$$z = x+y \quad \Rightarrow \quad p_{x+y}(z) = \int_{-\infty}^{\infty} p_{xy}(x, z-x)\,dx = \int_{-\infty}^{\infty} p_{xy}(z-y, y)\,dy$$

The two integrals must be equivalent because each is just the integral of $p_{xy}(x, y)$ along the line $z = x + y$ (with specified, fixed z) in the x-y plane. If the two random variables x and y are independent, then the integral becomes a *convolution* of their two independent PDFs (the roles of x and y in the convolution may be exchanged without affecting the result):

PDF of a sum of two independent random variables

1.14
$$p_{x+y}(z) = \int_{-\infty}^{\infty} p_x(x)\,p_y(z-x)\,dx$$

If the sum is of $N>2$ independent random variables x_i, then the resulting PDF may be found by iteration: for example, if $w = x + y + z$, then first find the PDF of $u = x + y$ and use the result to find the PDF of $w = u + z$. Continue to iterate this procedure to include more terms, resulting in a sequence of convolutions: if $p_{N-1}(z)$ is the PDF for the sum of the first $N-1$ variables (with $p_1(z) \equiv p_{x_1}(x_1)$), then:

1.15
$$p_N(z) = \int_{-\infty}^{\infty} p_{x_N}(x_N)\,p_{N-1}(z-x_N)\,dx_N$$

Examples of this process will be discussed later. The average of N random variables is their sum divided by N. The width of the resulting distribution is reduced by a factor of N, and for the PDF to remain normalized it must be everywhere increased by a factor of N. If $\bar{x}$ is the random variable representing this average, then using (1.15) $\bar{x} = z/N$, and $p_{\bar{x}}(\bar{x}) = Np_N(z)$. Therefore $p_{\bar{x}}(\bar{x}) = Np_N(N\bar{x})$, with the sum PDF p_N calculated by iteration using (1.15).

Independent samples of a single random variable

Given the results for the distribution statistics of several different random variables considered jointly, we are now ready to define the statistics of a finite set of samples of a random variable. A set of N noisy measurements $\{x_i\}$, $i \in \{1 \ldots N\}$, is assumed to be part of a sequence of samples of a random variable x with PDF p_x, as already described. The ensemble PDF of each of the individual samples is also given by p_x, so we could consider each measurement x_i to be a single sample of a random variable x_i, all of which have identical PDFs. This implies that the distribution statistics of a set of N measurements is described by the joint statistics of N identical random variables.

Therefore, all of the relations found previously for the distribution statistics of functions of N distinct random variables will also apply to N samples of a single random variable x, assumed to describe our noisy measurements. *Our measurements are defined to be independent* if the joint PDFs of the N random variables x_i factor: $p_{x_j x_k} = p_{x_j} p_{x_k}$. In this case they are also uncorrelated, and their pairwise covariances vanish: $\sigma^2_{x_j x_k} = 0$. In practice, we must be especially careful with our experimental technique if we want statistical independence of our measurements. This issue will be examined in a later chapter.

> The statistics of N independent samples of a single random variable are the same as those of single samples from each of N independent random variables, all of which have identical distributions. Therefore we can use the formulas from this chapter to describe the statistics of multiple independent samples drawn from a single random variable.

NOISE AND THE NORMAL DISTRIBUTION

The normal distribution

The most important and studied continuous random distribution is the *Gaussian*, or *normal distribution*, thoroughly analyzed by J. Carl Friedrich Gauss in 1809.* The PDF for this distribution, (1.16), exhibits the familiar "bell curve" shape and was the distribution used for Figure 1-2 on page 6.

1.16 $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$ **Normal Distribution**

The two free parameters μ and σ in (1.16) are the distribution's *mean* and *standard deviation*, respectively, and completely define the distribution; the leading coefficient

* Gauss still remains probably the most brilliant and important mathematician to have ever lived. Along with the normal distribution, his 1809 text also introduced the *maximum likelihood* and *least-squares* methods, which we examine later.

normalizes the area under the PDF to 1. A plot of the normal distribution is shown in Figure 1-5, and some of its properties of are listed in Table II on page v. The normal distribution PDF falls quite rapidly away from its mean, dropping to $<14\%$ of the peak 2σ away.

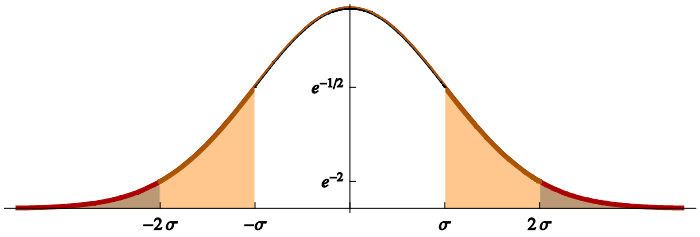


Figure 1-5: PDF plot of the normal distribution. The vertical axis is at the distribution mean μ; its ticks show the magnitudes (relative to the peak) at $\pm\sigma$ and $\pm 2\sigma$ from μ.

The sum or average of several independent Gaussians

Consider the sum of a set of N independent, normally-distributed (Gaussian) random variables x_i, with various means μ_i and standard deviations σ_i. From equations (1.11) and (1.12), we know that the sum distribution will have $\mu = \Sigma\mu_i$ and $\sigma^2 = \Sigma\sigma_i^2$. The sum PDF may be derived using the sequence of convolutions (1.15). In particular, we derive $p_{x_1+x_2}(z)$ in order to get the sequence started:

$$p_{x_1+x_2}(z) = \int_{-\infty}^{\infty} p_{x_1}(x)\, p_{x_2}(z-x)\, dx = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} \exp\left[\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right]\exp\left[\frac{-(z-x-\mu_2)^2}{2\sigma_2^2}\right] dx$$

$$p_{x_1+x_2}(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[\frac{-(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right] \quad \Rightarrow \begin{cases} \mu = \mu_1+\mu_2 \\ \sigma^2 = \sigma_1^2+\sigma_2^2 \end{cases}$$

Thus the distribution of the sum of two independent Gaussians, $p_{x_1+x_2}(z)$, is also Gaussian, with mean and variance equal to those given by equations (1.11) and (1.12). At each iteration in the sequence (1.15) to evaluate the sum distribution, the integral is a convolution of two Gaussians, resulting in another Gaussian. Thus the sum or average of several independent Gaussians is also Gaussian, with mean and variance given by (1.11) and (1.12).

The central limit theorem

In many situations wherein a large number of independent, relatively tiny effects combine to result in observed noise or scatter in a set of measurements, the resulting distribution of values is found to be modeled quite well by assuming that *the measurements are samples of a normally-distributed random variable*. This seemingly mysterious behavior was found to be explained by the aptly-named *central limit theorem*, one of the foundations of modern statistics and statistical mechanics. The gist of the theorem, for our purposes, is stated in the highlighted box on the next page.

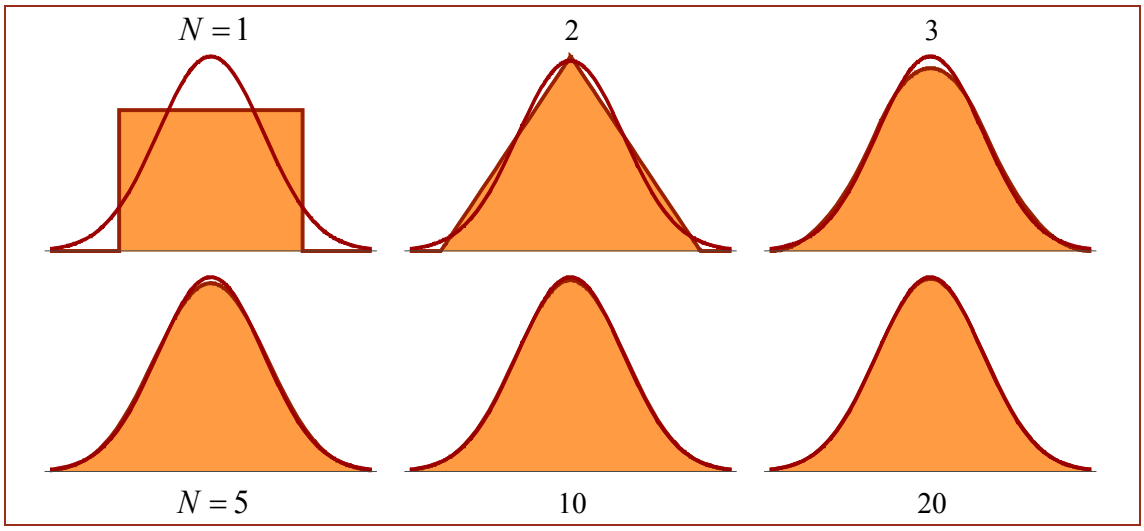


Figure 1-6: Comparisons of uniform-sum distribution PDFs to Gaussians for various numbers of independent terms in the sum. The N = 1 case shows the uniform distribution between ±1/2 assumed for each term. Each Gaussian has μ = 0 and $\sigma^2 = N/12$, the same as its corresponding uniform-sum (note that the horizontal and vertical scales differ from plot to plot). Note the rapid convergence of the sum PDFs to a Gaussian shape.

The Central Limit Theorem*

If a new random variable is created from the sum (or arithmetic mean) of a large number of independent random variables, then in the limit that the number of variables becomes infinite, ***the probability distribution of the new random variable is that of the normal distribution***. This remains true under very liberal conditions on the distributions of the constituent random variables being averaged.

No matter what may be the probability distributions of the individual, underlying random variables (and they may all have different distributions), then, as long as they are statistically independent of one another, the sum of a large number of them will be well-approximated by a normally distributed random variable! The only real requirements on the individual distributions are that they have well-defined, finite means and variances, although even these conditions may be further relaxed for some limited classes of distributions.

Infinity, after all, is a large number! Just how many random variables must be added for the resulting distribution to be well-represented by a Gaussian? For many common distributions of the underlying variables, the answer may be "less than you think," at least as long as we don't look too many standard deviations away from the mean (that's where the word "central" comes in). Consider this example: N random variables $x_1 \ldots x_N$, all with the same

* The theorem's more than 200 year history includes the works of many mathematicians from Abraham de Moivre in 1733 to Alan Turing in 1934. A 1922 work by the Finnish mathematician Jarl Lindeberg cast it into its modern form.

uniform probability distribution: $|x| \le 1/2$: $p(x) = 1$; $|x| > 1/2$: $p(x) = 0$ (so $\mu = 0$ and $\sigma^2 = 1/12$).

Form a new random variable z given by the sum of the N original, independent random variables: $z = \Sigma x_i$. The resulting PDF is that of a *uniform-sum* or *Irwin-Hall* distribution.* As N increases it rapidly converges on a Gaussian distribution, as was illustrated in Figure 1-6. The sum standard deviation $\sigma_N = \sqrt{N/12}$ (and thus the mean of N such variables has standard deviation $\sigma_N = 1/\sqrt{12N}$).

The χ^2 distribution

The variance of a random distribution x is its second moment about its mean: $\sigma^2 = \mathrm{E}[(x-\mu)^2]$. One may ask what the PDF of this squared deviation $(x-\mu)^2$ might be for some distribution of interest. Here we consider this question for the case of a Gaussian distribution ξ with mean 0 and variance 1, so that is:

$$p_\xi(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2}$$

(one can always transform a Gaussian variable x into ξ by setting $\xi^2 = (x-\mu_x)^2/\sigma_x^2$). Now transform ξ into $y(\xi) = \xi^2 \ge 0$, and investigate the PDF of this derived variable, $\chi^2 \equiv y(\xi)$. Using (1.6) on page 6 to get the PDF of y, which has two branches to its inverse function $\xi(y)$ and an integrable singularity at $y = 0$:

1.17

$$p_{\chi^2}(y>0) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}$$

χ^2 for a single Gaussian

This is known as the *chi-squared distribution with 1 degree of freedom*, which describes the distribution of the squared deviation from the mean of a single Gaussian random variable. More interesting is the distribution of a sum of ν independent χ^2 variables drawn from that same underlying Gaussian ξ. The result is known as the *chi-squared distribution with ν degrees of freedom*, χ^2_ν, and it will turn out to be very important when quantitatively evaluating theories or for fitting functions to measured data (see Chapter 3 and Chapter 4).

To get the PDF for χ^2_ν from (1.17), one iterates through the sequence of convolutions (1.15) on page 10 (not a trivial task):

1.18

$$p_{\chi^2_\nu}(y>0) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \sqrt{\frac{y^{\nu-2}}{2^\nu}}\ e^{-y/2}$$

χ^2 for ν degrees of freedom

The χ^2_ν PDF is identically 0 for $y < 0$. $\Gamma(\,)$ is the *gamma function*, related to the factorial function for positive arguments. For $\nu > 1$, the χ^2_ν PDF is nonsingular. See Figure 1-7 for

* Statistician Joseph Irwin (1898–1982) and mathematician Philip Hall (1904–1982), both British.

plots of a few χ^2_ν PDFs. The mean and variance of the χ^2_ν distribution are:

1.19 $$\mu_{\chi^2_\nu} = \nu \; ; \qquad \sigma^2_{\chi^2_\nu} = 2\nu$$ **χ^2 mean and variance**

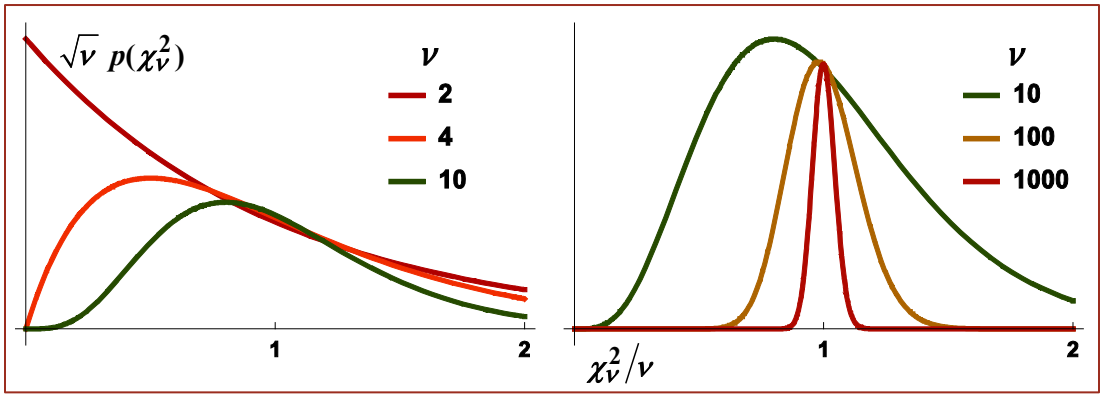


Figure 1-7: χ^2 distribution PDF plots for various degrees of freedom ν. The distribution approaches a Gaussian for large ν, as would be expected from the central limit theorem. For small ν, however, the distribution is quite asymmetric (skew). All PDFs vanish identically for $\chi^2 < 0$. The vertical and horizontal axes for each function are scaled as shown so that the functions are more easily compared.

The reduced chi-squared

Dividing χ^2_ν by its degrees of freedom ν results in the mean of ν independent, random χ^2 variables, rather than their sum. This new random variable is the *reduced chi-squared*, $\tilde{\chi}^2_\nu$:

1.20 $$\tilde{\chi}^2_\nu \equiv \chi^2_\nu / \nu$$ **Reduced chi-squared**

From (1.19) we can easily see that the $\tilde{\chi}^2_\nu$ distribution has a mean of 1 and a standard deviation of $\sqrt{2/\nu}$. The PDFs for various $\tilde{\chi}^2_\nu$ are what are actually plotted in Figure 1-7 (note the scaling of the x-axis). In terms of the reduced chi-squared, numerically integrating the PDF in (1.18) will show that the probability is greater than 84% that a $\tilde{\chi}^2_\nu$ variate will be less than one standard deviation above its mean, $P(\tilde{\chi}^2_\nu < 1+\sqrt{2/\nu}) > 0.84$, and the probability is less than 5% that $\tilde{\chi}^2_\nu$ will be more than 2 standard deviations above 1 (if $\nu = 1$ the probability is 94.96%). These observations will prove to be quite important when comparing measured data to theoretical predictions (Chapter 3 and Chapter 4).

Chapter 2
From samples to statistics

When one performs a measurement during the course of an experiment, it is essential to have some estimate of the accuracy of the resulting value. For example, after considering all potential sources of error, is an experimenter's measurement of an atomic emission line wavelength likely to be accurate to within 10 nanometers, 1 nanometer, or 0.1 nanometer? In this chapter we begin to tackle the problem of estimating the magnitudes of the errors in a set of measurements, which we refer to as the measurements' *uncertainties*. Once we have these uncertainties of our data points in hand, we must understand how these uncertainties as well as those introduced by other sources of error propagate through the mathematics required to compare our results to theoretical predictions. The framework presented in the previous chapter will be essential to our approach, an approach whose development and potential consequences occupy much of the remainder of this text. This chapter starts that development and is concerned with two fundamental topics: (1) how to properly estimate the uncertainties in a collection of data points, and (2) how to propagate uncertainties in numerical values through various mathematical calculations.

Determining measurement uncertainty

Estimating the magnitude of the possible error in a single, isolated, measured value only by considering experimental technique and instrument calibration and resolution, although sometimes unavoidable, can be notoriously inaccurate. We need to determine a data point's uncertainty in a way that is logically sound and defensible. To do this we begin by considering the types of error sources which can limit the accuracy of a measurement. Error sources have a seemingly infinite variety, but for now we can divide them into three broad categories: (1) *noise* and *drift*, which cause differences in the values obtained by repeated measurements; (2) *systematic* errors, such as instrument calibration errors and dimensional errors in the apparatus, which do not cause observable fluctuations in the values obtained by repeated measurements; and (3) measurement *precision* or *resolution*, which determines the smallest difference between measurement values which can be detected.

Noise vs. systematic error

Noise and *drift* are general categories of random errors which manifest themselves as a lack of repeatability in experimental measurements. The scatter in the values of repeated measurements illustrated in Figure 1-1 on page 1 serves as a typical example. This behavior has been the motivation for the development of the mathematical concept of samples of a random number explored in Chapter 1. The general distinction between *noise* vs. *drift*, as used in this text, is that *noise* describes those sources of random measurement errors which vary rapidly, so that the correlations in the errors of successive measurements can be made

small, whereas *drift* is caused by sources which vary slowly during the course of an experiment, causing a noticeable correlation in the errors in a set of successive measurements. Error due to drift can commonly also be seen when an experiment is repeated after several hours or days, generating new results which are incompatible with the observed scatter due to noise during the original experiment.

Data errors introduced by noise which cause independent fluctuations in successive measurements are the most tractable. The model introduced in the last chapter of data points as independent samples of a random variable will allow us to estimate the statistics of the underlying random variable in order to estimate uncertainties. This chapter's next section on ***Point estimation*** provides the details. Drift, on the other hand, because it leads to correlated errors in successive data points, usually requires special handling of some sort. Its effects may often be addressed using techniques similar to those needed to handle sources of systematic error, so we put off its analysis to Chapter 5, *Dealing with systematic errors*.

Systematic error is the general term used to describe errors introduced during the design, construction, and data acquisition phases of an experiment which affect the accuracies of all measurements in strongly correlated ways. Sources of such errors are everywhere you look in an experiment: the calibration error in a voltmeter, the angular alignment error of the fixed arms of an interferometer, the position and alignment errors in the placement of particle detectors around a particle collision site, the machining errors in the dimensions of a resonant cavity, particle trajectory errors introduced by improperly analyzed fringe fields of an electromagnet, changes in the dimensions or electrical characteristics of the apparatus caused by changes in laboratory temperature or humidity, etc. You can probably easily think of many more examples.

Because such fixed, systematic errors do not lead to seemingly random scatter in repeated measurements (as in Figure 1-1), the magnitudes of their individual effects on an experiment's accuracy can be much harder to determine. In many experiments of fundamental importance, unfortunately, these errors may be the dominant determiners of the experiments' accuracies. As mentioned in the previously, we delay the detailed consideration of such errors until Chapter 5.

Finite precision, resolution, and round-off errors

Lastly, consider the errors introduced by the *finite precision* of the measuring instruments used. By precision, we mean the "number of decimal digits" obtainable during a measurement. The closely related concept of *resolution* describes the smallest change in a measured value which is consistently detectable by an instrument. This precision or resolution may be limited by the number of digits available on a digital instrument's display, the number of bits used in an analog to digital converter (*ADC*), or the smallest length divisions scribed on a precision caliper, to name just a few. Errors introduced by such a truncation to a finite number of digits are variously called round-off errors, quantization

errors, or quantization noise, but all are basically the same in their effects on the data. The experimenter should, of course, choose a measuring instrument with a precision appropriate for the expected overall accuracy of the experiment, because high-resolution, precision instruments can be much more expensive to acquire and maintain than their less precise counterparts. If a measurement's accuracy will be dominated by one or more sources of systematic error, paying for instrument precision beyond that accuracy limit may be a waste, as would be the reduction of noise fluctuations to well below the accuracy limit imposed by that systematic error.

Properly estimating data uncertainties introduced by round-off or truncation can be subtle. A single measurement which is rounded off or truncated to a finite precision could represent any actual numerical value within the interval determined by the resulting measurement resolution. For example, measuring a length to the nearest millimeter implies that the actual length could be anywhere within $\pm 0.5\,\text{mm}$ of the recorded value. With nothing else to go on, one must assume that the actual length is equally likely to be anywhere within this interval, a uniform PDF within the interval and 0 outside it.* The *uniform distribution* described in the last chapter (see Figure 1-6 on page 13) characterizes this situation; it has a standard deviation of $1/\sqrt{12}$ times the resolution, or about $0.3\,\text{mm}$.

