Statistical Analysis of Data
for PHY4803L

Robert DeSerio

University of Florida — Department of Physics
PHY4803L — Advanced Physics Laboratory

Contents

Chapter 1

Introduction

Data obtained through measurement always contain random error. Random error is readily observed by sampling—making repeated measurements while all experimental conditions remain the same. For various reasons, the measured values will vary and a histogram like that in Fig. 1.1 might be used to display a *sample frequency distribution.* Each histogram bin represents a possible value or a range of possible values as indicated by its placement along the horizontal axis. The height of each bar gives the frequency, or number of times a measurement value falls in that bin.

The measurements are referred to as a *sample* or as a *sample set* and

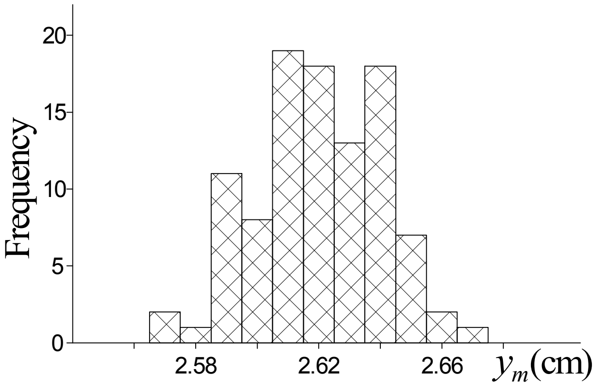


Figure 1.1: A sample frequency distribution for 100 measurements of the length of a rod.

the number of measurements N is called the *sample size.* Dividing the frequencies by the sample size yields the bin fractions or the *sample probability distribution.* Were new sample sets taken, the randomness of the measurement process would cause each new sample distribution to vary. However, as the sample sizes grow larger, variations in the distributions grow smaller and as $N \to \infty$, the *law of large numbers* says that the sample distribution converges to the *parent distribution*—a distribution containing complete statistical information about the particular measurement.

Thus, a single measurement should be regarded as one sample from a parent distribution—the sum of a non-random signal component and a random noise component. The signal component would be the center value or mean of the measurement's parent distribution, and the noise component would be a random error that scatters individual measurement values above or below the mean.

Briefly stated, measurement uncertainty refers to the distribution of random errors. The range of likely values is commonly quantified by the distribution's *standard deviation.* Typically, about 2/3 of the measurements will be within one standard deviation of the mean.

With an understanding of the measuring instrument and its application to a particular apparatus, the experimenter gives physical meaning to the signal component. For example, a thermometer's signal component might be interpreted to be the temperature of the system to which it's attached. Obviously, the interpretation is subject to possible errors that are distinct from and in addition to the random error in the measurement. For example, the thermometer may be out of calibration or it may not be in perfect thermal contact with the system. Such problems give rise to systematic errors—non-random deviations between the measurement mean and the physical variable.

Theoretical models provide relationships for physical variables. For example, the temperature, pressure, and volume of a quantity of gas might be measured to test various equations predicting specific relationships among those variables. Devising and testing theoretical models are typical experimental objectives.

Broadly summarized, the analysis of many experiments amounts to a compatibility test for the following two hypotheses.

Experimental: For each measurement the uncertainty is understood and any systematic error is sufficiently small.

Theoretical: The physical quantities follow the predicted relationships.

Experiment and theory are compatible if the deviations between the measurements and predictions can be accounted for by reasonable measurement errors.

If compatibility can not be achieved, at least one of the hypotheses must be rejected. The experimental hypothesis is always first on the chopping block because compatibility depends on how the random measurement errors are modeled and it relies on keeping systematic errors small. Only after careful assessment of both sources of error can one conclude that predictions are the problem.

Even when experiment and theory appear compatible, there is still reason to be cautious—one or both hypotheses can still be false. In particular, systematic errors are often difficult to disentangle from the theoretical model. Sorting out the behavior of measuring instruments from the behavior of the system under investigation and designing experimental procedures to verify all aspects of both hypotheses are basic goals of the experimental process.

In Chapter 2 the basics of random variables and probability distributions are presented and the Law of Large Numbers is used to highlight the differences between expectation values and sample averages.

Four of the most common probability distributions are introduce in Chapter 3 and in Chapter 4 the central limit theorem and systematic errors are discussed so that the discussions to follow can be restricted without losing too much generality. Chapter 5 introduces the idea of correlation in the random errors associated with pairs of random variables.