Uncertainty due to round-off

A single measurement rounded off to a finite increment has a distribution standard deviation (uncertainty) of $1/\sqrt{12}$ (about 0.3) of the round-off interval. If additional rounded off measurements yield the same value each time, ***then their mean cannot be assumed to have uncertainty decreasing as*** $1/\sqrt{12N}$.

As far as our data analysis procedures are concerned, a set of multiple measurements, each of which is rounded off or truncated to the same value, is completely equivalent to taking only a single measurement: no advantage is gained by taking multiple measurements.

Taking more measurements, even if they all round off to the same result, might tempt one to think that their mean might have a standard deviation of only $1/\sqrt{12N}$ of the round-off interval, but this conclusion is rarely justifiable. For example, assume that a resonant cavity has a length of $12.35\,\text{mm}$. The experimenter rounds off each of 20 measurements of its length to the nearest millimeter, $12\,\text{mm}$. Each measurement would then have a value and an uncertainty (standard deviation) of $12.0 \pm 0.3\,\text{mm}$, but the value and uncertainty of the mean of the 20 measurements, and thus of the recorded length of the cavity, are certainly not $12.0 \pm 0.06\,\text{mm}$! The reason for this is that the error of each measurement is $-0.35\,\text{mm}$, so

* The *likelihood PDF* is uniform within the interval and 0 outside it. We define the concept of likelihood in the next chapter.

that the errors in the measurements are strongly correlated and not due to independent, random noise (with $\sigma = 0.3\,\text{mm}$). This situation invalidates the formula for the uncertainty in the mean of multiple independent measurements presented in Chapter 1.

If the accuracy of the data is expected to be limited by noise, then one should choose an instrument which has a resolution sufficient to detect these noise-induced fluctuations in the data. As a result, repeated measurements will display fluctuations due to the noise. The experimenter can use the observed scatter in the measurements to analyze the noise and improve the accuracy of the experiment's results, as explained in this and the next few chapters. In other words, this lack of repeatability will turn out to be an asset! This fact is so important that digital instrument manufacturers often improve the performance of their products by artificially adding random noise to the data prior to measurement. The added noise is enough to cause observable fluctuations in the truncated or rounded off measurements limited by the instrument's basic resolution. By averaging the results of multiple, slightly noisy measurements, the ultimate precision of the instrument can then be improved.[*]

In the case of the length measurement example presented earlier, the experimenter should attempt to carefully interpolate each measurement to a better precision that the nearest millimeter. Multiple measurements could then show random scatter subject to the analysis techniques presented here, potentially improving the precision of the experiment's results.

> Always try to interpolate the markings on any continuous (analog) scale used to perform a measurement, such as on a ruler or analog voltmeter. Multiple measurements may then show fluctuations which can be averaged to improve the precision of the result.

Determining the overall uncertainty in an experiment's result

Inaccuracies introduced by noise and those introduced by systematic errors must be handled very differently. In the end, the effects of both categories of error sources must be included when determining the final uncertainty to be assigned to an experiment's result. We can say, however, that a general approach to estimate overall experimental accuracy is a two-step process:

1. Random, independent measurement fluctuations (lack of repeatability) introduced by noise are characterized and used to assign uncertainties to individual data point results. These data points are then compared to theoretical models whose free parameters are optimized using the techniques presented in this and the next two chapters. Provisional agreement with the optimized model or models can then be evaluated and initial uncertainties in the model's parameter values may be assigned.

[*] Among the techniques used are those referred to as *dithering*, *oversampling*, and *stochastic modulation*.

2. The effects of uncertainties due to the various systematic error sources are determined and those effects are combined with the uncertainties determined in (1) using error propagation techniques discussed later in this chapter. The final results will include uncertainties due to noise, the fitting and optimization of theory free parameters, and, finally, systematic errors.

POINT ESTIMATION

The most accurate and logically justifiable way to estimate a measurement's uncertainty due to lack of repeatability (noise) is by collecting a set of multiple data samples while attempting to maintain the same experimental conditions. In this section we develop the techniques to properly analyze a set of data samples to determine the best value for a measurement and its associated uncertainty introduced by this lack of exact repeatability. Since we are attempting to determine the value of a single physical quantity, our process is called *point estimation*.

One very important caveat concerning the method presented here: it can only deal with errors in the data which manifest as random, independent scatter (noise) in the values obtained following repeated measurements of the quantity. Other sources of error, such as calibration errors and gain drift, can be quite important and must be treated differently. See Chapter 5: *Dealing with systematic errors* for details.

Estimating the distribution mean and variance

As we attempt to accurately measure a single physical quantity, we repeat the measurement several times to obtain a set of samples with a variety of different measured values. If we were to continue to repeat the measurement many, many times, keeping the experimental conditions as constant as we can, we expect that the samples would be more or less clustered about some average, with a relative frequency which would converge to some smooth distribution, as in Figure 1-1 on page 1. This assumption formed the foundation of the analyses presented in Chapter 1.

As described in Chapter 1 we model our noisy data as independent samples generated by a random variable. If we knew the distribution of the random variable, then we could calculate probabilities and expected values from its PDF. In reality, however, all we have to work with is our finite set of measurements. This section describes simple techniques we may use to analyze our samples and generate an approximate model of an appropriate random variable, as well as how to estimate our uncertainty in the model's accuracy.

Assume that a noisy data set is well-modeled as samples of a random variable y. If y is a distribution with a well-defined mean and variance, as are the uniform and normal

distributions, then the expected value of any sample of y would be the distribution mean, μ. The noise-induced deviations of the various sample values away from μ will be characterized by the distribution variance, σ^2. If we are careful with our experimental technique, then we can model our data as independent samples of this distribution. From the section *The sum and arithmetic mean of a set of random variables* starting on page 9, we know that the arithmetic mean $\overline{y}$ of N independent data samples y_i will also have a distribution mean of μ, but it will have an expected variance about μ of σ^2/N. By the central limit theorem, we also expect that the distribution of $\overline{y}$ will be more like a Gaussian than the distributions of the y_i.

If the noise inherent in our measurement process is likely to introduce errors distributed symmetrically about the actual physical quantity Y we're interested in measuring, then the distribution mean μ of the sample set $\{y_i\}$ is appropriately identified with Y. If, on the other hand, the error distribution is expected to be somewhat asymmetric, then whether or not to use μ as our value for Y is more problematic. For this elementary discussion, however, we set aside this question and assume that μ is the value we seek. Consequently, we want the best *estimator* of μ derivable from our N-sample data set $\{y_i\}$.

In fact, the *best estimator of μ is our sample set's arithmetic mean* $\overline{y} = \Sigma y_i/N$. By "best" we mean that (1) $\overline{y}$ is *unbiased*: its expected value is μ no matter how large or small N may be, and (2) $\overline{y}$ *has a distribution with lower variance than any other unbiased estimator* of μ constructible from N samples drawn from a Gaussian random variable.[*] From (1.13) and the discussion following that equation we know that $\sigma_\mu^2 = \sigma^2/N$.[†] Therefore to properly estimate the uncertainty in $\overline{y}$ we require σ^2. Unfortunately, we have only our N particular samples y_i. We can, however, determine the expected value of $\Sigma(y_i - \overline{y})^2$ in terms of N and the distribution's hypothesized σ^2. This relation can then be inverted to provide us with an unbiased estimate of σ^2 using $\Sigma(y_i - \overline{y})^2$:

$$(y_i - \overline{y})^2 = (y_i - \mu - \overline{y} + \mu)^2 = (y_i - \mu)^2 + (\overline{y} - \mu)^2 - 2(y_i - \mu)(\overline{y} - \mu)$$

$$\therefore \mathrm{E}[(y_i - \overline{y})^2] = \mathrm{E}[(y_i - \mu)^2] + \mathrm{E}[(\overline{y} - \mu)^2] - 2\,\mathrm{E}[(y_i - \mu)(\overline{y} - \mu)]$$

$$= \sigma^2 + \frac{\sigma^2}{N} - \frac{2}{N}\mathrm{E}[(y_i - \mu)\Sigma(y_j - \mu)] = \sigma^2 + \frac{\sigma^2}{N} - \frac{2}{N}\left(\sigma^2 + \Sigma_{j \neq i}\, \sigma_{y_i y_j}^2\right)$$

$$= \sigma^2\left(1 - \frac{1}{N}\right) = \frac{N-1}{N}\sigma^2$$

[*] The *median* of N samples from a normal distribution, for example, has variance $(\pi/2)\sigma^2/(N-1)$. The proof that the sample mean $\overline{y}$ is a *minimum variance unbiased estimator* (MVUE) of the distribution mean is beyond the scope of this text. $\overline{y}$ will not be a MVUE if the underlying distribution has very large "wings," implying that its variance is not defined. In such cases, another average such as the median may be a better choice.

[†] By σ_μ^2 we really mean the variance of $\overline{y}$, our estimate of μ, which is $\mathrm{E}[(\overline{y} - \mu)^2]$. This is the uncertainty in our estimate of the distribution mean μ.

The final expression obtains because the samples are assumed to be independent, so their covariances vanish. Because this result will be the same for each of the samples, the sum $\Sigma(y_i - \bar{y})^2$ is expected on average to be N times this result, or $(N-1)\sigma^2$. Thus an estimate of the underlying random variable y's variance σ^2 is:

2.1 $$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$ ***Sample variance***

Our estimate is called s^2, the *sample variance*. It turns out that this is the best unbiased estimator of σ^2 constructible from N independent measurements. Note that if there were only a single measurement y, then $\bar{y} = y$ and $N = 1$, so that $s^2 = 0/0$, leaving σ^2 completely indeterminate, as would be expected from only a single measured data value.

The uncertainties in the estimations of μ and σ_μ

From the results of the previous section we find that we can use our N independent samples y_i to estimate the mean μ and variance σ^2 of the parent distribution by calculating the sample arithmetic mean $\bar{y}$ and the sample variance s^2. Because we really are interested the distribution mean μ and our uncertainty in its value, we should use an estimate of the variance in $\bar{y}$, namely $\sigma_\mu^2 = \sigma^2/N \approx s^2/N$. The square root of this value could then estimate the expected standard deviation of $\bar{y}$ around the desired result μ.*

Our resulting point estimate of the physical quantity Y we attempted to determine with our N independent measurements y_i is then given by:

Point estimate from N samples

2.2
$$Y \pm \sigma_Y = \bar{y} \pm \sqrt{s^2/N}$$
$$\bar{y} = \frac{1}{N}\sum_i y_i \; ; \quad s^2 = \frac{1}{N-1}\sum_i (y_i - \bar{y})^2$$

The *uncertainty* of Y is to be interpreted as meaning that, should the experiment be repeated many times, the distribution of the estimates of Y will have a standard deviation of σ_Y about the actual value of the physical quantity.

* It turns out that the square root of the sample variance s^2 is *not* the expected value of the standard deviation of N samples, but the ratio is close to unity except for very small sample sizes. For samples of a Gaussian random variable, the sample variance is described by the *chi-squared distribution* introduced on page 14; the sample standard deviation, on the other hand, is described by the closely related *chi distribution*. Since the uncertainty in the determination of either the variance or the standard deviation is large if N is small, then the error introduced by using the square root of the sample variance as the distribution standard deviation is unimportant.

Just how accurately do we know the uncertainty σ_Y? The χ^2 *distribution* discussed on page 14 and illustrated in Figure 1-7 provides an answer. An N-sample sum $\Sigma(y_i - \mu)^2/\sigma^2$ of an underlying Gaussian distribution is distributed as a χ^2 variate with N degrees of freedom, but the sum $(N-1)\,s^2/\sigma^2 = \Sigma(y_i - \bar{y})^2/\sigma^2$ has only $N-1$ degrees of freedom (because $\bar{y}$ is the average of the N samples and not independently determined, adding a constraint on the y_i). The variance of this latter χ^2 variate away from its mean of $N-1$ is, from (1.19) on page 15, equal to $2(N-1)$, so the expected standard deviation of s^2 away from σ^2 (which is the uncertainty in our estimate of σ^2) may be estimated as follows:

$$\sigma^2_{(N-1)\,s^2/\sigma^2} = 2(N-1) \quad \Rightarrow \quad \sigma_{(N-1)\,s^2/\sigma^2} \approx \sqrt{2(N-1)}$$

$$\therefore \quad \sigma_{s^2/\sigma^2} \approx \frac{\sqrt{2(N-1)}}{N-1} = \frac{\sqrt{2}}{\sqrt{N-1}}$$

$$\therefore \quad \sigma_{s^2} \approx \frac{\sqrt{2}}{\sqrt{N-1}}\sigma^2 \quad \Rightarrow \quad \boxed{\frac{\sigma_{s^2}}{s^2} \sim \frac{\sqrt{2}}{\sqrt{N-1}}}$$

This expression is only a rough approximation because the y_i may not have a Gaussian distribution, and for small N the final substitution $\sigma^2 \to s^2$ can be quite inaccurate. It will do for our purposes, however.

As we will show in the next section of this chapter the standard deviation of the square root of s^2 will be only half as large fractionally:

$$\frac{\sigma_s}{s} = \frac{1}{2}\frac{\sigma_{s^2}}{s^2}\,; \quad \therefore \frac{\sigma_s}{s} \sim \frac{1}{\sqrt{2(N-1)}}$$

This then gives us an estimate of "the uncertainty of the uncertainty" in our value for the physical quantity Y:

2.3 Accuracy of the uncertainty

$$\boxed{\frac{\sigma_{s^2}}{s^2} \sim \frac{\sqrt{2}}{\sqrt{N-1}}\,; \qquad \frac{\sigma_{\sigma_Y}}{\sigma_Y} \sim \frac{1}{\sqrt{2(N-1)}}}$$

For only two samples, any uncertainty calculation is probably only an order-of-magnitude estimate; for 5 samples it is approximately accurate to 35% (and the variance's uncertainty is twice as large: a disappointing 71%). The moral of the story: don't take the uncertainty estimate in (2.2) as very accurate if the samples are few.

The estimation of σ^2 using the sample variance s^2 will be very important to the utility of the material presented in the following chapters of this text. You will want as accurate an estimation of σ^2 as you can get in order to properly interpret quantitative comparisons of theory to your results. The expression in (2.3) will be particularly important to keep in mind as you study the materials in Chapter 3 and Chapter 4.

Examples of point estimation

Consider first the data shown in Figure 2-1, consisting of six independent voltage measurements collected while attempting to maintain the experimental conditions constant. Using the equations (2.2), the estimates of the distribution mean and standard deviation are given by the arithmetic mean of the samples, 883.7V, and the sample standard deviation, 2.7V. The estimated uncertainty in the distribution mean is then $2.7\text{V}/\sqrt{6} = 1.1\text{V}$. These uncertainties, 2.7V and 1.1V, are illustrated by the dotted and dashed lines in Figure 2-1. From expression (2.3), the accuracy of the uncertainty estimate is ~32%.

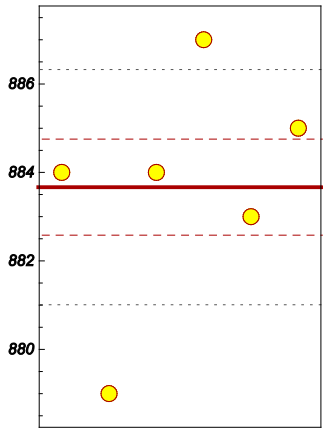


Figure 2-1: Plot of a sequence of six voltage measurements obtained while the experimenter carefully attempted to maintain the same conditions for each voltage determination. The thick red line is at the arithmetic mean of the voltage values. The two dotted gray lines are at ± one sample standard deviation from this mean. The dashed red lines are at ± the uncertainty in the distribution mean value. All were calculated using equations (2.2).

For the next example, consider the much larger data set in Figure 2-2. The data consist of gamma-ray event counts vs. energy channel number for a portion of a multi-channel analyzer spectrum output generated by a scintillation detector. The event rate is expected to be very nearly independent of gamma-ray energy for this region of the spectrum, and an estimate of the expected counts/channel and its uncertainty is desired. Using expressions (2.2) with the 88 samples results in a mean rate of 82.1 ± 1.0 counts/channel (shown by the red solid and

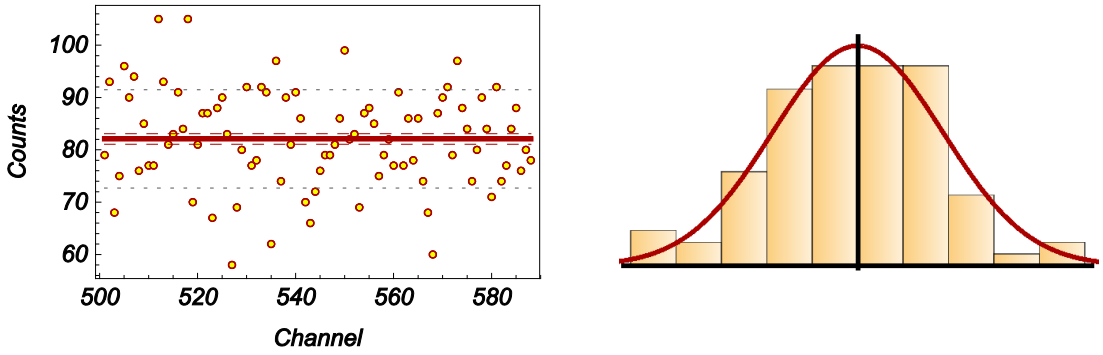


Figure 2-2: Gamma-ray detector event counts vs. energy channel number (both channel numbers and counts are integers for each data point). The right-hand plot is a histogram of counts/channel along with a Gaussian representing the sample data mean and the sample standard deviation.

dashed lines in the left-hand plot in Figure 2-2), and the calculated sample standard deviation about this mean is 9.4 counts/channel (the gray dotted lines in the same plot). Because there are 88 data samples, the uncertainty in the estimated mean rate (± 1.0) is an order of magnitude smaller that the observed spread in the data (± 9.4).

The right-hand plot in Figure 2-2 is a histogram of the relative frequencies of the various counts/channel values along with a plot of a Gaussian with mean 82.1 and standard deviation 9.4. It appears that the Gaussian distribution provides a fair representation of the observed data frequencies, at least by eye. A quantitative test of the consistency of a sample set with a Gaussian distribution is presented in the next chapter. Pure counting statistics would require that the distribution of counts/channel data be described by a *Poisson distribution* (see Chapter 6). If this were the case, then the expected standard deviation in the counts/channel data should be given by the square root of the expected count number, or, given the observed mean rate, 9.1. The estimated uncertainty in the distribution standard deviation calculated from the 88-sample standard deviation of 9.4 is, using (2.3), approximately 7.6% or about 0.7 counts/channel. Thus, given its expected accuracy, the sample standard deviation of 9.4 is quite consistent with the Poisson distribution's value of 9.1.

PROPAGATION OF UNCERTAINTIES

Assume an experiment results in the determination of one or more numerical values with associated uncertainties: $x \pm \sigma_x$, $y \pm \sigma_y$, etc., possibly by using the point estimation technique of the previous section. Now we wish to use these values in an algebraic expression in order to calculate some derived quantity $z = f(x, y, \ldots)$. What uncertainty σ_z should be assigned to the calculated value z? This is the problem addressed by the techniques of *error propagation*, which are really methods to *propagate uncertainties* through numerical calculations.

Functions of a single uncertain value

First consider the case of an expression $z = f(x)$ containing a single experimentally-determined value $x \pm \sigma_x$. This implies that the value of the expression will then have an associated uncertainty: $z \pm \sigma_z = f(x \pm \sigma_x)$. How do we estimate $z \pm \sigma_z$? For example, a sound wave with a very precisely known frequency ν has its wavelength measured to be $\lambda \pm \sigma_\lambda$; you wish to use this value to determine the speed of sound $c = \lambda \nu$ and its associated uncertainty σ_c. By the phrase "experimentally-determined value $x \pm \sigma_x$" we mean, for example, the value and uncertainty determined by a point estimate such as expressions (2.2) on page 22.* Our goal is then to estimate the expected value of the derived quantity $z = f(x)$ and its standard deviation. If we knew the correct probability density function p_x, we could

* Another example would be that σ_x is the standard deviation of the *likelihood function* about our experimentally-determined value for x. More about the likelihood function in the next chapter.

calculate any desired statistics of $z = f(x)$ in the standard way using (1.2) (assuming the integrals converge):

$$\mu_z = \mathrm{E}[f(x)] = \int_{-\infty}^{+\infty} f(\xi) p_x(\xi) d\xi; \qquad \sigma_z^2 = \left(\int_{-\infty}^{+\infty} f(\xi)^2 p_x(\xi) d\xi \right) - \mu_z^2$$

In the case we're considering, however, we don't know p_x in any detail (although we may, for example, expect that it is approximately Gaussian). To proceed in the face of our incomplete knowledge we must content ourselves with estimates. Assume that x is our best estimate of the distribution mean μ_x and that σ_x is its uncertainty about μ_x. Expand the function $z = f(x)$ in a Taylor series about μ_x and consider only the first few terms:

2.4 $$z = f(x) = f(\mu_x) + f'(\mu_x)(x - \mu_x) + \tfrac{1}{2} f''(\mu_x)(x - \mu_x)^2 + \ldots$$

The expected value of z is then found by taking the expected value of each $(x - \mu_x)^n$ factor in the expansion, where we've used an overscore symbol to indicate the averaging operation (1.2) to calculate the expected value of an expression containing x:

$$\mu_z = \overline{f(x)} = f(\mu_x) + f'(\mu_x) \cancel{\overline{(x - \mu_x)}} + \tfrac{1}{2} f''(\mu_x) \overline{(x - \mu_x)^2} + \ldots$$

$$\mu_z = f(\mu_x) + \tfrac{1}{2} f''(\mu_x) \sigma_x^2 + \ldots$$

The first derivative term vanishes because the expected value of x is μ_x. Thus our estimate of the expected value of $z = f(x)$ is given by the final expression above. Since the measured value x is our best estimate of μ_x, and σ_x is our best estimate of its uncertainty, we substitute these values to determine our best estimate of the value of $\mu_z \to z$:

2.5 $$\boxed{z \approx f(x) + \tfrac{1}{2} f''(x) \sigma_x^2}$$ **Expected value of $z = f(x \pm \sigma_x)$**

Again, x is our best estimate of its distribution mean μ_x, and σ_x is the experimental uncertainty of that estimate. Then (2.5) provides our estimate of z's resulting expected value. Often the second derivative term is relatively small and can be ignored, so that a simple, adequate estimate is $z \approx f(x)$; this is the approximation used by default by the data analysis package *CurveFit*.

> One obvious exception wherein it would be incorrect to discard the second derivative term in (2.5) is the important case $z = x^2$ with $\mu_x = 0$. In this case $\mu_z = \sigma_x^2$. Such a situation arises, for example, when we are interested in the average power generated by a noise source.

Now for the estimate of σ_z. By equation (1.5) $\sigma_z^2 = \mathrm{E}[z^2] - \mu_z^2$. Squaring the previous expansion (2.4) and keeping terms up to order 2,

$$z^2 = f(x)^2 = \left[f(\mu_x) + f'(\mu_x)(x-\mu_x) + \tfrac{1}{2} f''(\mu_x)(x-\mu_x)^2 + \ldots \right]^2$$
$$= f(\mu_x)^2 + 2 f(\mu_x) f'(\mu_x)(x-\mu_x) + f'(\mu_x)^2 (x-\mu_x)^2 + f(\mu_x) f''(\mu_x)(x-\mu_x)^2 + \ldots$$

Taking the expected value of this expression,

$$\mathrm{E}[z^2] = f(\mu_x)^2 + f'(\mu_x)^2 \sigma_x^2 + f(\mu_x) f''(\mu_x) \sigma_x^2 + \ldots$$

Subtracting the square of expression (2.5) for μ_z and keeping terms of second order and below, the only surviving term gives our estimate of the uncertainty of z: (where we again assume that 'x' is our best estimate of the distribution mean μ_x):

$$\sigma_z^2 \approx f'(\mu_x)^2 \sigma_x^2$$

2.6 $$\boxed{\sigma_z \approx \left| f'(x) \right| \sigma_x}$$ **Uncertainty of $z = f(x \pm \sigma_x)$**

The expression (2.6) may be a bit naïve in some situations, such as the simple case $z = x^2$ with $\mu_x = 0$ mentioned earlier (for which (2.6) would give vanishing σ_z). A more accurate expression, applicable if the distribution of the estimate of x is Gaussian, is:

2.7 $$\sigma_z^2 \approx f'(\mu_x)^2 \sigma_x^2 + \left[\tfrac{1}{2} f''(\mu_x)^2 + f'(\mu_x) f'''(\mu_x) \right] (\sigma_x^2)^2$$

In the case of $z = x^2$ about $\mu_x = 0$, this more accurate expression gives $\sigma_z \approx \sqrt{2}\, \sigma_x^2$. The more nonlinear the function $f(x)$, the wiser it would be to consider the more accurate expressions, but in many cases the simple, naïve error propagation formulas (2.5) and (2.6) will be adequate. Table I on page iii provides examples of naïve uncertainty propagation formulas for some common functions; Table III on page v provides moments about the mean of the normal distribution useful for developing higher-order formulas like (2.7).

Functions of two or more uncertain values

We now extend our naïve error propagation formulas (2.5) and (2.6) to the case of a function involving two uncertain values: $z \pm \sigma_z = f(x \pm \sigma_x, y \pm \sigma_y)$. The obvious extension of (2.5) for the expected value is $\mu_z = f(\mu_x, \mu_y)$. The same caveats apply for cases wherein $f(\mu_x, \mu_y) = 0$, but not all second partial derivatives of f vanish. Assuming that this is not the case, we will use the naïve formula of simply inserting the expected value estimates of x and y (our measured values) into the function $f(x, y)$ for z.

An estimate of the variance σ_z^2 is then found using the first terms of a two-dimensional Taylor expansion about $f(\mu_x, \mu_y)$:

$$z^2 = f(x,y)^2 = \left[f(\mu_x,\mu_y) + (x-\mu_x)\frac{\partial f}{\partial x} + (y-\mu_y)\frac{\partial f}{\partial y} + \ldots \right]^2$$
$$= f(\mu_x,\mu_y)^2 + 2(x-\mu_x)f(\mu_x,\mu_y)\frac{\partial f}{\partial x} + 2(y-\mu_y)f(\mu_x,\mu_y)\frac{\partial f}{\partial y}$$
$$+ \left(\frac{\partial f}{\partial x}\right)^2 (x-\mu_x)^2 + \left(\frac{\partial f}{\partial y}\right)^2 (y-\mu_y)^2 + 2(x-\mu_x)(y-\mu_y)\frac{\partial f}{\partial x}\frac{\partial f}{\partial y} + \ldots$$

The partial derivatives are evaluated at (μ_x, μ_y). Taking the expected value of this expression and then subtracting $\mu_z^2 = f(\mu_x, \mu_y)^2$ to get the variance of z:

$$\sigma_z^2 = \overline{z^2} - \mu_z^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\sigma_{xy}^2 + \ldots$$

Note that the last term in this expression includes the product of the two first partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$, not the mixed second derivative $\partial^2 f/\partial x \partial y$. The covariance σ_{xy}^2 will generally be nonzero unless x and y are independent. If x and y are parameter values jointly found by fitting a set of experimental data (see Chapter 4), then it will almost certainly be the case that the covariance σ_{xy}^2 is nonzero and should be included in the estimation of σ_z.

The extension of this result to more than two independent variables is straightforward. If z is a function of n uncertain variables then the lowest order, naïve estimates of the expected value of z and its uncertainty are given by:

Naïve uncertainty propagation for a function of several variables

$$z \approx z(x_1, x_2, \ldots, x_n)$$
$$\sigma_z \approx \sqrt{\sum_{i=1}^{n}\left(\frac{\partial z}{\partial x_i}\right)^2 \sigma_{x_i}^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\frac{\partial z}{\partial x_i}\frac{\partial z}{\partial x_j}\sigma_{x_i x_j}^2} \qquad (2.8)$$

If the arguments are all independent random variables, then the sum over the covariance terms in (2.8) is not needed, because the covariances vanish. In this case the uncertainty in z is just given by a *Pythagorean sum* (square root of the sum of the squares) of the contributions to it from each of the independent variables $(x_1, x_2, \ldots, x_n)$.

An uncertainty propagation example

Measuring the force between two charged plates as a function of the voltage applied between them can provide an experimental determination of the permittivity ε of the air

separating the plates. The force is set using a laboratory balance to apply an opposing force mg to one plate with a small test mass m and acceleration due to gravity g. The applied voltage V required to just overcome this opposing force is then measured. The theoretical formula relating ε to V and m becomes:

2.9

$$\varepsilon = \frac{2gd^2}{\pi r^2}\left(\frac{dV^2}{dm}\right)^{-1}$$

The slope dV^2/dm relating the square of the applied voltage V to the test mass m is determined, along with its uncertainty, from the experiment's data using the methods described in Chapter 4. The other parameters are the plates' separation d and effective radius r. These parameters were carefully measured and their uncertainties estimated using the point estimation techniques described earlier. The results of these efforts are summarized here:

$$\begin{aligned} dV^2/dm &= 7.126 \pm 0.018\ (\text{kV})^2/\text{gram} \\ d &= 0.309 \pm 0.001\ \text{cm} \\ r &= 3.073 \pm 0.001\ \text{cm} \\ g &= 979.6\ \text{cm/sec}^2 \end{aligned}$$

We must now calculate the experimentally-determined value of ε and its uncertainty. The wisest way to propagate uncertainties is to do it incrementally using the propagation tables starting on page iii. We first note that the uncertain values on the RHS of (2.9) are raised to various powers, so Table I tells us that the various factors have uncertainties of:

$$\begin{aligned} \frac{\sigma_{d^2}}{d^2} &= 2\frac{\sigma_d}{d} = 0.00647 \\ \frac{\sigma_{r^{-2}}}{r^{-2}} &= 2\frac{\sigma_r}{r} = 0.00065 \\ \frac{\sigma_{dV^2/dm}}{dV^2/dm} &= 0.00253 \end{aligned}$$

Now use the formula for the propagation of a product of uncertain values:

$$\text{if:}\quad z = a\,x\,y, \quad \text{then:}\quad \frac{\sigma_z}{|z|} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$$

Applying this to (2.9),

$$\frac{\sigma_\varepsilon}{\varepsilon} = \sqrt{0.00647^2 + 0.00065^2 + 0.00253^2} = 0.007, \text{ or just under 1\%.}$$

Chapter 3
Likelihood and hypothesis testing

Assume that you have performed some experiment to determine, for example, q_e/m_e, the charge/mass ratio of the electron. Your data values may have been meant to provide direct measurements of q_e/m_e and were analyzed using the point estimation method of Chapter 2, or they may have consisted of pairs of values which were then fitted by a function that included q_e/m_e as a parameter. Your objective is to state that you have determined the fundamental constant of nature, q_e/m_e, to have some uncertain value $Y \pm \sigma_Y$, that is, $q_e/m_e = Y$ with an experimental uncertainty of σ_Y. This chapter explores the *maximum likelihood method* for determining the result of such an experiment and discusses how the quoted uncertainty should be interpreted.

SELECTING AMONG HYPOTHESES USING MAXIMUM LIKELIHOOD

To illustrate the concepts of the *likelihood function* and *the maximum likelihood method*, let us further consider the above example of the experimental determination of a fundamental constant of nature. Given the "true" (but unknown) value of some physical parameter Y, one may ask what would be the probability density of obtaining an experimentally measured value y for that parameter: $p(y\,|\,Y)$ (to be read: "probability density of obtaining y given actual condition Y"). If you have a good idea of how the experiment behaves, then you could theoretically estimate $p(y\,|\,Y)$ for any particular postulated "true" value of Y. For example, the left-hand graph in Figure 3-1 shows possible plots of $p(y\,|\,Y)$ as a function of experimental result y (the x-axis) for three different postulated values of the underlying parameter Y. A vertical line through our actual experimental result y (on the x-axis) intersects the various $p(y\,|\,Y)$ curves at different resulting probability density values, depending on

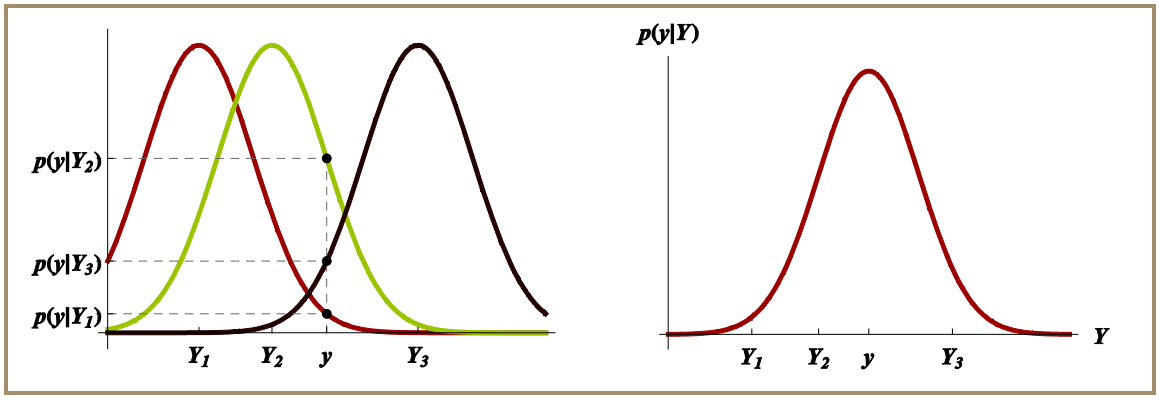


Figure 3-1: Generating the *likelihood function* $p(y|Y)(Y)$: (Left) plots of various PDFs for the expected distribution of experimental results given three choices Y_1, Y_2, Y_3 for the value of the unknown constant of nature Y. The actual experimental result y would have the PDF value shown for each choice of Y. (Right) plot of the resulting likelihood function $p(y|Y)$ as the value chosen for Y is varied.

how close our experimental result y is to each assumed value for the underlying constant Y.

Alternatively, with an actual, particular experimental result y in hand, a single plot of the expected $p(y \mid Y)$ *as a function of various postulated "true" values* Y could be generated, as in the right-hand plot in Figure 3-1. This new, single probability density $p(y \mid Y)$ plotted as a function of Y (and not y) is called the *likelihood function of Y given an experimental result y*. In our initial example of an experimental measurement of q_e/m_e, the condition Y would be a postulated actual value of q_e/m_e, and y would be our measured q_e/m_e.*

As an example of the determination of the likelihood function, assume that noise in the measurement is such that experimental results are expected to be normally distributed around the true value Y with standard deviation σ (independent of the value of Y). In this case the Gaussian PDF $p(y \mid Y)$ will depend on $(y-Y)^2$, symmetric in y and Y, depending only on the magnitude of their difference. Now consider Figure 3-1 again: as the assumed value for Y is varied the resulting *likelihood function* PDF $p(y \mid Y)(Y)$ must also be a Gaussian with standard deviation σ, but with mean y, as shown in the right-hand plot of the figure.

$Y_{\max}$, the assumed value of Y which maximizes $p(y \mid Y)$, is then the *most likely value* for Y given our particular experimental result y. In the example shown in Figure 3-1, $Y_{\max} = y$. For this simple example of an experimental measurement of q_e/m_e we would, naturally, pick the experiment's determination of q_e/m_e as the value we would quote for our result, corresponding also to the likelihood PDF's maximum, $Y_{\max}$. The result's uncertainty σ_Y would then be the standard deviation of the likelihood PDF, which in this case is the same as the measurement distribution standard deviation σ caused by the added noise of the measurement process.

Choosing among different values for Y represents the consideration of *different, alternative hypotheses*, in this case concerning the value of q_e/m_e. By choosing the maximum of the likelihood function as the value for Y most consistent with the experiment's results, we have used the *maximum likelihood method* for selecting among alternative hypotheses or theories (in this case, the hypothesized value of Y). In other words, *we selected the hypothesis that, if correct, would maximize the probability of obtaining our experimental result*.

> The *maximum likelihood method* chooses among alternative hypotheses or competing theories by selecting that hypothesis which maximizes the likelihood function of the experiment's result.

* Note that, given our fractional or frequency definition of probability presented in Chapter 1, we do not refer to the likelihood function as a converse conditional probability density $p(Y \mid y)$. In our view, this PDF would be a *Dirac delta function* located at the actual, "true" value of Y, independent of any experimental result y. The author does not consider the popular *Bayesian interpretation* of probability, "degree of rational belief," as particularly rigorous (English statistician and philosopher Thomas Bayes, c. 1701–1761), especially in the context of the scientific method as used in the physical sciences. Some may counter that Bayesian probability theory is no less rigorous than our assumed limiting cases of infinite processes and statistical ensembles.

If the various hypotheses differ only by their choices of values for one or more fundamental constants or other numerical parameters, then the likelihood function takes the form of a surface in a multi-dimensional space of likelihood vs. these parameter values. In this case we choose the global maximum of the likelihood function. Our previous example was of this type: one fundamental constant q_e/m_e, the alternative hypotheses being the various possible numerical values of this constant.

The main justification for using the maximum likelihood method may simply be to ask, "Why not?" Why wouldn't one choose the hypothesis that made your experimental result the most likely one? Another reason is that if the likelihood function is differentiable, then its maximum may be found by examining the zeroes of its derivatives, simplifying the math. These are both very powerful reasons for using the maximum likelihood method as a heuristic for choosing among alternative hypotheses, and this method will form the foundation of the methods you will use to analyze your experiments' results.

If the likelihood distribution is symmetric about its mean and has a single maximum, then its maximum coincides with its mean. In this case the maximum likelihood method also finds the experiment's *mean likelihood*, which is often considered to be another appropriate indicator of the best selection among alternative hypotheses. One must always bear in mind that if the likelihood distribution for an experimental result is strongly asymmetric about its mean (a *skew* distribution), maximizing likelihood may not be the wisest strategy. For the results expected from undergraduate physics lab experiments this should not be a huge worry, because the expected distributions of experimental measurements should usually be nearly symmetric about their means. Thus we will continue under the assumption that the maximum likelihood method is appropriate.

CHI-SQUARED AND MAXIMUM LIKELIHOOD

Weighted mean of several measurements

Consider this application of the maximum likelihood method: determine a properly-weighted average (weighted mean) of several uncertain numerical values. For example, assume that there are N independent experimental measurements of q_e/m_e with various uncertainties: $y_i \pm \sigma_i$. You wish to combine these measurements to determine a single $q_e/m_e = (\bar{y} \pm \sigma_{\bar{y}})$, making optimal use of the set of N measurements. We therefore use the maximum likelihood method to choose that value for $\bar{y}$ which maximizes the probability of having obtained the set of N experimental results $y_i \pm \sigma_i$.

Assume that Gaussian noise in each of the various measurements led to the scatter in the observed results, and that the measurements are samples from distributions with a common mean μ, corresponding to the actual value of the fundamental constant. Some experiments were noisier than others, and the various experimental uncertainties σ_i associated with each

of the measurements y_i reflect this fact. Because the noise is Gaussian, the distribution probability density for each individual experimental result would be given by the normal distribution, equation (1.16). The set of N independent values y_i would then have a joint probability density given by the product of their individual densities:

$$p(y_1, y_2, \ldots, y_N) = p(y_1)p(y_2)\ldots p(y_N)$$

3.1

$$\propto \prod_{i=1}^{N} \exp\left[\frac{-(y_i-\mu)^2}{2\sigma_i^2}\right] = \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{\sigma_i^2}\right]$$

In (3.1) we've assumed that the variances σ_i^2 associated with the measurements may be different, but all distributions have the same mean μ. This common value of μ in (3.1) is what we want to estimate from the data set. Given any hypothesized value for μ, expression (3.1) would then give the likelihood function for that choice. We will use the maximum likelihood method to choose our best estimate $\bar{y}$ for μ by maximizing expression (3.1) with respect to it.

Since the expression (3.1) for the joint PDF is differentiable, to find the maximum we could take its derivative with respect to μ and set it to zero: the solution $\bar{y}$ for μ would then be the maximum likelihood value. Rather than doing that calculation directly, however, we solve a simpler problem. We realize that the maximum of $p(y_1, y_2, \ldots, y_n)$ is also the maximum of its logarithm, so to simplify the math we instead maximize $\log p(y_1, y_2, \ldots, y_n)$ with respect to μ. Taking the logarithm of (3.1) and finding the value of μ where its slope vanishes:

3.2

$$0 = \frac{d}{d\mu}\log\left\{\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{\sigma_i^2}\right]\right\} = \frac{d}{d\mu}\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{\sigma_i^2}\right]\Bigg|_{\mu=\bar{y}}$$

$$= \sum_{i=1}^{N}\frac{y_i-\bar{y}}{\sigma_i^2} = \left(\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2} - \bar{y}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)$$

$$\therefore\ \bar{y} = \left(\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}\right) \div \left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)$$

The uncertainty in $\bar{y}$ may be derived from the uncertainties in the y_i using the uncertainty propagation formula (2.8). The resulting weighted mean and uncertainty estimates are then:

Weighted mean of several independent measurements

3.3

$$\bar{y} = \left(\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}\right) \div \left(\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right); \qquad \frac{1}{\sigma_{\bar{y}}} = \sqrt{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}$$

This, then, is the set of equations we would use to combine the several experimental results $y_i \pm \sigma_i$ into a maximum likelihood estimate of an underlying physical constant's value $Y = (\bar{y} \pm \sigma_{\bar{y}})$ such as a determination of q_e/m_e from the results of several independent measurements. Expressions (3.3) are also a generalization of the point estimate expressions (2.2), so they could be used to combine several measurements acquired during a single experiment (if those measurements were taken in different ways, which could lead to a variety of uncertainty estimates s_i for the various data points).

Chi-squared minimization

The above example demonstrated a maximum likelihood calculation. In that example the data were assumed to be independent samples of Gaussian random variables, and the calculation (3.2) to maximize the logarithm of the likelihood PDF with respect to the unknown mean μ was to find the zero of its derivative with respect to μ:

$$\frac{d}{d\mu}\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{\sigma_i^2}\right] = -\frac{1}{2}\frac{d}{d\mu}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{\sigma_i^2} = -\frac{1}{2}\frac{d}{d\mu}\chi^2_{N-1}$$

where $\chi^2_{(N-1)}$ is a sample of a chi-squared distribution with $N-1$ degrees of freedom, as described in the section *The χ2 distribution* starting on page 14.*

Because of the minus sign in front of the summation, finding the value of the unknown parameter μ which maximizes the likelihood is clearly equivalent to minimizing χ^2. Thus, *if the data points are independent and each is a sample of a Gaussian distribution, then* χ^2 *minimization of the experimental result with respect to an unknown parameter yields the parameter's maximum likelihood value*. Chapter 4 examines this point in greater detail. The next section discusses the importance of the χ^2 distribution when comparing experimental results to theoretical predictions.