Chapter 6 provides *Propagation of Error* formulas for determining the uncertainty in variables defined from other variables. Chapter 7 discusses the Principal of Maximum Likelihood and its implications regarding the sample mean and sample variance. Chapter 8 covers *Regression Analysis* for comparing measurements with independent theoretical predictions and determining fitting parameters and their uncertainties.

Chapter 9 discusses evaluation of regression results and the chi-square random variable. Typically used to evaluate the "goodness of fit," chi-square is a measure of the difference between experiment and theoretical predictions. The chi-square test and other methods are presented for checking if those differences are reasonable in relation to the uncertainties involved.

Chapter 10 provides a guide to using Excel for linear and nonlinear regression.

Chapter 2

Random Variables

The experimental model treats each measurement as a *random variable*—a numerical quantity having a value which varies randomly as the procedure used to obtain it is repeated. Each possible value for a random variable occurs with a fixed probability as described next.

When the possible outcomes are discrete, their probabilities are governed by a *discrete probability function* or dpf. For example, the number of clicks from a Geiger counter over some time interval is limited to the discrete set of nonnegative integers. Under unchanging conditions, each possible value occurs with a probability given by the Poisson dpf, which is discussed in more detail shortly. A dpf is the complete set of values of $P(y_i)$ for all possible y_i, where each $P(y_i)$ gives the probability for that y_i to occur.

When the possible outcomes cover a continuous interval, their probabilities are governed by a *probability density function* or pdf as follows. With the pdf $p(y)$ specified for all values y in the range of possible outcomes, the differential probability $dP(y)$ of an outcome between y and $y + dy$ is given by

$$dP(y) = p(y)dy \tag{2.1}$$

Probabilities for outcomes in any finite range are obtained by integration. The probability of an outcome between y_1 and y_2 is given by

$$P(y_1 < y < y_2) = \int_{y_1}^{y_2} p(y)\,dy \tag{2.2}$$

Both discrete probability functions and probability density functions are referred to as probability distributions.

Continuous probability distributions become effectively discrete when the variable is recorded with a chosen number of *significant digits.* The probability of the measurement is then the integral of the pdf over a range ±1/2 of the size of the least significant digit.

$$P(y) = \int_{y-\Delta y/2}^{y+\Delta y/2} p(y')\, dy' \tag{2.3}$$

For example, a current I recorded to the nearest hundredth of an ampere, say 1.21 A, has $\Delta I = 0.01$ A and its probability of occurrence is the integral of its (as yet unspecified) pdf $p(I)$ over the interval from $I = 1.205$ to 1.215 A. Note how the values of $P(y)$ for a complete set of non-overlapping intervals covering the entire range of y-values would map the pdf into an associated dpf.

Many statistical analysis procedures will be based on the assumption that $P(y)$ is proportional to $p(y)$. For this to be the case, Δy must be small compared to the range of the distribution. More specifically, $p(y)$ must have little curvature over the integration limits so that the integral becomes

$$P(y) = p(y)\, \Delta y \tag{2.4}$$

Law of Large Numbers

$P(y)$ for an unknown distribution can be determined to any degree of accuracy by histogramming a sample of sufficient size.

For a discrete probability distribution, the histogram bins should be labeled by the allowed values y_j. For a continuous probability distribution, the bins should be labeled by their midpoints y_j and constructed as adjacent, non-overlapping intervals spaced Δy apart and covering the complete range of possible outcomes. The sample, of size N, is then sorted to find the frequencies $f(y_j)$ for each bin

The law of large numbers states that the sample probability $f(y_j)/N$ for any bin will approach the predicted $P(y_j)$ more and more closely as the sample size increases. The limit satisfies

$$P(y_j) = \lim_{N\to\infty} \frac{1}{N} f(y_j) \tag{2.5}$$

Sample averages and expectation values

Let y_i, $i = 1..N$ represent sample values for a random variable y having probabilities of occurrence governed by a pdf $p(y)$ or a dpf $P(y)$. The *sample average* of any function $g(y)$ will be denoted with an overline so that $\overline{g(y)}$ is defined as the value of $g(y)$ averaged over all y-values in the sample set.

$$\overline{g(y)} = \frac{1}{N}\sum_{i=1}^{N} g(y_i) \tag{2.6}$$

For the function $g(y) = y$, application of Eq. 2.6 represents simple averaging of the y-values

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.7}$$

$\bar{y}$ is called the *sample mean*.