Reduced chi-squared tests

Continue to consider the problem of comparing an experiment's results to a theoretical model. Assume that the noise present in the data is Gaussian, and that you have collected a total set of N independent data points y_i and have characterized the noise so well that you can assign an associated uncertainty (standard deviation) σ_i to each point. A theory predicts that your experimental result should be some value Y. If your data are consistent with the theory's prediction, then it is clear that the common mean of the Gaussian distributions of which your data points are samples should be $\mu = Y$.

* The number of degrees of freedom is one less than the number of data points N because μ is calculated from them, a single constraint reducing the remaining degrees of freedom by one. More will be said about this in Chapter 4.

Since the variance of the distribution associated with a data point y_i is σ_i^2, then if $\mu = Y$,

3.4

$$\mathrm{E}[(y_i - Y)^2] = \sigma_i^2 \; ; \qquad \therefore \; \mathrm{E}\left[\frac{(y_i - Y)^2}{\sigma_i^2}\right] = 1$$

On average, if the theory were correct, the expected value of each data point's deviation from the theoretical prediction Y would be described by (3.4), and the sum of these terms over the N data points would be expected to have value N. As discussed in Chapter 1, the χ^2 distribution describes the distribution of a sum of the squared deviations from the mean of a set of independent samples of a Gaussian with unit variance. Thus, a sum over the N data points of the terms (3.4) would be a sample of a χ^2 variate* with N degrees of freedom†, if it were indeed the case that $\mu = Y$:

3.5 **Chi-squared test calculations**

$$\chi_N^2 = \sum_{i=1}^{N} \frac{(y_i - Y)^2}{\sigma_i^2}$$

$$\tilde{\chi}_N^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{(y_i - Y)^2}{\sigma_i^2}$$

As was described in Chapter 1, the expected value of the χ_N^2 distribution is N, and it has a standard deviation of $\sqrt{2N}$ (expression (1.19) on page 15). Its companion *reduced chi-squared*, $\tilde{\chi}_N^2$, also shown in (3.5), is defined as χ_N^2 / N, and is therefore a weighted average of the squared deviations of the y_i from their common mean. The reduced chi-squared has an expected value of 1 and a standard deviation of $\sqrt{2/N}$. From the characteristics of its distribution listed in Table IV on page vi, we would expect a $\tilde{\chi}_N^2$ sample to exceed 1 by more than two standard deviations only about 5% of the time.

Thus the calculations (3.5) provide a useful test of the compatibility of a data set with a theory's prediction. The next chapter will extend this analysis to the more comprehensive and important case of comparing the functional relationship between two experimental quantities to a theoretical prediction.

The *reduced chi-squared test* using (3.5) provides a simple, quantitative comparison of a theoretical prediction to an experimental result: if the theory $\mu = Y$ were correct, then one would expect that more than 95% of the time, $\tilde{\chi}_N^2 - 1 < \sqrt{8/N}$.

Beware, however! Such comparisons may be trusted only if: (1) the distributions of the measurement errors in the y_i are reasonably Gaussian, and (2) we accurately know the distribution standard deviations σ_i.

---

* Because each term in the sum is scaled by dividing by its variance, each scaled term has $\sigma^2 = 1$.

† If the hypothesized common mean Y is calculated from the data values, as in the weighted mean example, then the number of degrees of freedom $\nu = N - 1$. More will be said about this in the next chapter.

If a calculation of (3.5) results in a very small value, $\tilde{\chi}^2_N \ll 1$, then the analysis is quite problematic. If $N > 5$, the probability is less than 4% that one would obtain $\tilde{\chi}^2_N < 0.2$. For any $N < 40$ the probability of a $\tilde{\chi}^2_N$ value more than two standard deviations below 1 is less than 1%:

$$P\left[\tilde{\chi}^2_N < 1-\sqrt{8/N}\right] < 1\%$$

Even in the limit of *very* large N, the probability is only approximately 2.25%. Therefore it is quite unlikely to obtain an experimental result which closely matches a theoretical prediction (more closely than the uncertainties σ_i would indicate). A calculated $\tilde{\chi}^2_N \ll 1$ is often due to a mistake on the part of the experimenter: either the data point measurements were not independent of each other, or the standard deviation estimates σ_i are inaccurate (too large). We'll have more to say about this issue later in the text.

Comparing competing theoretical predictions

Again consider N experimental results $y_i \pm \sigma_i$, and assume that each result has been perturbed by Gaussian noise from the data points' common mean μ. Given an independent theoretical prediction Y (which may not necessarily coincide with μ), then what would be the expected mean squared deviation of an experimental result from Y if $\mu \neq Y$? In this case:

$$\begin{aligned} \mathrm{E}[(y_i - Y)^2] &= \mathrm{E}[(y_i - \mu + \mu - Y)^2] \\ &= \mathrm{E}[(y_i - \mu)^2] + (Y-\mu)^2 - 2(Y-\mu)\,\cancel{\mathrm{E}[(y_i - \mu)]} \\ &= \sigma_i^2 + (Y-\mu)^2 > \sigma_i^2 \end{aligned}$$

3.6

$$\therefore \quad \mathrm{E}\left[\frac{(y_i - Y)^2}{\sigma_i^2}\right] = \mathrm{E}\left[\frac{(y_i-\mu)^2}{\sigma_i^2}\right] + \frac{(Y-\mu)^2}{\sigma_i^2} = 1 + \frac{(Y-\mu)^2}{\sigma_i^2}$$

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - Y)^2}{\sigma_i^2}\right] - 1 = (Y-\mu)^2\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right)$$

Equations (3.6) show that if the theoretical value Y actually differed from the common mean of the y_i distributions, then the weighted mean squared deviation of the data from Y calculated using the $\tilde{\chi}^2$ expression in (3.5) is expected to be greater than 1, the difference being proportional to $(Y-\mu)^2$. In this case the actual distribution of the weighted sum of the $(y_i - Y)^2$ would be described by a generalized form of chi-squared called the *noncentral chi-squared distribution*.

If there are two or more alternative theories which attempt to predict the result of our experiment, e.g.: $\mu = X$, $\mu = Z$, or $\mu = W$, then we may choose between these hypotheses by calculating the weighted mean squared deviation of the data from each hypothetical value using the $\tilde{\chi}^2$ expression in (3.5). The alternative which yields the smallest value would be that which is most consistent with our experiment's result, in the sense that the hypothetical

value with the smallest difference from the actual data distributions' mean μ would be expected to on average have the smallest $\tilde{\chi}^2$ value calculated using (3.5). We may use each calculated value as a quantitative score (with a smaller score being better) to rate the relative success of each of the various alternative hypotheses. In fact, from the discussion concerning the derivation (3.2), we also know that each score would be proportional to the negative of the logarithm of the likelihood of each alternative. Picking the minimum score then corresponds to using the maximum likelihood method for choosing among the alternative theories.

Note, however, that if the predictions Y, Z, and W are not very different, then our experiment must be precise enough to clearly distinguish between them. This can be accomplished by reducing the noise (thus decreasing the σ_i's) or by collecting more data (increasing N). Since the width of the $\tilde{\chi}^2$ distribution decreases proportional to $\sqrt{1/N}$ it takes, for example, 100 times as many data points to increase an experiment's precision by a factor of 10. Working to reduce the noise may turn out to be a quicker, more cost-effective solution.

Testing the consistency of a set of measurements

In the earlier example of calculating the weighted mean of a set of N independent experimental results $y_i \pm \sigma_i$, the assumption was made that all values were samples of distributions with a common mean μ. The maximum likelihood method was then used to determine an estimate $\bar{y}$ for μ which is most consistent with the measurements, resulting in equations (3.3) on page 33 for $\bar{y} \pm \sigma_{\bar{y}}$. Now one may question whether this assumption of a common mean for all the measurements is warranted. A reduced chi-squared calculation using $Y = \bar{y}$ in (3.5) may then be performed to test this assumption.

Since the estimate $\bar{y}$ is calculated from the set of measurements $y_i \pm \sigma_i$, the number of degrees of freedom in the resulting χ^2 calculation is $\nu = N-1$, as will be explored in the next chapter. Consequently, the relevant reduced chi-squared is given by $\tilde{\chi}^2_{N-1} = \chi^2/(N-1)$. Given the expected $\tilde{\chi}^2$ distribution if the measurements' distributions share a common mean μ, we would expect that $\tilde{\chi}^2 - 1 < \sqrt{8/(N-1)}$ more than 95% of the time. Thus, if $\tilde{\chi}^2 - 1$ exceeds this value, then it is likely that one or more of the measurements might be samples of distributions which do not share a mean with the others. A later chapter will further explore this topic.

A common situation is to test the consistency of a single pair of results, $x \pm \sigma_x$ and $y \pm \sigma_y$. Using (3.3) and (3.5) to calculate $\tilde{\chi}^2$, realizing that for this case $\nu = N-1 = 1$, so $\tilde{\chi}^2 = \chi^2$,

3.7

$$\tilde{\chi}^2 = \chi^2 = \frac{(x-y)^2}{\sigma_x^2 + \sigma_y^2}$$

With only one degree of freedom, we would expect that this $\tilde{\chi}^2 < 4 \approx 1 + \sqrt{8}$ over 95% of the time. In fact, for Gaussian x and y with a common mean μ and standard deviations σ_x and σ_y,

respectively, then the difference $z = x - y$ is Gaussian with zero mean and standard deviation:

$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2}$$

We would then expect that 95.5% of the time a sample of z would lie within $2\sigma_z$ of its mean of zero. We see that the $\tilde{\chi}^2$ calculation in (3.7) is just z^2/σ_z^2, so one test that two experimental results are consistent (may agree to within the level of their uncertainties) is:

3.8 $$|x - y| < 2\sqrt{\sigma_x^2 + \sigma_y^2}$$ **Consistency check of 2 measurements**

Testing for normally-distributed data scatter

Our final example of the use of a χ^2 calculation is to test whether the scatter in a set of measurements is consistent with a Gaussian distribution, so that the sample mean and its uncertainty can be expected to provide a maximum likelihood estimate of the underlying distribution's mean. There are actually many such tests, but the one we present here is a version of *Pearson's chi-squared test.*[*] Consider again the point estimation example from the last chapter, whose data are repeated in Figure 3-2 below. In the figure's right-hand graph a histogram of the relative frequencies of the various counts/channel values are compared to a Gaussian with mean and standard deviation given by the data set sample mean and sample standard deviation. How confident may we be that the data are actually consistent with samples drawn from a normal distribution?

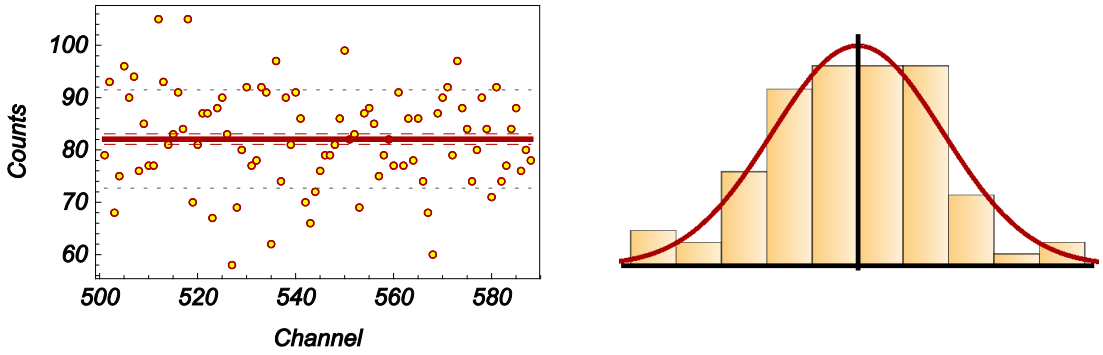


Figure 3-2: Gamma-ray detector event counts vs. energy channel number (both channel numbers and counts are integers for each data point). The right-hand plot is a histogram of counts/channel along with a Gaussian representing the sample data mean and the sample standard deviation.

The idea behind our simple test to answer this question goes like this: given the sample mean $\bar{y}$ and sample variance s^2 of our N data points, form the normalized sample set $(y_i - \bar{y})/s$, which should have mean 0 and variance approximately equal to 1. What is the

[*] The same English mathematician Karl Pearson (1857–1936) cited in Chapter 1.

probability that a set of N random samples drawn from a Gaussian distribution with mean 0 and variance 1 would be distributed at least as unevenly as the normalized data set? If the probability is not too small, then we may be satisfied that the actual data can be modeled as drawn from a Gaussian distribution. We use the χ^2 distribution to calculate this probability as follows: partition the proposed Gaussian distribution into Q *quantiles*: Q intervals each with a total integrated probability of $1/Q$. We find the values of y which delimit the quantiles by numerically integrating the probability expression (1.1) with the Gaussian PDF (1.16). For example, to find its *first quartile* ($Q = 4$, with Gaussian $\mu = 0$, $\sigma^2 = 1$):

$$1/4 = \int_{-\infty}^{y} p(\xi)\,d\xi \quad \rightarrow \quad y = -\sqrt{2}\,\mathrm{erfc}^{-1}(1/2) = -0.67448975\ldots$$

The *complementary error function* is defined as $\mathrm{erfc}(y) \equiv (2/\pi)\int_y^\infty \exp(-t)^2 dt$. The points $\{-0.674\ldots,\ 0,\ +0.674\ldots\}$ divide the Gaussian distribution into quartiles, so that a random sample drawn from it has probability $1/4$ for being included in any particular one of them. The general formula for the boundary between two quantiles is given on page vi.

The N data points are expected to be evenly spread among the Q quantiles of the parent distribution. The actual numbers of points found in a quantile will, of course, vary randomly, distributed according to a *binomial distribution* with N trials and success probability $1/Q$. Each quantile should then contain an expected number $(1/Q)N$ of the data points with variance $(1/Q)(1-1/Q)N$.* Now we have collected all of the ingredients needed to perform a χ^2 analysis of the actual distribution of the data points over the Q quantiles of the candidate Gaussian:

3.9 $$\chi^2_{(Q-3)} = \sum_{i=1}^{Q} \frac{(N_i - N/Q)^2}{(1/Q)(1-1/Q)N} = \frac{1}{(Q-1)N}\sum_{i=1}^{Q}(QN_i - N)^2$$ **Quantile χ^2 test**

The sum is over the quantiles, with N_i as the count in the ith quantile. $Q-3$ degrees of freedom for this χ^2: Q quantiles, but N, $\bar{y}$, and s^2 (used to define or normalize the Gaussian distribution) are derived from the data, providing 3 constraints.

With $Q = 4$ for the example shown in Figure 3-2, $N = 88$ and the quartile counts are 19, 26, 21, and 22. Equation (3.9) then gives $\chi^2 = 1.576$. With 1 degree of freedom $\chi^2 > 1.576$ has the fairly high probability of 21% (this probability is called the *p-value* of the test), so one may reasonably conclude that the actual data points are not inconsistent with samples of a Gaussian. A common rejection criterion is that if the test's p-value is 5% or less, then it is very unlikely that the data are consistent with the Gaussian distribution. With one degree of

* Many references would instead use the *Poisson distribution*, discussed in Chapter 6, to estimate each quartile's variance. The Poisson distribution approximates the binomial distribution only in the limit that the success probability becomes small. The Poisson distribution's variance is equal to its mean, which would be N/Q. For the quartile case, this is a 33% error over that of the binomial distribution. In the limit that Q is large, so that the probability of a sample falling into any particular division becomes small, the binomial distribution is well-approximated by the Poisson distribution.

freedom (using quartiles), this would be the case if $\chi^2 > 3.84$. In many cases, a particular choice for Q may result in a small p-value, whereas for another choice the p-value is much higher. It is therefore probably a good idea to perform this test using at least a few different values for Q. Table V on page vii provides quantile boundaries and limiting χ^2 values for various numbers of quantiles.

INTERPRETING THE UNCERTAINTY OF A RESULT

The standard deviation of the likelihood distribution determines the uncertainty in our experimentally measured value of a physical constant. Once we have decided that the physical constant has some uncertain value $Y \pm \sigma_Y$, the question remains: how exactly should one interpret the uncertainty σ_Y? Yet again consider our example of the experimental determination of the electron charge/mass ratio, q_e/m_e. Is σ_Y the standard deviation of the probability distribution of possible q_e/m_e values around our estimate Y? Is there such a thing as a "probability distribution" of the value of a fundamental *constant* of nature?

As of this writing, the NIST (National Institute of Standards and Technology) website* gives the experimentally-determined value for q_e/m_e to be:

$$-(1.758\ 820\ 024 \pm .000\ 000\ 011) \times 10^{11} \text{ coulomb/kilogram}$$

The NIST website describes their quoted uncertainty as the estimated standard deviation of the determination of $q_e/m_e = Y$. It goes on to define the meaning of this phrase (their uncertainty $u(y) \equiv \sigma_Y$ in our notation):

> *If the probability distribution characterized by the measurement result … is approximately normal (Gaussian), … then the interval … implies that it is believed with an approximate level of confidence of 68% that Y is greater than or equal to $y - u(y)$, and is less than or equal to $y + u(y)$, which is commonly written as $Y = y \pm u(y)$.*†

The problem with this definition, at least for a statement about the value of a fundamental constant of nature, is that at best it may be poorly worded and potentially ambiguous. The actual value of q_e/m_e either lies somewhere within the stated interval (and therefore has probability $P[\,Y = y \pm u(y)] = 1\,)$, or it doesn't (with probability $P[\,Y = y \pm u(y)] = 0\,)$, so what is the correct interpretation of "level of confidence of 68%?"

A more precise interpretation of the stated uncertainty addresses the experimental methods used to measure of the constant: one should expect that approximately 68% of repeated, similar experiments should yield a value of q_e/m_e within $\pm\sigma_Y$ of the true, but unknown, value of the constant.

* http://physics.nist.gov/cgi-bin/cuu/Value?esme|search_for=atomnuc

† http://physics.nist.gov/cgi-bin/cuu/Info/Constants/definitions.html

This interpretation is an inference about *the expected scatter in the results of similar experiments,* and not about the probability that the actual value of q_e/m_e falls within some range. It addresses the standard deviation of the likelihood function used to choose that value for q_e/m_e considered to be most consistent with available experimental results.

Given the methods used to determine q_e/m_e, the resultant scatter about the true value due to various sources of measurement error is expected to be characterized by a standard deviation of σ_Y. If the distribution of the errors is Gaussian about a mean equal to q_e/m_e, then we would estimate that approximately 68% of a large set of similar experimental results could be expected to lie within $\pm\sigma_Y$ of the actual value of q_e/m_e. Thus the likelihood PDF derived from the available experimental results would also be Gaussian with maximum (and mean) at $q_e/m_e = Y$ and with standard deviation σ_Y. The area under the likelihood PDF within $\pm\sigma_Y$ of the experimentally-determined value would then also be 68%. That is the actual meaning of a *68% confidence interval* about the experimentally-determined value; $\pm 2\sigma_Y$ would represent a 95% confidence interval.

Chapter 4
Curve fitting and optimizing free parameter values

The last chapter demonstrated that for data whose random errors are independent and normally-distributed, a maximum likelihood estimate of the distribution mean may be found using *chi-squared minimization*. That result may be generalized to more complicated and important problems of experimental data analysis: selecting that theory which best describes the functional relationship between measureable quantities, or optimizing a given theory's free parameter values to best match an experiment's results. This chapter addresses some of the practical mathematical details of this approach and gives some pointers on how to properly set up the optimization process and interpret its results.

CHI-SQUARED MINIMIZATION

Functions with several parameters; the degrees of freedom of χ^2

Assume that a theory predicts a certain algebraic functional relationship between two experimentally accessible quantities: $y = f(x)$. Assume further that the theoretical relationship between x and y depends on M numerical parameters requiring experimental determination (they are free parameters of the theory). We label these parameters as $a_1, a_2, \ldots, a_M$, and restate the theory as $y = f(x; a_1, a_2, \ldots, a_M)$, making the dependence on these parameters explicit. Note that the values of the parameters are not in principle to be considered to be functions of the variables x and y, but rather to define the numerical relationship $y = f(x)$. The M parameters must be *independent*: we can't replace them with a smaller set of parameters and still maintain the same relation $y = f(x)$. Here are a couple of examples from undergraduate laboratory experiments, each expressed as functions containing two free parameters:

The rate of decay of a radioactive sample decreases exponentially with time.	$r(t) = r_0 \exp(-t/\tau)$	r_0 : decay rate at time 0 τ : mean lifetime
The square of the magnetic field required to bend the trajectory of a high-speed charged particle along a fixed-radius path varies quadratically with the particle's kinetic energy.	$B^2(T) = (qR)^{-2}\left(T^2 + 2mT\right)$	qR : charge-path radius product m : particle rest energy

Assume that an experiment is designed to test the relation $y = f(x; a_1, a_2, \ldots, a_M)$, and that the experiment's resulting data consists of N ordered pairs of measured values (x_i, y_i) collected as the experimental conditions are varied. Assume further that each pair has an associated, experimentally-determined standard deviation estimate σ_i due to random,

independent, Gaussian noise in the measurements (how to determine σ_i is discussed later in this chapter). Construct a χ^2 variate which sums the differences between the measured data and the theory's predictions, each term weighted by its variance:

4.1

$$\chi^2_{(N-M)} = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_M) \right)^2}{\sigma_i^2}$$

χ^2 for $y = f(x)$

The individual data point errors must be independent!

The data point variances σ_i^2 in the χ^2 expression (4.1) represent the expected scatter (noise) in the various measured data points (x_i, y_i). ***It is quite important that the errors in the data point values are independent***, so that the covariances between different data points all vanish (otherwise, the provided expression for χ^2 is incorrect). ***Errors that may have been introduced into the data points which are* not *independent*** (such as a calibration error in an instrument used to collect the data) ***must not be included in the σ_i^2 estimates of expression* (4.1).** These additional sources of *correlated errors* are *systematic*, and will be dealt with in a completely different manner (see Chapter 5: *Dealing with systematic errors*).

The number of *degrees of freedom* of the χ^2 variate in (4.1) is $N-M$ and not simply N as was often the case in the previous chapter. The chi-squared minimization process to optimize the M parameter values uses the measured data point values (x_i, y_i). The resulting expressions for the optimized parameter values provide M *equations of constraint*, reducing the number of degrees of freedom. Alternatively, because we use the actual acquired data values while minimizing the sum (4.1), the residuals $y_i - f(x_i; a_1, a_2, \ldots, a_M)$ can be made a bit smaller on average than their corresponding σ_i values. It turns out that for normally-distributed residuals the sum then has an expected value of $N-M$, rather than N. This implies that the χ^2 variate resulting from the minimization of the sum has $N-M$ degrees of freedom.* The corresponding reduced chi-squared variate $\tilde{\chi}^2$ is then $\chi^2/(N-M)$:

4.2

$$\tilde{\chi}^2_{(N-M)} = \frac{1}{N-M} \sum_{i=1}^{N} \frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_M) \right)^2}{\sigma_i^2}$$

Reduced χ^2 for $y = f(x)$

* Note the connection between (4.1) and our previous solution for the specific case of estimating a distribution's mean and variance from a set of N data points given in expressions (2.2) on page 18. In that case the model for the y values was simple: $y = f(x) =$ constant $= \mu$. The single free parameter in the function was μ, and the number of degrees of freedom was therefore $N-1$. All points were assumed to be samples of a single distribution, so all the σ_i were equal to the distribution's unknown standard deviation σ. This approach is discussed further in the next two sections.

The set of values for the parameters which minimizes (4.1) and (4.2) is called the *least-squares solution* for the parameters. If the N residuals $y_i - f(x_i; a_1, a_2, \ldots, a_M)$ are normally distributed, then this set of values will be the maximum likelihood solution for the parameters $a_1, a_2, \ldots, a_M$.

Linear regression

If the function $f(x; a_1, a_2, \ldots, a_M)$ is a *linear combination* of the M parameters a_j, that is, it is of the form shown in equation (4.3), then an algebraic expression for the minimum of the sum (4.1) for χ^2 can be found (each of the functions $g_j(x)$ in (4.3) must be independent of all of the M parameters).

4.3 $$f(x; a_1, a_2, \ldots, a_M) = \sum_{j=1}^{M} a_j \, g_j(x)$$ **Linear combination of the a_j**

In this case the optimization of the M parameter values using (4.1) is a form of *linear least-squares* minimization or *linear regression*, and its solution is an exercise in linear algebra. Assuming that $f(x; a_1, a_2, \ldots, a_M)$ has the form (4.3), setting the partial derivative of the expression for χ^2 in (4.1) with respect to one of the M parameters a_j to 0 results in the equation:*

Linear regression system of equations

4.4 $$\sum_{k=1}^{M} a_k \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2} g_j(x_i) g_k(x_i) \right] = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} g_j(x_i) \, y_i$$

There will be M such equations (for $j = 1 \ldots M$), and the left-hand side of each is a linear combination of the M parameters, whereas the right-hand side is independent of the parameters. This system of M linear equations in the M unknown parameters provides a unique solution for the set of parameter values which minimizes χ^2, as long as the parameters are independent (the determinant of the coefficients of the a_k doesn't vanish).

The simplest example of a linear regression is the case wherein the theory predicts that the y_i data should all equal the same, but unknown, constant a, independent of the x_i values: $y = f(x) = a$. Noise in the measurements, of course, spoils this perfect scenario, and the uncertainties in the various y_i measurements are given by the σ_i. In other words, the theory

* In deriving (4.4) we assume that the data point variances σ_i^2 are constants (independent of the parameter values). This will be true if they are associated with measurements of the y_i values only; the corresponding x_i values are considered to be equivalent to experimental "control settings" which have no associated measurement uncertainty. This assumption leads to what is commonly called the *ordinary* weighted least-squares method. We relax this restriction later in the chapter.

predicts that the distributions of the y_i share the common mean a. Applying the single linear regression equation for a given by (4.4) with the single function $g(x) = 1$,

$$a\left[\sum_{i=1}^{N}\frac{1}{\sigma_i^2}g(x_i)g(x_i)\right] = a\sum_{i=1}^{N}\frac{1}{\sigma_i^2} = \sum_{i=1}^{N}\frac{1}{\sigma_i^2}g(x_i)\,y_i = \sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}$$

$$\text{or, simply:}\quad a = \left[\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}\right]\div\left[\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\right]$$

This is the formula (3.3) for the weighted mean of a set of N experimental results derived in Chapter 3.

Next consider the only slightly more complicated example of a two-parameter, linear model for the set of data: $y = f(x) = a_1 + a_2x$. From (4.3) we have $g_1(x) = 1$ and $g_2(x) = x$. To keep things nice and simple, assume that the data point uncertainties are all equal: $\sigma_i^2 = \sigma^2$. Thus the σ_i^2 divide out of the linear regression system of equations in (4.4), and the pair of coupled equations for a_1 and a_2 may be written in matrix form and solved, giving (4.5):

$$\begin{pmatrix} N & \Sigma x_i \\ \Sigma x_i & \Sigma {x_i}^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma(x_i y_i) \end{pmatrix}$$

4.5 with the solution:

$$a_1 = \frac{\Sigma {x_i}^2 \Sigma y_i - \Sigma x_i \Sigma(x_i y_i)}{N\Sigma {x_i}^2 - (\Sigma x_i)^2};\quad a_2 = \frac{N\Sigma(x_i y_i) - \Sigma x_i \Sigma y_i}{N\Sigma {x_i}^2 - (\Sigma x_i)^2}$$

How the uncertainties in the parameter values are determined is addressed in a later section.

Estimating the uncertainty from experimental data; unweighted least squares

If the uncertainties σ_i of the data points are all the same, $\sigma_i = \sigma$, then the common variance may be factored out of the χ^2 expressions (4.1) and (4.2). The resulting problem is that of the minimization of a simple sum of the squared residuals, $\Sigma(y_i - f(x_i))^2$, known as the *unweighted least squares* problem. This was the case for the simple linear regression example in the last section leading to the solution (4.5). Because the uncertainty σ can be factored out of the sum, the χ^2 minimization and the resulting optimized parameter values are independent of the value of σ.

Unweighted least squares problems most often arise in situations wherein the variance of the noise-induced scatter in the individual y_i measurements is unknown, but is assumed to be the same for all of them. At the same time it is assumed that any noise in the x_i

measurements is completely negligible.* If the optimized model function $f(x; a_1, a_2, \ldots, a_M)$ is consistent with the data, and the noise in the y_i is Gaussian, then the residuals $y_i - f(x_i)$ will be normally distributed with mean 0 and variance determined by σ. Thus the optimization can not only provide maximum likelihood estimates of the parameter values, but also provides an unbiased estimate of σ.

Under the circumstances outlined in the previous paragraph, the expected value of χ^2 is $N-M$, so the expected value of σ^2 must be:

4.6 $$s^2 \equiv \mathrm{E}[\sigma^2] = \frac{1}{N-M}\sum_{i=1}^{N}\left(y_i - f(x_i; a_1, a_2, \ldots, a_M)\right)^2$$ **Estimated variance of the y_i**

When evaluating (4.6), the parameters take on their optimized values. This equation provides a generalization of the sample variance defined by expression (2.1) on page 22. We can estimate the uncertainty in this estimation of σ by generalizing (2.3):

4.7 $$\frac{\sigma_s}{s} \sim \frac{1}{\sqrt{2(N-M)}}$$ **Accuracy of the uncertainty**

To reiterate, the basic idea behind the estimations in (4.6) and (4.7) is that if the model is correct and the noise is Gaussian, then the reduced chi-squared $\tilde{\chi}^2_{N-M}$ is expected to be 1. Assuming that it is in fact equal to 1 and solving for the corresponding variance σ^2 then gives these expressions. The same caveat applies to this uncertainty estimate as did that of (2.3), namely that the substitution of s^2 for $\mathrm{E}[\sigma^2]$ may be quite inaccurate if the number of degrees of freedom is small. Using uncertainty propagation, we know that the fractional uncertainty in the variance σ^2 will be twice that of (4.7).

Nonlinear regression; the Hessian matrix of χ^2

Often the theoretical model function $f(x; a_1, a_2, \ldots, a_M)$ in (4.1) is not of the form (4.3), making it is a nonlinear function of one or more of the parameters. For example, consider $f(x) = a_1\exp(-a_2 x) + a_3$. In most such cases an analytic solution for the parameters cannot be found, and the χ^2 minimum must be determined numerically. If χ^2 is a differentiable function of each of the M parameters a_j in a region about its minimum, then it must be the case that all of the M first partial derivatives $(\partial/\partial a_j)\chi^2$ vanish at that minimum. Now, $f(x; a_1, a_2, \ldots, a_M)$ is a real-valued function of its real-valued argument and parameters. If it is twice-differentiable at the χ^2 minimum, then the *Hessian matrix*† **H** of χ^2 will be real-

* An *ordinary least squares* problem.

† The *Hessian matrix* **H** of a scalar function $g(a_1, a_2, \ldots, a_M)$ is the square matrix formed from all of its second partial derivatives: $H_{jk} = (\partial^2/\partial a_j \partial a_k)\, g$. It is named for the 19th century Prussian mathematician Ludwig Hesse.

valued and symmetric at that minimum. Therefore **H** will have only real-valued *eigenvalues*. At a minimum (local or global) of χ^2, **H** must be *positive definite*, that is, its eigenvalues must all be positive.* These conditions on the first and second partial derivatives of χ^2 with respect to the parameters characterize the *nonlinear least-squares* minimization problem and form the basis of algorithms to solve it. Additionally, the Hessian matrix evaluated at the χ^2 minimum is closely related to the uncertainties in the optimized parameter values, as we investigate later in this chapter.

It is not unusual for there to be several local minima of χ^2, as well as local maxima and saddle points. The solver must be prepared to evaluate the Hessian matrix **H** at each candidate critical point (a point where the partial derivatives $(\partial/\partial a_j)\chi^2$ all vanish). Mutual independence of the parameters implies that the determinant of the Hessian matrix, det(**H**), will not vanish at a critical point. The value of the determinant of **H** is equal to the product of its eigenvalues, and Tr(**H**), the *trace* of **H** (the sum of its diagonal elements), equals the sum of its eigenvalues. If either det(**H**) < 0 or Tr(**H**) < 0, then at least one of the eigenvalues must be negative, implying that the critical point cannot be a minimum. If both det(**H**) > 0 and Tr(**H**) > 0, however, the eigenvalues must still be checked to ensure that they are all positive. The solver should also be careful not to select a local minimum rather than the global minimum of χ^2.

DETERMINING σ_i^2 THE FROM *X* AND *Y* UNCERTAINTIES

From (4.1) it should be clear that each of the N residuals in the χ^2 sum is weighted by its variance, so that more uncertain (noisy) data points are weighted less than those with smaller uncertainties. Thus χ^2 minimization will tend to favor points with small uncertainties when optimizing the parameter values in the theoretical model function $y = f(x; a_1, a_2, \ldots, a_M)$. It is important that these uncertainties be properly calculated so that the data points are properly weighted and that the resulting χ^2 value is accurately determined.

The uncertainty σ_i in each term of the χ^2 sum represents *the uncertainty in the corresponding residual* $y_i - f(x_i)$ introduced by noise (lack of repeatability) in the experimental measurement of (x_i, y_i). The random noise in these measured quantities, characterized by their uncertainties σ_{x_i} and σ_{y_i} and their covariances $\sigma^2_{x_i y_i}$, propagate through the residual expression to produce the uncertainty σ_i. We can use the naïve error propagation formula (2.8) on page 28 to calculate the residual's variance σ_i^2 from the variances in x and y:

* If **H** is *positive semidefinite*, i.e. has at least one vanishing eigenvalue, so that det(**H**) = 0, then the critical point may be a minimum. In this case, however, the parameters are not all independent, at least in a small neighborhood of parameter space containing that critical point, so one or more of them can be eliminated without changing the function $f(x)$. This should be accomplished, and the minimization repeated.

4.8
$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2 - 2\left(\frac{df}{dx}\right)\sigma_{x_i y_i}^2$$

The derivatives of $f(x)$ in (4.8) should be evaluated during χ^2 minimization at $x = x_i$ using the current estimates of the parameter values $a_1, a_2, \ldots, a_M$. We shall spend some time analyzing the implications of this expression on our understanding of the χ^2 minimization process.

Ordinary least-squares

In many experiments, only one of the two numerical values of a data point pair (x_i, y_i) is actually measured; the other may be a value set on some instrument control or otherwise established as part of the experiment's procedure. In this case, only the measured value will have an uncertainty associated with it, because observable scatter in the data from multiple samples will only be present for a measured quantity. Of course, there may be an error in the value set for the unmeasured quantity, but when minimizing χ^2 we want the uncertainties σ_i to reflect random, independent fluctuations in measured quantities.

If the only measured member of a data point pair is y_i, then $\sigma_i^2 = \sigma_{y_i}^2$. This variance is independent of changes to the function $f(x_i)$ as the χ^2 minimization proceeds, so the weight of each residual term in the χ^2 sum remains constant. This situation results in what is called the *ordinary least-squares* problem, and very efficient and accurate numerical algorithms are available to solve it. It can be shown that the results of this procedure will provide minimum-variance, unbiased estimates of the function parameters $a_1, a_2, \ldots, a_M$, at least if the noise fluctuations in the y_i are Gaussian. This is the best-case scenario for the maximum likelihood analysis of experimental data.

Uncertainties in the x_i: total least-squares

The theoretical physical relationship between the experimentally accessible quantities x and y is represented by the model function $y = f(x)$, but this expression doesn't necessarily require that x be an "independent" experimenter control input which then generates the "dependent" response y. Rather the function just expresses a numerical relationship between the quantities x and y, without assigning "cause" and "effect." If the only measured value of a data point pair (x_i, y_i) is x_i, resulting in the associated uncertainty $\sigma_{x_i}^2$, then the wisest course of action is to swap the roles of x and y in the residual calculations: $x_i - f^{-1}(y_i)$. The inverse function $f^{-1}(y; b_1, b_2, \ldots, b_M)$ will involve a set of M new parameters $b_1, b_2, \ldots, b_M$ which will generally not be related in a simple fashion algebraically to the ones in the original function $f(x; a_1, a_2, \ldots, a_M)$. The advantage of this approach, however, is that the variances of the terms in the χ^2 sum remain constant throughout the minimization, leading again to an ordinary least-squares problem.

If some other overriding consideration requires that the χ^2 residuals remain in the original form $y_i - f(x_i; a_1, a_2, \ldots, a_M)$, then by (4.8) the variances become $\sigma_i^2 = (df/dx)^2 \sigma_{x_i}^2$, and these are now functions of the parameters. As the χ^2 minimization proceeds and the estimates of the parameter values are adjusted, not only do the term weights vary, but the term partial derivatives with respect to the parameters must include factors involving $(\partial/\partial a_j)\sigma_{x_i}^2$. This added complication can be quite serious, and leads to the use of so-called *total least-squares* algorithms.*

There are two important consequences which must be considered when the total least-squares approach is chosen:

(1) Even if the model function $y = f(x; a_1, a_2, \ldots, a_M)$ is linear in the M parameters as in (4.3), χ^2 minimization will require an iterative, nonlinear algorithm rather than a closed-form, algebraic solution characteristic of linear least-squares problems.

(2) Algorithms are not guaranteed to find minimum-variance, unbiased estimates of the function parameters $a_1, a_2, \ldots, a_M$, even when the noise is Gaussian. In fact, the resulting parameter estimates are often biased, and their uncertainties are often underestimated.

Uncertainties in both x_i and y_i

If both of the data point values in (x_i, y_i) are measured, then noise will add fluctuations to each of them, resulting in nonzero estimates for both $\sigma_{x_i}^2$ and $\sigma_{y_i}^2$. Thus the full expression (4.8) must be used to determine σ_i^2, and a total least-squares algorithm will be needed to minimize χ^2 (with all of its disadvantages, as outlined above). An extra, very important complication arises in this case, however, because of possible correlation in the fluctuations of x and y, leading to a nonzero covariance $\sigma_{x_i y_i}^2$.

To illustrate the problem caused by a nonzero covariance, consider this experimental situation: x is a control variable set to various values as the experiment proceeds; the physics, as modeled by the equation $y = f(x)$, predicts that the observed value for y will vary in response to changes in x (i.e. changes in x result in changes in y, as predicted by the model $f(x) \rightarrow y$). Of course, as with any experiment, random variations introduce noise into the experiment, so fluctuations in the values of measured quantities are evident as multiple samples are collected. Assume that as the control variable x is adjusted, its value is then measured along with measurements of y, generating the data points (x_i, y_i). Since both x and y are measured, each will show fluctuations and have associated variances.

Now the problem of correlation rears its ugly head: is the observed noise in x introduced *solely through its measurement*, or does it indicate that the underlying control value x is *actually fluctuating*? If x really is varying, then the physics $f(x) \rightarrow y$ would imply that y

* Other approaches may also be used, such as various *models of measurement error* methods.

should also vary in response. To help understand the situation, we may model the noise introduced into the experiment's data as shown in Figure 4-1.

If noise is independently introduced into each measurement chain only (leaving the underlying control value x alone), then the fluctuations observed during measurements of x and y will originate in independent sources and therefore be uncorrelated. In this case, $\sigma_i^2 = \sigma_{y_i}^2 + (df/dx)^2\sigma_{x_i}^2$, and a total least-squares algorithm is appropriate.

If, on the other hand, a significant fraction of the variance observed during measurements of x indicate fluctuations of the underlying control value input to the system, then these fluctuations will generate corresponding fluctuations in y. To be more specific, assume that the only significant noise introduced to the determination of x in the experiment illustrated in Figure 4-1 is at the point c, and that the noise introduced at point a is negligible. Assume further that the function $f(x)$ correctly describes how y responds to the input x. Now these fluctuations in x, carefully measured and correctly described by its observed variance σ_x^2, cause corresponding fluctuations in y, which add to the independent fluctuations introduced at point b. The overall fluctuation in y is also correctly measured and described by σ_y^2. Because of the noise injected at c, the instantaneous fluctuations in x and y away from their respective means are strongly correlated. Let the instantaneous noise fluctuations at c and b be designated δ_c and δ_b. If x_i and y_i data are measured nearly simultaneously, so that any particular noise fluctuation in the value of x affects both the observed values of x_i and y_i for any particular data point (x_i, y_i), then (to lowest order, with independent errors δ_c and δ_b), we have (see next page):

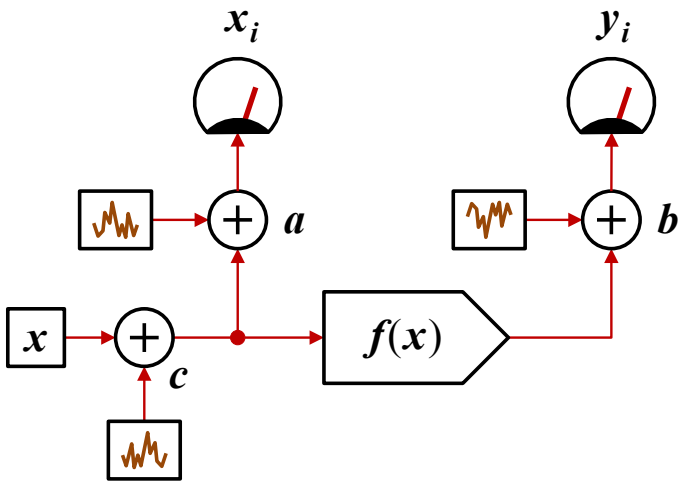


Figure 4-1: Noise added to measurements of *x* and *y* may or may not lead to correlated errors, depending on where the noise sources are located in the experiment's causal chain. In this example, the physics requires that changes in *x* result in changes in *y* through some functional relationship *f*(*x*) → *y*. Noise added at points *a* and *b* independently affect the measurements x_i and y_i, but noise added at *c* affects both measurements in a correlated manner.

$$(y_i - \mu_y) = \delta_b + \left(\frac{df}{dx}\right)\delta_c$$

$$\sigma_{x_i}^2 = \overline{(x_i - \mu_x)^2} = \overline{\delta_c{}^2}$$

$$\sigma_{y_i}^2 = \overline{(y_i - \mu_y)^2} = \overline{\delta_b{}^2} + \left(\frac{df}{dx}\right)^2 \overline{\delta_c{}^2} = \overline{\delta_b{}^2} + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2$$

$$\sigma_{x_i y_i}^2 = \overline{(x_i - \mu_x)(y_i - \mu_y)} = \overline{\delta_c \delta_b + \left(\frac{df}{dx}\right)\delta_c{}^2} = \left(\frac{df}{dx}\right)\sigma_{x_i}^2$$

and

$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2 - 2\left(\frac{df}{dx}\right)\sigma_{x_i y_i}^2$$

$$\therefore \boxed{\sigma_i^2 = \sigma_{y_i}^2 - \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2}$$

Because the function $f(x)$ correctly describes the functional relationship between the values of x and y, at any instant $y_i = f(x_i) + \delta_b$, and the residual $y_i - f(x_i) = \delta_b$. Thus the above calculation correctly results in σ_i^2 equal to the variance of the residual, which is just the extra noise injected at point b during the measurement of y; in this case the noise in x injected at c does not contribute to errors in the residuals and must be subtracted from the observed σ_y^2. Already we see that correct calculation of the data point variance can be subtle.

If the measurements x_i and y_i are not performed nearly simultaneously, then the noise signal δ_c injected at point c in Figure 4-1 will have time to change to some different value δ'_c between the two measurements. If the interval between measurements of x_i and y_i is long enough, the instantaneous fluctuations δ_c (affecting x_i) and δ'_c (affecting y_i) will be uncorrelated. Consequently,

$$(y_i - \mu_y) = \delta_b + \left(\frac{df}{dx}\right)\delta_c'$$

$$\sigma_{x_i}^2 = \overline{(x_i - \mu_x)^2} = \overline{\delta_c{}^2} = \overline{(\delta_c')^2}$$

$$\sigma_{y_i}^2 = \overline{(y_i - \mu_y)^2} = \overline{\delta_b{}^2} + \left(\frac{df}{dx}\right)^2 \overline{(\delta_c')^2} = \overline{\delta_b{}^2} + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2$$

$$\sigma_{x_i y_i}^2 = \overline{(x_i - \mu_x)(y_i - \mu_y)} = \overline{\delta_c \delta_b + \left(\frac{df}{dx}\right)\delta_c \delta_c'} = 0$$

$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2 - 2\left(\frac{df}{dx}\right)\sigma_{x_i y_i}^2$$

and

$$\boxed{\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2}$$

In this case, the full contributions of the observed x and y variances must be included in σ_i^2.

If the x value is not measured, but a nominal value is used for x_i, then the problem becomes an ordinary least-squares minimization, and $\sigma_i^2 = \sigma_{y_i}^2$. Thus, depending on the scenario established by the experimental setup, the value to be used for σ_i^2 can be hard to determine. Equation (4.8) for the residual variance evaluates to:

4.9

$$\sigma_i^2 = \sigma_{y_i}^2 + \rho\left(\frac{df}{dx}\right)^2 \sigma_{x_i}^2; \quad -1 \le \rho \le 1$$

The value of the coefficient ρ of the $\sigma_{x_i}^2$ term in (4.9) can be positive or negative and must be calculated following a careful study of exactly how the x_i and y_i are measured. For most situations encountered in undergraduate physics experiments, the x and y values are not measured simultaneously, and their observed fluctuations are uncorrelated. In this case the appropriate calculation of σ_i^2 would be (4.9) with $\rho = 1$. This is the choice assumed by the *CurveFit* data analysis package when both x and y uncertainties are present.

Derived values are used for x and y

In some cases it may be convenient to calculate the quantities x and y used in χ^2 minimization from some other pair of actually measured quantities u and v, whose noise fluctuations are assumed to be independent. Thus $x_i = x(u_i, v_i)$ and $y_i = y(u_i, v_i)$, for functions $x(u, v)$ and $y(u, v)$. How is σ_i^2 then determined from the experimental uncertainties in u_i and v_i? For example, an experiment may measure the magnetic field strength $B \pm \sigma_B$ needed to focus high-energy electrons on a detector and the corresponding kinetic energy $T \pm \sigma_T$ of the focused electrons. The most convenient representation of the theory relating B and T may be $(B^2/T) = f(T)$, where $f(T)$ is a first-order polynomial function of T. The experimenter then chooses $x_i = T_i$ and $y_i = (B_i^2/T_i)$ as the data variables used in the least-squares procedure.

χ^2 calculated using (4.1) with the derived data pairs (x_i, y_i) and the model $y = f(x)$ should be the same as that calculated using the actual data pairs (u_i, v_i) and the corresponding model $v = g(u)$, where $g(u)$ is implicitly defined by $y = f(x)$: $y(u, g(u)) = f(x(u, g(u)))$. Without proof, it turns out that the χ^2 term for $y_i - f(x_i; a_1, a_2, \ldots, a_M)$ has σ_i^2 given by:

$$\sigma_i^2 = \left(\frac{\partial y}{\partial v} - \frac{df}{dx}\frac{\partial x}{\partial v}\right)^2 \sigma_{v_i}^2 + \left(\frac{\partial y}{\partial u} - \frac{df}{dx}\frac{\partial x}{\partial u}\right)^2 \sigma_{u_i}^2$$

Now we can form the correct expression for χ^2 in the case where x and y are derived from the measured variables u and v (whose noise errors are assumed to be independent). The derivatives are all evaluated at (x_i, y_i). Because σ_i^2 involves df/dx, a total least-squares algorithm must be used to minimize χ^2:

4.10

$$\chi^2_{(N-M)} = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_M) \right)^2}{\left(\frac{\partial y}{\partial v} - \frac{df}{dx}\frac{\partial x}{\partial v} \right)^2 \sigma_{v_i}^2 + \left(\frac{\partial y}{\partial u} - \frac{df}{dx}\frac{\partial x}{\partial u} \right)^2 \sigma_{u_i}^2}$$

x = x(u,v); y = y(u,v)

Returning to the example in which magnetic field strength B and electron kinetic energy T are measured, but for the analysis $(x_i, y_i) = (T_i, B_i^2/T_i)$, then the χ^2 terms become:

$$\frac{\left(y_i - f(x_i; a_1, a_2, \ldots, a_M) \right)^2}{4\frac{B_i^2}{T_i^2}\sigma_{B_i}^2 + \left(\frac{B_i^2}{T_i^2} + \frac{df}{dx} \right)^2 \sigma_{T_i}^2}$$

The expression for σ_i^2 may alternatively be put into the form (4.8) by propagating uncertainties using (2.8) and the functions $x(u, v)$ and $y(u, v)$ to determine $\sigma_{x_i}^2$, $\sigma_{y_i}^2$, and $\sigma_{x_i y_i}^2$. The results for these variances are:

$$\sigma_{x_i}^2 = \left(\frac{\partial x}{\partial v} \right)^2 \sigma_{v_i}^2 + \left(\frac{\partial x}{\partial u} \right)^2 \sigma_{u_i}^2 \; ; \quad \sigma_{y_i}^2 = \left(\frac{\partial y}{\partial v} \right)^2 \sigma_{v_i}^2 + \left(\frac{\partial y}{\partial u} \right)^2 \sigma_{u_i}^2$$

$$\sigma_{x_i y_i}^2 = \frac{\partial x}{\partial v}\frac{\partial y}{\partial v}\sigma_{v_i}^2 + \frac{\partial x}{\partial u}\frac{\partial y}{\partial u}\sigma_{u_i}^2$$

Clearly, since x and y both are functions of the same two measured quantities u and v, their covariance will generally be nonzero. Substituting the above expressions into (4.8) and simplifying will result in (4.10), as it must.

PARAMETER UNCERTAINTIES AND THE COVARIANCE MATRIX

If the errors in the measurements are Gaussian, then, as discussed previously, minimization of χ^2 in (4.1) provides the maximum likelihood solution for the values of the parameters a_1, a_2, …, a_M. To determine the expected uncertainties in the estimates of the parameter values, we consider the PDF of the experiment's likelihood function as the hypothesized parameter values are varied: $p_L(\langle\text{data set}\rangle|\langle\text{parameter values}\rangle)$. Consider first the case of a single parameter value: $\chi^2 = \chi^2_{(N-1)}(a)$. If the errors are Gaussian, then so will be the likelihood function, at least for a small range of values of the parameter a around its maximum likelihood value a_0. Therefore, the likelihood PDF will vary as:

4.11
$$p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a) \propto \exp\left[\frac{-(a-a_0)^2}{2\sigma_a^2}\right]$$

When $|a - a_0| = \sigma_a$, the exponent in (4.11) will equal $-1/2$, and

$$\frac{p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a_0 \pm \sigma_a)}{p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a_0)} = e^{-1/2}$$

But the likelihood PDF may also be determined from the independent data point values as

$$p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a) \propto \prod \exp\left[\frac{-(y_i - f(x_i, a))^2}{2\sigma_i^2}\right] = \exp\left[-\frac{1}{2}\chi^2(a)\right]$$

With $p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a_0 \pm \sigma_a) = e^{-1/2} p_{\mathrm{L}}(\langle \text{data set} \rangle \,|\, a_0)$, then $\chi^2(a_0 \pm \sigma_a) = \chi^2(a_0) + 1$.

Thus it must be the case that $\chi^2(a_0 \pm \sigma_a) = \chi^2(a_0) + 1$, and σ_a may be determined by equating it to the Δa required to increase χ^2 by 1 above its optimized, minimum value.

This method is also appropriate if there are several parameters $a_1, a_2, \ldots, a_M$. In this case, however, $\chi^2_{(N-M)}$ must be continually minimized with respect to all of the other $M-1$ parameters as the parameter whose σ is sought is varied. In this way the likelihood function remains a function of the single parameter being evaluated, because the other $M-1$ parameters then become implicit functions of it. An example to illustrate the results of this process is shown in Figure 4-2 for the case of a two-parameter χ^2 minimization: if, for instance, $\chi^2_{(N-2)}$ were not kept minimized with respect to a_2 as a_1 was varied, then σ_{a_1} would be identified with the points where the $\Delta\chi^2 = 1$ contour crosses the a_1 axis rather than with the correct, larger σ_{a_1} value associated with the projection of the contour onto the a_1 axis. This figure is sometimes referred to as the *error ellipse* for the two parameter estimates. The next subsection will address this issue in more detail.

The *Error Ellipse*

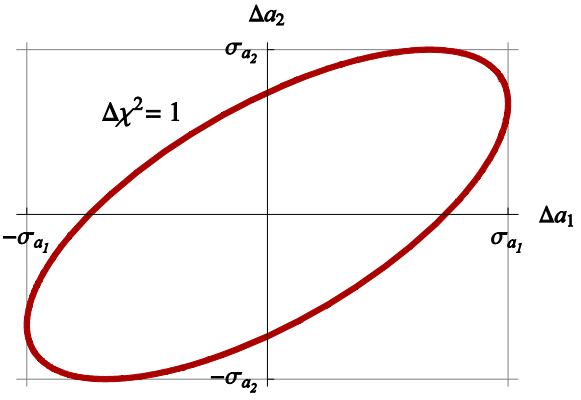


Figure 4-2: A hypothetical example of the correct determination of the uncertainties of two parameter values. The axis origin is at those values for a_1 and a_2 which minimize χ^2. The elliptical contour shows the locus of parameter value combinations which increase χ^2 by 1 from its minimum value. The proper estimates of the parameter uncertainties are at the extremes of the contour as projected on the a_1 and a_2 axes, and not where the contour crosses each axis.

The covariance matrix and its relationship to the Hessian matrix

Consider a case in which the theoretical relationship between x and y is expected to be a simple proportion: $y = f(x) = ax$. The expression (4.1) for χ^2 has $f(x_i) = ax_i$, which is clearly linear in the single unknown parameter a. χ^2 is a quadratic function of a, and the χ^2 minimization problem becomes a simple linear regression, with (4.4) reduced to a single equation for the value a_0 at which $(d/da)\,\chi^2 = 0$:

4.12

$$\chi^2 = a^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} - 2a \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} + \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2}$$

$$\left.\frac{d(\chi^2)}{da}\right|_{a=a_0} = 0 \quad \Rightarrow \quad a_0 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

Since (4.12) is a simple quadratic in a, $(d^2/da^2)\,\chi^2$ is everywhere constant and is equal to twice the coefficient of the quadratic term in (4.12). Note that this term is also the coefficient of a_0 in the equation above and is also nonnegative. Thus this equation does indeed identify the unique minimum of χ^2 with respect to the single parameter a. The uncertainty of the value a_0 is given by the $\Delta a = a - a_0$, which makes $\Delta\chi^2(\Delta a) = \chi^2(a_0+\Delta a) - \chi^2(a_0) = 1$. From (4.12) the equation for $\Delta\chi^2(\Delta a)$ is:

$$\Delta\chi^2 = (\Delta a)^2 \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} = (\Delta a)^2 \left(\frac{1}{2}\frac{d^2(\chi^2)}{da^2}\right) \quad \Rightarrow \quad 1 = \sigma_a^2 \left(\frac{1}{2}\frac{d^2(\chi^2)}{da^2}\right)$$

4.13

$$\boxed{\sigma_a^2 = 2\left(\frac{d^2(\chi^2)}{da^2}\right)^{-1}}$$

The variance of the estimated value of a single fit parameter using linear regression is twice the reciprocal of the second derivative of χ^2 with respect to the parameter. This result may be extended to the multi-parameter fit situation using a bit of straightforward but somewhat messy linear algebra. The results are shown in (4.14):

Covariances of the fit parameter estimates from χ^2 minimization

4.14

$$\boxed{\boldsymbol{\Sigma} = 2\,\mathbf{H}^{-1}}$$

Covariance matrix $\boldsymbol{\Sigma}$: $\Sigma_{jk} = \sigma^2_{a_j a_k}$ **Hessian matrix** $\mathbf{H}$: $H_{jk} = \left.\dfrac{\partial^2(\chi^2)}{\partial a_j\,\partial a_k}\right|_{\min(\chi^2)}$

The Covariance Matrix

The *covariance matrix* (or *error matrix*) of the fit parameter estimates is given by twice the inverse of the *Hessian matrix* (evaluated at the χ^2 minimum). The diagonal elements of this matrix are the individual parameter estimate variances. The off-diagonal elements are the covariances between the various parameter pairs.

Use of the inverse of the Hessian matrix to determine the parameter uncertainties ensures that they are calculated properly (Figure 4-2 on page 54).

The Hessian matrix is especially easy to calculate for the ordinary linear regression problem, equation (4.3) on page 44 with $\sigma_i^2 = \sigma_{y_i}^2$. In (4.4) the array of coefficients of the M parameters ($k = 1 \ldots M$) in the M equations ($j = 1 \ldots M$) form a $M \times M$ matrix which you can easily show to be half the Hessian matrix $\mathbf{H}$ of $\chi^2_{(N-M)}$. Using (4.14), it must be the case that the covariance matrix of the parameter estimates $\mathbf{\Sigma}$ is the inverse of the $M \times M$ coefficient matrix (4.15).

Hessian and covariance matrices of the ordinary linear regression problem

4.15

$$\mathbf{H}: \quad H_{jk} = \frac{\partial^2(\chi^2)}{\partial a_j \, \partial a_k} = 2\sum_{i=1}^{N} \frac{1}{\sigma_i^2} g_j(x_i) g_k(x_i)$$

$$\mathbf{\Sigma} = 2\mathbf{H}^{-1} = \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2} g_j(x_i) g_k(x_i) \right]^{-1} \qquad (j,k = 1 \ldots M)$$

For the general χ^2 minimization problem with the model function $f(x; a_1, a_2, \ldots, a_M)$ nonlinear in one or more of the parameters, then the expressions for the elements of the $M \times M$ Hessian matrix can be much more complicated.

If the regression is ordinary, with $\sigma_i^2 = \sigma_{y_i}^2$, then:

Ordinary nonlinear regression problem Hessian matrix

4.16

$$H_{jk} = \frac{\partial^2(\chi^2)}{\partial a_j \, \partial a_k} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial^2}{\partial a_j \, \partial a_k} \left(y_i - f(x_i; a_1, a_2, \ldots, a_M) \right)^2$$

$$= 2\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial f(x_i)}{\partial a_j} \frac{\partial f(x_i)}{\partial a_k} - \left(y_i - f(x_i) \right) \frac{\partial^2 f(x_i)}{\partial a_j \, \partial a_k} \right]$$

For a total regression problem (either linear or nonlinear), the Hessian matrix will include terms involving the derivatives of σ_i^2.

Correlations among the parameter estimates

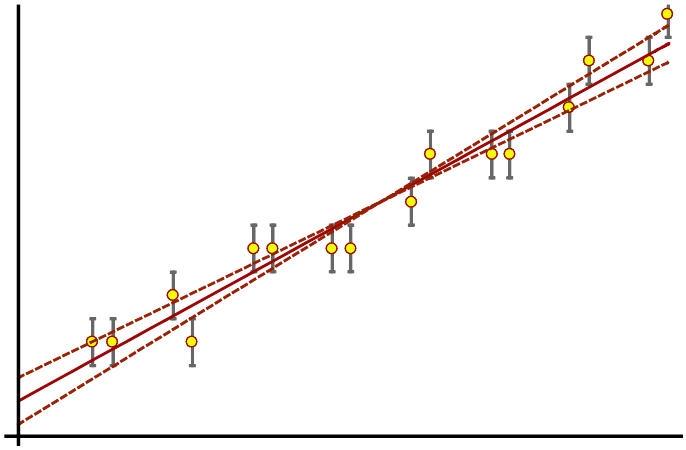
Figure 4-2 illustrates an issue which will often arise when more than one parameter is involved: the parameters' distributions are correlated, and therefore their covariances do not vanish. For example, consider again the simple 2-parameter linear regression problem, $y = f(x) = a_1 + a_2 x$, with equal uncertainties for all data points $\sigma_i^2 = \sigma^2$. The solution of the χ^2 minimization problem is given in (4.5) on page 45. The covariance matrix of the parameter estimates is σ^2 times the inverse of the coefficient matrix in (4.5), since σ^2 was divided out of both sides of (4.4) to get (4.5):

4.17

$$\begin{pmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2}^2 \\ \sigma_{a_1 a_2}^2 & \sigma_{a_2}^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} N & \Sigma x_i \\ \Sigma x_i & \Sigma {x_i}^2 \end{pmatrix}^{-1} = \sigma^2 \left[N \Sigma {x_i}^2 - (\Sigma x_i)^2 \right]^{-1} \begin{pmatrix} \Sigma {x_i}^2 & -\Sigma x_i \\ -\Sigma x_i & N \end{pmatrix}$$

The covariance $\sigma_{a_1 a_2}^2$ has a sign opposite to that of Σx_i. Clearly it will be nonzero unless $\Sigma x_i = 0$ (that is, the arithmetic mean of the x_i vanishes). That the covariance between the slope and y-intercept of the line is usually nonzero is illustrated in Figure 4-3 below.

Figure 4-3: A linear fit to the data points shown has a *y*-intercept outside the data's range. ±2σ errors in the fit's slope (shown by the dashed lines) result in corresponding errors in the fit's intercept. Consequently these errors in the fit's slope and intercept are correlated.



Uncertainties in model predictions

Once the model function's free parameter values have been optimized and the parameters' covariance matrix has been obtained, these results may be used to calculate the expected value and uncertainty of the model $y = f(x)$ for any particular input value x. The model predicts that $y = f(x; a_1, a_2, \ldots, a_M)$, where the parameters $a_1, a_2, \ldots, a_M$ take on their optimized values.

To calculate the uncertainty in this prediction, we will use the "naïve" uncertainty propagation formulas—in particular the second of equations (2.8) on page 28. Let us now introduce a more compact notation for that expression. The gradient vector of $y = f(x)$ with respect to the M parameters, $\nabla_{\mathbf{a}} f$, has components

$$\nabla_{\mathbf{a}} f \;=\; \left(\frac{\partial f}{\partial a_1}, \frac{\partial f}{\partial a_2}, \ldots, \frac{\partial f}{\partial a_M} \right)$$

The components of $\nabla_{\mathbf{a}} f$ are evaluated at $(x; a_1, a_2, \ldots, a_M)$, using the parameters' optimized values. $\nabla_{\mathbf{a}} f$ then combines with the parameter estimates' covariance matrix $\boldsymbol{\Sigma}$ to determine the uncertainty in the model prediction $y = f(x)$:

4.18 $$y \pm \sigma_y \;=\; f(x) \pm \sqrt{\nabla_{\mathbf{a}} f \cdot \boldsymbol{\Sigma} \cdot \nabla_{\mathbf{a}} f}$$ **Uncertainty in the model *y* = *f*(*x*)**

The dot products shown in the uncertainty calculation in (4.18) represent matrix multiplications wherein the left-hand $\nabla_{\mathbf{a}} f$ is written as a row matrix and the right-hand $\nabla_{\mathbf{a}} f$ as a column matrix. An example of the result of the calculation (4.18) is shown in Figure 4-4. Note how the model becomes more uncertain when extrapolated beyond the measured data range, as you might expect.

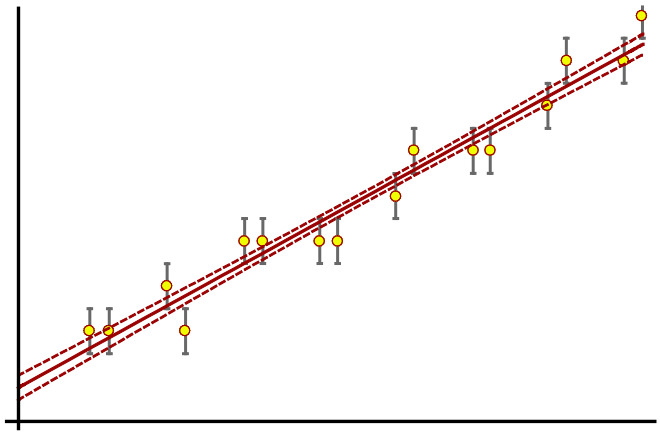


Figure 4-4: A linear fit to the same data points as those shown in Figure 4-3. The solid line shows the model *y* = *f*(*x*) which minimizes χ^2, and the dashed lines show the ±1σ boundaries on the model calculated using (4.18).

EVALUATING FIT RESIDUALS

Once chi-squared minimization has been used to fit a theoretical model $y = f(x)$ to a set of experimental data points (x_i, y_i), you must judge how well the optimized function actually describes the measured data and how trustworthy are the resulting free parameter estimates. An important tool to assist with these deliberations is an inspection of the fit residuals $y_i - f(x_i)$ along with a careful evaluation of the final reduced chi-squared value. This section addresses some issues typically encountered when using fit residuals and reduced chi-squared to assess fit results.

Data consistent with the model, and accurate uncertainty estimates

Chi-squared minimization using an ordinary least-squares algorithm produces a maximum likelihood result if the scatter in the fit residuals $y_i - f(x_i)$ is normally distributed. If the

optimized model is consistent with the measured data, then the resulting $\tilde{\chi}^2$ is expected to be 1. In this case, the data are well-described by the model (at least to the level of the noise in the data), and the optimized parameter values along with their covariance matrix provide maximum likelihood estimates and uncertainties for the theory's free parameters.

Consider the example shown in Figure 4-5, a laboratory ^{57}Fe Mössbauer nuclear absorption spectrum of a stainless steel sample. As a radioactive ^{57}Co source is moved toward and away from the absorber, it emits 14.4 keV (kilo-electron-volt) gamma rays which are very slightly Doppler shifted in energy by the instantaneous velocity of that relative motion. The data consist of the observed rates at which these gamma rays pass through the thin absorber as a function of the relative velocity of source and absorber. The apparatus divided the relative velocity range into discrete channels and recorded the number of detection events in each velocity channel. The count rate uncertainties (shown by the error bars in Figure 4-5) were derived by assuming that Poisson counting statistics describe the expected variations in the observed number of counts in a channel.

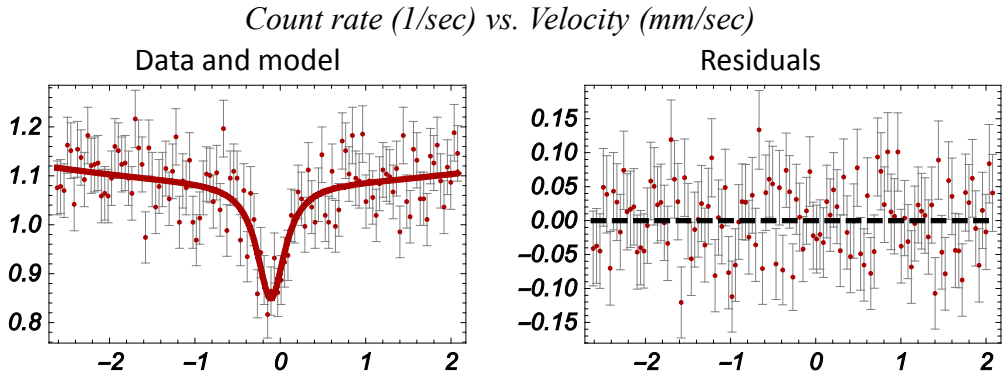


Figure 4-5: Iron-57 Mössbauer nuclear absorption spectrum of a type 302 stainless steel sample. The data were collected for 6 minutes as the cobalt-57 source was oscillated with a constant acceleration toward and away from the absorber. The source relative velocity profile was first calibrated using a standard natural iron absorber. Count rate uncertainties were estimated using Poisson count statistics and are shown by the error bars on each data point. The theoretical model used to fit the data had 5 free parameters: 3 fundamental and 2 experiment-specific. The optimized model $y = f(x)$ is plotted in the left-hand graph, and the residuals $y_i - f(x_i)$ are shown in the right-hand graph.

The model used to fit the data set shown in Figure 4-5 had 5 free parameters: the fundamental physical theory of the nuclear gamma ray interaction predicts the absorption line position and shape using 3 free parameters, and the specific experimental setup required 2 more. A nonlinear, ordinary least-squares algorithm resulted in the optimized model plotted in Figure 4-5 along with its residuals $y_i - f(x_i)$. The 119 data points and 5 optimized parameters left 114 degrees of freedom for the $\tilde{\chi}^2$ variate. Following the minimization, $\tilde{\chi}^2 = 0.980$, well within one standard deviation of unity, the value expected for a model consistent with the observed data.

The optimized values of the parameters represent maximum likelihood estimates only if the residuals $y_i - f(x_i)$ are consistent with independent, normally distributed samples. There is no obvious systematic pattern to the residuals in the right-hand graph of Figure 4-5, and the $\tilde{\chi}^2$ value so near unity implies that the magnitude of their scatter about the model $y = f(x)$ is consistent with the data point uncertainties. As a further test, we apply the simple, quartile Pearson's chi-squared test, equation (3.9) on page 39. We histogram the residuals and compare to a Gaussian model; the results are shown in Figure 4-6. The encouraging match of the histogram to a normal distribution indicates that the optimized free parameter values and their associated uncertainties (or covariance matrix) may be reasonably identified with maximum likelihood estimates, and that the experimental data are completely consistent with the theoretical model and its optimized parameter values.

Experiments such as the one described here are often intended not to test theoretical models but rather to use well-established models in order to better refine empirical estimates of physical constants, such as the electron charge/mass ratio example discussed in the previous chapter. Results analogous to those in Figure 4-5 indicate that an experiment designed to use a particular theoretical model is successful, and that the measured value of a parameter and its associated uncertainty can be trusted, except for possible, lingering systematic errors not properly accounted for by the experimenter (see Chapter 5). In fact, conversely, such an experiment may use well-established theory and precisely-measured values of physical constants to *perform a calibration* of a scientific apparatus or instrument. In this case, the experiment is designed to measure relevant parameters which can then be used to correct for systematic errors and thereby greatly improve the instrument's accuracy.

Choosing between two optimized theoretical models

Scientific experiments are often performed to attempt to choose between two or more competing theories. If these theories provide incompatible predictions of an experiment's results, even when each has had its free parameter values optimized, then a precisely measured data set may provide the evidence needed to determine which theory is more likely to correctly model the phenomena explored by the experiment.

For example, consider a crystalline semiconductor sample into which impurities have been

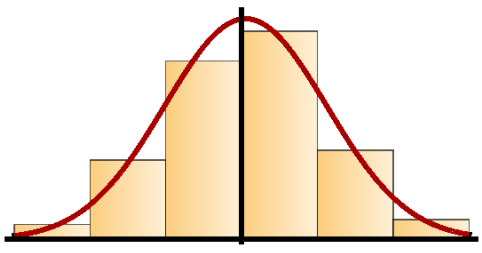


Figure 4-6: Histogram of the fit residuals shown in Figure 4-5. The residuals are quite accurately characterized as samples of a normal distribution, shown by the superimposed curve. The reduced χ^2 of 0.98 indicates that the data are consistent with the model, and therefore the optimized parameter values provide reasonable maximum likelihood estimates of the free parameters and their uncertainties.

introduced to provide a fixed cadre of charge carriers. At moderately low temperatures, these charge carriers can completely dominate the conductivity of the material, and the variation of its conductivity with temperature may then provide important information concerning the kinematics of charge carrier motion within the material. The measured temperature dependence of the resistance of a semiconductor crystal is shown in Figure 4-7.

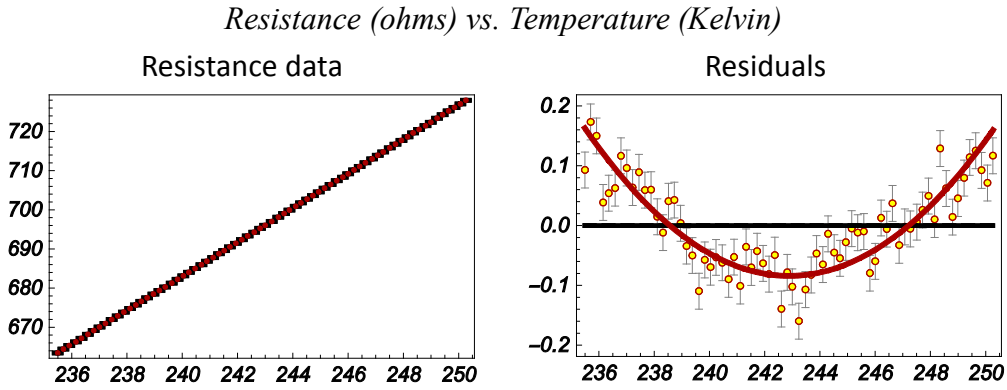


Figure 4-7: Temperature dependence of the resistance of a semiconductor crystal rod "doped" with impurity atoms to make it conductive at temperatures well below those which would ionize a significant fraction of the semiconductor atoms. The left-hand graph shows the measured data, each with resistance uncertainty of ±0.03 ohm. Although the data in the left-hand graph appear to be quite linear, a fit to a linear model shows a small, but significant, systematic pattern to the residuals (points and black line in the right-hand graph). Fitting to a model varying with temperature as $T^{3/2}$, on the other hand, results in much smaller, seemingly random residuals (red line in right-hand graph). Data measured during an undergraduate physics laboratory experiment.

At temperatures low enough to thermally ionize only an insignificant fraction of the semiconductor atoms of the crystal (but not so low that the impurity-provided charge carriers are frozen out), the charge carrier density in the crystal is very nearly independent of temperature, as in a normal conductor (a metal). If these charge carriers behave kinematically like those in a normal conductor, then the material's resistance should rise linearly with temperature. Alternatively, the charge carriers might behave kinematically more like the particles in a classical gas (or plasma), in which case the material's resistance should increase as the 3/2 power of the temperature. As shown in the residual plot in the right-hand graph of Figure 4-7, this latter alternative appears to be much more consistent with the measured data.

The linear model, $y = f(x) = a_1 + a_2 x$, when fit to the 70 data points resulted in a least-squares-minimized $\tilde{\chi}^2 = 6.22$. If this were the correct model, we would expect such a large reduced chi-squared value (over 30 standard deviations >1) an insignificant fraction of the time ($\approx 10^{-52}$). This conclusion is further supported by the observation that the residuals from the fit to the linear model in Figure 4-7 show a clear, systematic variation across the data set. The alternative model, $y = g(x) = a_1 + a_2 x^{3/2}$, fares much better. Its $\tilde{\chi}^2 = 1.51$, about 3 standard deviations >1. Although still seemingly rather improbable for an appropriate theoretical model, it is much more reasonable than the linear model result. Of the two

models, this latter is much more consistent with the measured data. The residuals $y_i - g(x_i)$ are shown in Figure 4-8; the model agrees with the measured data to within 0.1 ohm (out of an average of 700 ohms) over the tested temperature range (the standard deviation of the residuals is 0.037 ohm).

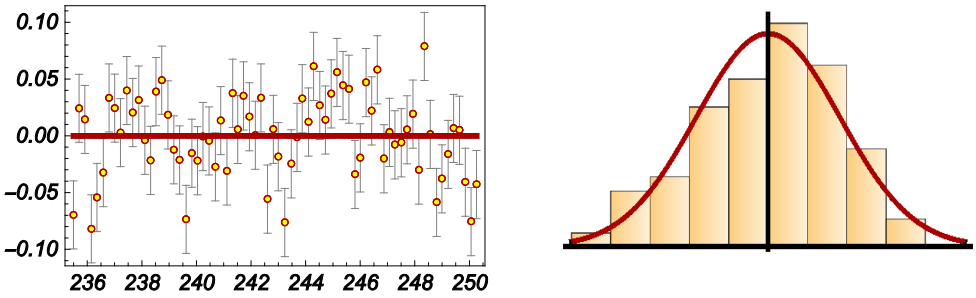


Figure 4-8: Residual plot and histogram of the residuals for the semiconductor data of Figure 4-7 fit to a $T^{3/2}$ model. The reduced χ^2 is 1.51, and the standard deviation of the set of residuals is .0366 ohm, compared to the resistance uncertainty of 0.03 ohms assumed for the data. These values are approximately 5×10^{-5} of the measured resistance values, near the instrument's precision limit.

The residuals of the $y = a_1 + a_2 x^{3/2}$ model fit shown in Figure 4-8 appear to have a small, systematic pattern with a period of about 3.5 Kelvin, or about every 15 ohms rise in resistance. This pattern may explain the $\tilde{\chi}^2 = 1.51$, and it may be better understood by considering the details of the instrumentation used and how the measurement uncertainties were determined. The data point uncertainty of 0.03 ohm was determined by analyzing the scatter in a set of resistance data acquired while the temperature was held constant, using the point estimation techniques of Chapter 2. An additional source of error as the temperature is varied could be due to nonlinearity in the resistance measurement. The manufacturer-specified maximum error due to nonlinearity of the 22 bit, 4-wire resistance measuring instrument is stated as 0.0006% of the 1000 ohm scale used, or 0.006 ohm.* Adding this nonlinearity error to the 0.03 ohm measurement uncertainty yields 0.036 ohm, tantalizingly close to the residuals' standard deviation of 0.037 ohm. Although not conclusively proven by this argument, it provides evidence that the $\tilde{\chi}^2 = 1.51$ can be at least partially understood by the small instrument nonlinearity. Chapter 5 further discusses the measurement problems introduced by systematic errors, of which this instrument nonlinearity is an example.

Note that the above observations do not invalidate the conclusion that a $T^{3/2}$ model of the dependence of the resistance is more consistent with the experimental result than a simple, linear temperature dependence. The smooth, systematic pattern to the linear model residuals

* This error due to nonlinearity is inherent to the instrument's measurement hardware and cannot be improved by a simple calibration procedure. On the other hand, *calibration errors* result in linear scale factor and 0-offset errors which would not affect the observed pattern of the residuals. Errors due to instrument nonlinearity, scale, and offset are examples of *systematic error*, discussed in the next chapter.

shown in Figure 4-7 is quite accurately captured by the $T^{3/2}$ model, as is demonstrated by the red curve in Figure 4-7.

Evaluating the accuracy of a theory

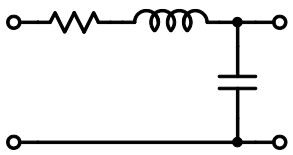


Figure 4-9: RLC voltage divider (resonant filter). Input voltage at left, output at right.

As illustrated in the previous section, precise data permits a quantitative evaluation of the accuracy of a theory's prediction of an experiment's results. Let us further examine this issue. Consider, for example, the carefully-measured frequency response of a series-connected *RLC* (resistor-inductor-capacitor) filter circuit shown in Figure 4-9 at right. When a sinusoidal voltage is input to the left-hand pair of terminals, the corresponding sinusoidal voltage generated at the right-hand terminals (across the capacitor) displays a typical resonant response described (to lowest order) by a version of the classic linear, second-order, damped harmonic oscillator theory. That theory, when applied to this experiment, has only two free parameters: the resonant frequency f_0 and the dimensionless quality factor (Q) of the resonance, both of which should be determined by the circuit's R, L, and C values. The theory predicts that at low frequency the circuit's voltage gain will approach unity with no phase shift between the output and input sinusoids; at high frequencies the voltage gain will become proportional to the inverse square of the signal frequency and the output's phase shift will approach −180°. For values of $Q \gg 1$ the theory predicts that the transition between these asymptotic behaviors is mainly confined to within a few Qths of the circuit's resonant frequency f_0. At f_0 the voltage gain should be Q and the phase shift from input to output should be −90°.

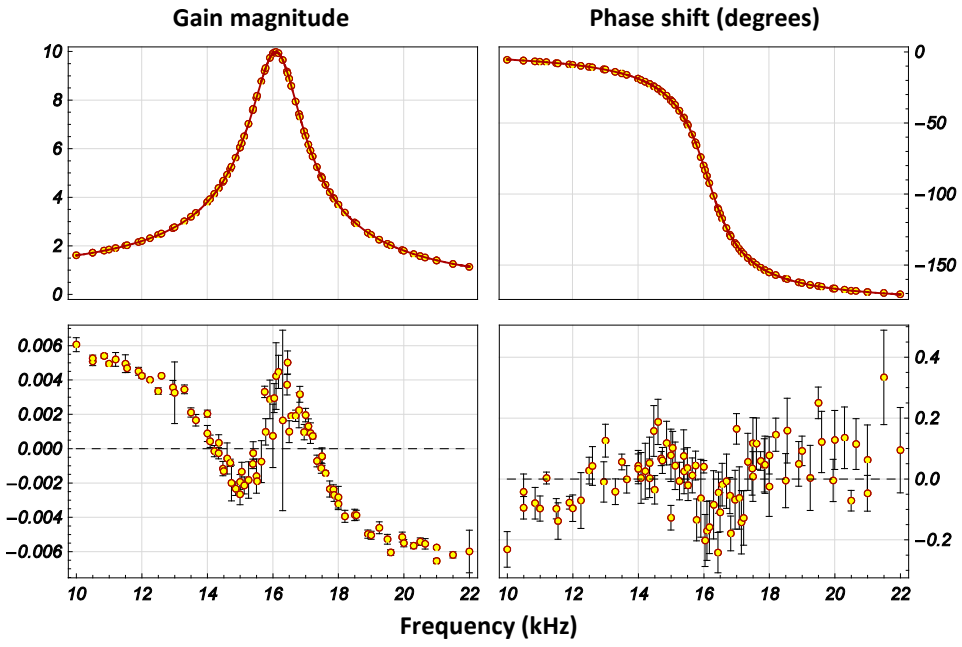
The experiment was performed by a computer-controlled data acquisition system that precisely measured the amplitudes and phases of the input and output voltage waveforms as it swept the input signal (generated by a high-quality, stable signal source) through a real *RLC* circuit's expected resonance frequency. At each selected frequency the input and output waveforms were measured and analyzed multiple times to provide resulting amplitude and phase values along with associated uncertainties using the techniques described in Chapter 2. The resulting data points' amplitude gains and phases were fit to the theoretical models shown in (4.19), whose parameters A, f_0, ϕ_0, and Q were optimized using χ^2 minimization.

$$\left|\frac{V_{out}}{V_{in}}\right| = \frac{A}{\sqrt{\left[1-(f/f_0)^2\right]^2 + (f/Qf_0)^2}} \qquad \text{4.19}$$

$$\phi_{out} - \phi_{in} = \arctan\left[Q\left(\frac{f_0}{f} - \frac{f}{f_0}\right)\right] - (90° + \phi_0)$$

The additional parameters A and ϕ_0 provide possible optimized corrections to the theory's predictions of the circuit's asymptotic gain and resonant phase shift, respectively. They should equal 1 and 0, respectively, if the unmodified, simple theory were to perfectly predict the experiment's results.

Plots of 87 data points and optimized model fits are shown below. Separate fits of the theory's predictions (4.19) to the amplitude and phase shift data were performed, resulting in independent determinations of the fundamental parameters f_0 and Q. These determinations should, of course, agree if the theory adequately describes the circuit's behavior.



$\tilde{\chi}^2 = 298$
$f_0 = 16.1327 \pm 0.0002\,\text{kHz}$
$Q = 10.020 \pm 0.004$
$A = 0.9945 \pm 0.0002$

$\tilde{\chi}^2 = 2.9$
$f_0 = 16.1358 \pm 0.0003\,\text{kHz}$
$Q = 10.016 \pm 0.003$
$\phi_0 = -0.426° \pm 0.007°$

Figure 4-10: Measured frequency response (gain and phase) of an RLC circuit (Figure 4-9). Using chi-squared minimization, the data were fit to a simple model of a damped harmonic oscillator driven by a sinusoidal input. The results of the fit including plots of the fit residuals are shown. The precision of the measured data provides for an accurate quantitative analysis of the accuracy of the chosen theoretical model. The definitions of the fit parameters used are described in the text.

The most striking feature of the χ^2 minimization result is the obvious pattern in the optimized model fit's gain residuals along with the its very large $\tilde{\chi}^2_{84}$ (87 data points and 3 optimized parameters leaves 84 degrees of freedom): the optimized model's deviations from the measured data are very much larger than most of the points' uncertainties. The optimized model of the phase data, on the other hand, has a 100 times smaller $\tilde{\chi}^2$, although there is also

a noticeable upward trend with frequency to its fit residuals, and a reduced chi-squared of 3 is still quite large for a data set with 84 degrees of freedom. The added fit parameters A and ϕ_0 also show significant differences from their ideal theoretical values of 1 and 0, differences which are many times larger than their estimated uncertainties. Although fits to the two data sets provide Q values which agree quite well, their estimates of the resonant frequency f_0 differ by about 10 times their uncertainties.

One should note, however, that despite the discrepancies noted above, the theoretical model (4.19) describes the observed circuit behavior quite accurately. Again examine Figure 4-10: visually, the model fits (solid red lines) in the upper plots appear remarkably consistent with the measured data. The gain residuals plot demonstrates that as the observed gain varied from 1.5 to 10 (more than a factor of 6) over a ±38% frequency variation around the resonance, the maximum deviation of the measured gain data from the model was only about 0.006, less than 0.1% of the maximum gain. The total deviation of the corresponding phase data around its optimized model was only 0.6° out of the nearly 180° phase variation, or 0.33%. The gain and phase model fits provided values for f_0 and Q which agree to within 0.02% and 0.004%, respectively. The optimized value of the gain "correction" parameter A is only 0.005 less than 1, and the low-frequency trend in the gain residuals shows that the observed gain did indeed approach very close to 1, as predicted by the theory. The phase "correction" ϕ_0 is only about 0.5° away from the theoretical value of −90°, again quite small. Clearly, this experiment has demonstrated that a very simple model can provide quite accurate predictions of the behavior of a real *RLC* system, at least within the parameter space investigated by this data set. The simple theory's two abstract concepts, resonant frequency f_0 and quality factor Q, can obviously be very useful when applied to specific, real world situations such as this one.

Thus the results of this experiment were precise enough to accomplish two important goals: (1) establish with high confidence just how accurately the theory was able to predict the behavior of the experimental system, and (2) provide accurate data regarding how the theory's predictions failed to explain fine-level details in the system's behavior. In particular, careful analyses of the patterns in the residuals shown in Figure 4-10 provide clues as to how the existing theory might be improved. Contrast this with situation illustrated in Figure 4-5 on page 59, where the residuals were completely dominated by random fluctuations in the data, resulting in a reduced chi-squared very near 1.

> High precision, accurate data is the key to revealing new physics not adequately described by existing theory. Scientific progress happens because of reduced chi-squared results $\gg 1$, which is thus the goal of many experimental investigations.

Of course, it is not altogether unlikely that a pattern in the residuals such as that evident in Figure 4-10 was the consequence of some unanticipated behavior of the experimental apparatus or technique, and not because of the underlying physical phenomenon being

investigated. Such potential sources of systematic errors are addressed in the next chapter. If modification of the apparatus or the data gathering technique leads to a significant change in the pattern of the fit residuals, then the experiment itself may be the culprit, and not the fundamental physical behavior being investigated. It is therefore important that the experimenter not jump to conclusions prematurely, but remain skeptical and critically examine all aspects of the experiment and its analysis.

In the case of this particular experiment, the precisely-measured but small deviations from the simple damped harmonic oscillator theory are mainly caused by the circuit element modelled by the inductor in Figure 4-9. This "inductor" was a tightly wound coil of wire with many turns wrapped around a core of soft, powdered iron shaped into a toroid and embedded in an insulating epoxy matrix. The theoretical model of this structure as a simple, pure inductance has two major failings as far as this experiment is concerned:

(1) Adjacent pairs of closely-spaced wire windings have a small amount of parasitic capacitance which, to lowest order, represents a net capacitance in parallel with the element's inductance, complicating the circuit model.
(2) The powdered iron core's magnetization in response to the magnetic field generated by current through the wire coil may greatly increase the magnetic flux and thus the element's inductance, but it also adds hysteresis to the time-dependent relationship between the coil's current and the resulting magnetic flux. This hysteresis not only adds an additional source of dissipation (included in the model's resistor), but this effect is nonlinear, making the element's inductance as well as its effective resistance slight functions of both the frequency and the amplitude of the current through it.

The effect (1) accounts for much of the amplitude data's deviation from the model, but the small variations in the amplitude and phase residuals within a Qth of the resonant frequency are also evidence of effect (2). Interestingly, the general slope to the residuals in the fit to the phase data (Figure 4-10) and the nonzero value for phase fit parameter ϕ_0 is almost certainly due to the presence of some residual systematic error source in the experiment's phase measurement technique. This sort of problem and how to mitigate its effects is addressed in the next chapter.

Chapter 5
Dealing with systematic errors

THE NATURE OF SYSTEMATIC ERROR AND UNCERTAINTY

This chapter expands on the discussion of *systematic error*, first addressed in Chapter 2. Here we provide a few specific examples of systematic error sources and introduce basic techniques for how to properly handle the uncertainties they introduce into an experiment's results. Recall the definition of *systematic error*: it is the general term used to describe errors introduced during the design, construction, and data acquisition phases of an experiment which affect the accuracies of all measurements in strongly correlated ways. Because of the finite accuracies attainable when constructing and calibrating an experimental apparatus and the instruments used for its various measurements, residual errors in numerical parameters which characterize the apparatus will introduce errors into the measurements and analyses of all of an experiment's data points. As mentioned in Chapter 2, sources of systematic error lurk everywhere: the calibration error in a voltmeter, the angular alignment error of the fixed arms of an interferometer, the position and alignment errors in the placement of detectors around a particle collision site, the machining errors in the dimensions of a resonant cavity, trajectory calculation errors introduced by improperly analyzed fringe fields of an electromagnet, changes in the dimensions or electrical characteristics of the apparatus caused by slow, gradual changes in laboratory temperature or humidity, etc.

Because inaccuracies introduced by systematic errors are traceable to residual errors in the characterizations of the apparatus and the instruments themselves, the resulting data point errors will be strongly correlated and will not, in general, be characterized by noticeable fluctuation from measurement to measurement. For example, if improper calibration of a voltmeter introduces a 1% error in its readings (all voltages it reports are 1% larger than the actual values, say), then averaging multiple readings will result in a mean measured voltage with an expected value which will be off by 1%, no matter how many readings are averaged, and no matter to how small the resulting standard deviation of the mean value may seemingly be reduced.

Because systematic errors do not manifest themselves as seemingly random, independent fluctuations in the results of repeated measurements, many of the statistical analysis methods introduced in the previous chapters become inapplicable when analyzing their possible effects on the measured data. ***Data uncertainties introduced by systematic errors are not improved by averaging the results of repeated measurements.***

Because the errors in measurements introduced by systematic error sources represent a fundamental limit to the accuracy of an experiment's results, it is a waste of the experimenter's time to work hard to reduce the effects of noise fluctuations to levels much lower than this fundamental limit (using the techniques presented in earlier chapters). Instead, efforts to identify, understand, and reduce potential systematic errors must go hand in hand with efforts to reduce the effects of noise when designing and conducting an experiment. Thus it is important to estimate the *uncertainties* in these systematic numerical values, that is, the experimenter's *best estimates* of the magnitudes of the various systematic errors. With this information in hand, the experimenter may then evaluate the impacts of each of these errors on the accuracy of the experiment's results. Along with estimates of the noise fluctuations expected to be generated by various components in the apparatus, the experimenter can create an *error budget* for a given apparatus and instrument configuration. Comparison of the error budget to the accuracy requirements of the experiment becomes a key part of an experiment's design. The experimenter can then identify those sources which are the major limiters to the experiment's accuracy and consequently focus time and money on those aspects of the design, construction, and calibration of the apparatus which reduce their collective impact to an acceptable level.

INCORPORATING SYSTEMATIC UNCERTAINTY ESTIMATES

Because uncertainties due to possible systematic errors cannot generally be reduced by averaging repeated measurements, they establish a limit on the accuracy of an experimental result. Consequently, they must be incorporated into the experiment's uncertainty determination near the very end of the analysis and after following any of the curve fitting techniques addressed in Chapter 4. Those techniques are only appropriate in the face of errors which are independent from measurement to measurement, such as fluctuations due to noise. The various data point uncertainties σ_i appearing in expressions such as (4.1) and (4.2) on page 43 represent the magnitudes of *independent data fluctuations*, and not the strongly correlated errors introduced by systematic sources. Let me emphasize this point:

Whenever calculating individual data point uncertainties that will then be used for curve fitting and parameter optimization, NEVER INCLUDE SYSTEMATIC ERROR SOURCES!

Failure to follow this rule is one of the most common mistakes made by beginning data analysts, and it is the most common reason for unacceptably small reduced chi-squared results ($\tilde{\chi}^2 \ll 1$).

Instrument calibration uncertainties

Our first example of how to properly incorporate systematic uncertainty into an experiment's result is a particularly simple, but illustrative, one: including the calibration uncertainty of a voltage measurement. Here's the scenario: the data are pairs of values (x_i, y_i) along with uncertainties σ_i determined from the scatter in the measured data points. The associated theory predicts that $y = a\,x,$ and the interesting physics happens to be in the value of the slope a. The y_i data consist of measured voltages, and the voltmeter has calibration uncertainties provided by its manufacturer. Because these calibration uncertainties are systematic, *they must not be incorporated into the individual data point σ_i determinations*. The x_i values are not measured using that same instrumentation.

The experimenter fits the data to the function $y = a\,x + b,$ rather than the simpler theoretical form, because there is probably an unknown, systematic offset error in the voltage measurements. This offset error could include a contribution from the voltmeter as well as from electrical grounding imperfections within the apparatus. Using the additional parameter b will result in a fit which can accommodate this source of error as described in a subsequent section of this chapter (starting on page 70). Once the data have been fit using the methods of Chapter 4, resulting in estimates of the slope $a \pm \sigma_{fit}$, incorporation of the remaining systematic uncertainties may proceed. Note that we have denoted the uncertainty in a's fit result as σ_{fit} to help keep the various uncertainties organized.

To make it clear how we proceed to include the voltmeter uncertainty, realize that a measured voltage v_m can be related to the actual voltage v using the simple relation $v = c\,v_m + e,$ where c and e are the voltmeter calibration gain (slope) and offset.[*] A perfect calibration would require that $c = 1$ and $e = 0$, but, of course, some residual errors must be present. These errors are characterized by calibration uncertainties, so that $c = 1 \pm \sigma_c$, and $e = 0 \pm \sigma_e$. For example, the Agilent 34410A has a specified accuracy on its 100 mV scale of 0.0050% of the voltage reading plus 0.0035% of the selected range scale (100 mV).[†] This should be interpreted as specifying that $\sigma_c = 5.0\times10^{-5}$ and $\sigma_e = 3.5\times10^{-5}\times100\,\text{mV} = 3.5\times10^{-6}\,\text{V}$. Because the experimenter's chosen fitting function $y = a\,x + b$ includes an offset b, the fit result should adequately correct for any actual nonzero calibration error e.

Thus we need only assess the impact of the calibration uncertainty σ_c on the fit's slope a, and, in particular, on the final value of σ_a. The actual y voltage values are related to measured values $y = c\,y_m + e,$ so that the fit's slope should be corrected to $a \to c\,a = a,$ because the expected value of c is 1. So in this case the uncertainty in the meter's calibration doesn't affect the final value quoted for the fit's slope parameter, *but it does affect the*

[*] For this elementary discussion we neglect any nonlinearities in the transformation between the actual and measured values.

[†] *Agilent 34410A/11A/L4411A User's Guide*, fifth edition, June 2012, © Agilent Technologies, Inc., 2005-2012. The stated accuracy is for within 1 year and 5°C of factory calibration for a DC voltage measurement.

parameter's uncertainty. Using the methods of Chapter 2 we must propagate the uncertainties of both c and a in the product ca to get the final experimental uncertainty. The resulting slope uncertainty, including the voltmeter calibration, is given in (5.1).

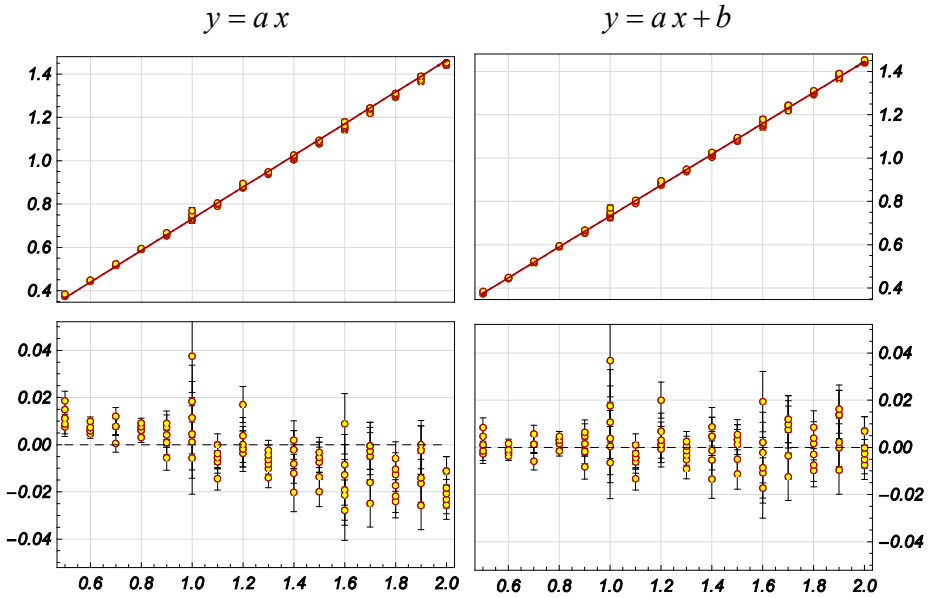
Including calibration uncertainty in the fitted slope:

$$\frac{\sigma_a}{a} = \sqrt{\left(\frac{\sigma_{fit}}{a}\right)^2 + \left(\frac{\sigma_c}{c}\right)^2} = \sqrt{\left(\frac{\sigma_{fit}}{a}\right)^2 + (\sigma_c)^2} \qquad \textbf{5.1}$$

As should be expected, the final fractional uncertainty in the experiment's value for the slope cannot be reduced below the voltmeter calibration's "% of reading" uncertainty σ_c.

INCLUDING UNKNOWN SYSTEMATIC ERRORS AS PARAMETERS IN AN AUGMENTED THEORY

Sometimes it happens that a possible systematic error value can be incorporated as an additional unknown free parameter in the functional representation of the theory the experiment is designed to test. A common example would be a constant offset error in the



$\tilde{\chi}^2 = 4.8$
$a = 0.7316 \pm 0.0004$

$\tilde{\chi}^2 = 1.17$
$a = 0.7127 \pm 0.0011$
$b = 0.0196 \pm 0.0011$

Figure 5-1: Including a parameter to account for a measurement offset. The fits and residuals are graphed for two different models; the right-hand model includes an additional parameter to accommodate the systematic offset error in the measurements.

measurement of, say, a data point y value. If the expected functional relationship is that y is proportional to x, i.e. $y = a\,x$, and the interesting physics happens to be in the value of the slope a, then fitting the data to the augmented function $y = a\,x + b$ would result not only in an estimate of the systematic offset b, but would also automatically reduce the effect of the offset on the experimentally obtained value for the slope a. The optimized value of the offset parameter b (along with its uncertainty) can then be compared to what was the expected magnitude of the systematic offset error, as shown in the example in Figure 5-1.

This example of a constant offset parameter added to the theoretical functional relationship $y = a\,x$ is a very common practice.

It is almost always a wise choice to augment the theoretical fitting function by including one or two additional parameters which can account for various systematic errors.

BACKGROUND SUBTRACTION

It is also important to keep this last method in mind when fitting a theory to data which may include points affected by the presence of *background events* not directly relevant to the phenomenon under investigation. Although not strictly an example of systematic error as we've defined it, in many experiments some level of background activity unrelated to the theory under investigation will be present. Assume that the y data values contain this background activity which is a function of the x values: $y = b(x)$, where b represents the background function. If y responds linearly to the presence of both the background b and the function of interest f, then $y = f(x) + b(x)$. Both f and b may depend on unknown free parameters, although the ones defining f are the more interesting ones. If the magnitude of b is significant, then fitting $y = f(x) + b(x)$ rather than just $y = f(x)$ may provide a more accurate result for f and its free parameter values.

This technique is appropriate in many, but not all, cases. Consider a simple example for which it works quite well. Figure 5-2 on page 72 shows part of a high energy photon spectrum captured using a scintillation detector. A "full energy" peak is isolated in the portion of the data shown, which plots the number of events recorded (y) in each detector energy channel (x). Although the *Poisson distribution* more properly describes the shape of the peak, at these energies this distribution is nearly indistinguishable from a Gaussian with the same standard deviation. The data are plotted two different ways in Figure 5-2: the left plot uses a linear vertical scale to highlight the Gaussian shape of the peak, whereas the logarithmic scale in the right plot emphasizes the low-intensity data on either side of the peak. Evidently, the Gaussian peak sits on top of a much weaker background event spectrum which appears to decrease with increasing energy. This background spectrum is generated by photon interactions in the detector which are unrelated to those generating the full energy peak (i.e., primarily by multiple Compton scattering of one or more additional photons).

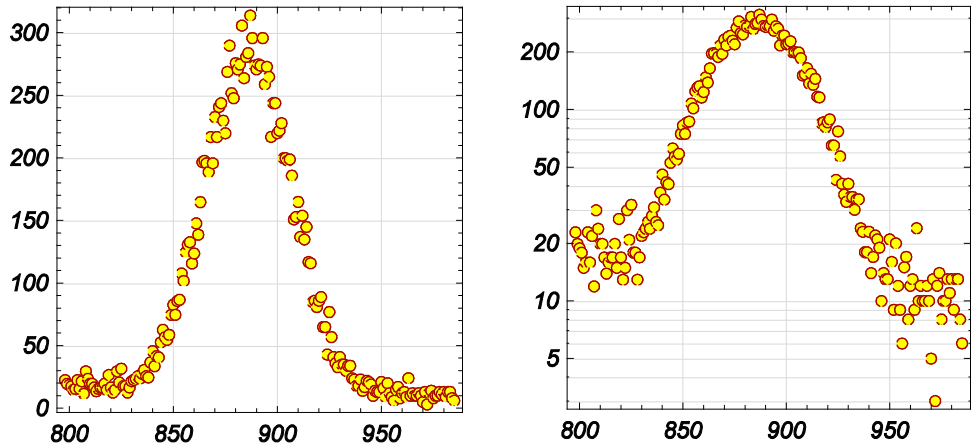


Figure 5-2: Two views of a gamma ray photon energy spectrum. The two graphs show the same histogram of event counts (vertical) vs. detector energy channel (horizontal). The Gaussian-shaped distribution has a width determined by the detector's energy resolution. The semi-log plot on the right emphasizes the background spectrum upon which the peak is superimposed.

Ignoring the background and fitting a Gaussian function to the data in Figure 5-2 may result in a poor fit. Adding a simple background function $b(x)$ can offer significant improvement, as demonstrated in Figure 5-3 on page 73. Just including a constant background to the fitting function resulted in a huge improvement to the fit and changed the estimated width of the peak by nearly 10%. Using a linearly-varying background changed the estimated peak position by nearly half a channel width, twice the estimated peak position's uncertainty.

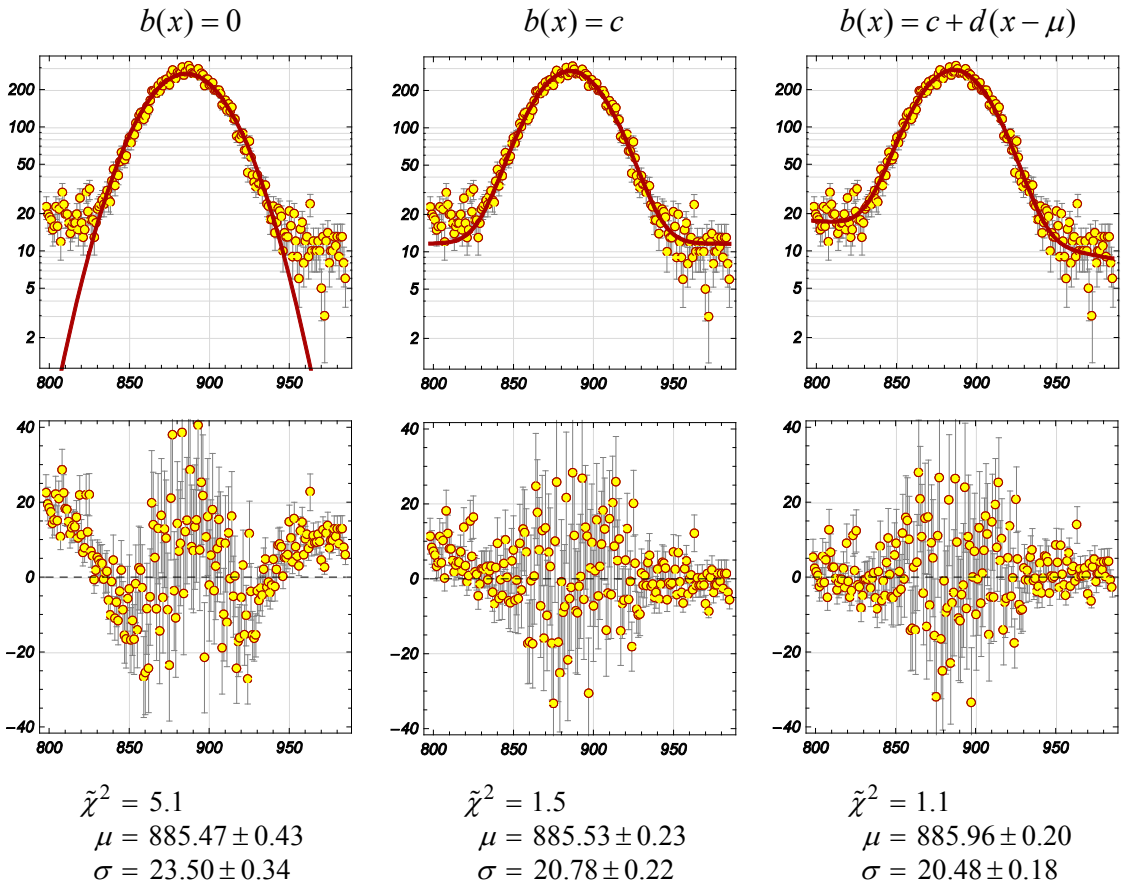


Figure 5-3: Including a function to model the background can greatly improve the fit to the data. Data point uncertainties were estimated using an approximation to Poisson counting statistics ($\sigma_y^2 \approx y$).

Chapter 6
Other important distributions