Note that $\bar{y}$, or the sample average of any function, is a random variable; taking a new sample set would produce a different value. However, in the limit of infinite sample size, the law of large numbers asserts that the average defined by Eq. 2.6 converges to a well defined constant depending only on the probability distribution and the function $g(y)$. This constant is called the *expectation value* of $g(y)$ and will be denoted by putting angle brackets around the function

$$\langle g(y)\rangle = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} g(y_i) \tag{2.8}$$

Equation 2.8 emphasizes the role of expectation values as "expected averages," or "true means" or simply "means" of $g(y)$. However, as this equation requires an infinite sample size, it is not directly useful for calculating expectation values.

Equation 2.8 can be cast into a form suitable for use with a known probability distribution as follows. Assume a large sample of size N has been properly histogrammed. If the variable is discrete, each possible value y_j gets its own bin. If the variable is continuous, the bins are labeled by their midpoints y_j and their size Δy has been chosen small enough to ensure that (1) the probability for a y-value to occur in any particular bin will be accurately given by $P(y_j) = p(y_j)\Delta y$ and (2) all y_i sorted into a bin at y_j can be considered as contributing $g(y_j)$— rather than $g(y_i)$—to the sum in Eq. 2.8.

After sorting the sample y_i-values into the bins, thereby finding the frequencies of occurrence $f(y_j)$ for each bin, the sum in Eq. 2.8 can be grouped by bins and becomes

$$\langle g(y)\rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{\text{all } y_j} g(y_j) f(y_j) \tag{2.9}$$

Note the change from a sum over all samples in Eq. 2.8 to a sum over all histogram bins in Eq. 2.9.

Moving the limit and factor of $1/N$ inside the sum, Eq. 2.5 can be used in Eq. 2.9 giving:

$$\langle g(y)\rangle = \sum_{\text{all } y_j} g(y_j)\, P(y_j) \tag{2.10}$$

Eq. 2.10 is a weighted average; each value of $g(y_j)$ in the sum is weighted by the probability of its occurrence $P(y_j)$.

Eq. 2.10 is directly applicable to discrete probability functions. For a continuous probability density function, $P(y_j) = p(y_j)\Delta y$. Making this substitution in Eq. 2.10 and then taking the limit as $\Delta y \to 0$ converts the sum to an integral and gives

$$\langle g(y)\rangle = \int_{-\infty}^{\infty} g(y)p(y)\, dy \tag{2.11}$$

Eq. 2.11 is a weighted integral with each $g(y)$ weighted by its occurrence probability $p(y)\, dy$.

Properties of expectation values

Some frequently used properties of expectation values are given below. They all follow from simple substitutions for $g(y)$ in Eqs. 2.10 or 2.11 or from the operational definition of an expectation value as an average for an effectively infinite data set (Eq. 2.8).

1. The expectation value of a constant is that constant: $\langle c\rangle = c$. Substitute $g(y) = c$ and use normalization condition. Guaranteed because the value c is averaged for every sampled y_i.

2. Constants can be factored out of expectation value brackets: $\langle cu(y)\rangle = c\,\langle u(y)\rangle$. Substitute $g(y) = cu(y)$, where c is a constant. Guaranteed by

the distributive property of multiplication over addition for the terms involved in the averaging.

3. The expectation value of a sum of terms is the sum of the expectation value of each term: $\langle u(y)+v(y)\rangle = \langle u(y)\rangle + \langle v(y)\rangle$. Substitute $g(y) = u(y)+v(y)$. Guaranteed by the associative property of addition for the terms involved in the averaging.

But also keep in mind the non-rule: The expectation value of a product is not necessarily the product of the expectation values: $\langle u(y)v(y)\rangle \neq \langle u(y)\rangle \langle v(y)\rangle$. Substituting $g(y) = u(y)v(y)$ does not, in general, lead to $\langle u(y)v(y)\rangle = \langle u(y)\rangle \langle v(y)\rangle$.

These properties will be put to use repeatedly. In the next section, they are used to get basic relationships involving parameters of any probability distribution.

Normalization, mean and variance

Probability distributions are defined so that their sum or integral over any range of possible values gives the probability for an outcome in that range. Consequently, if the range includes all possible values, the probability of an outcome in that range is 100% and the sum or integral must be equal to one. For a discrete probability distribution this *normalization* condition reads:

$$\sum_{\text{all } y_j} P(y_j) = 1 \tag{2.12}$$

and for a continuous probability distribution it becomes

$$\int_{-\infty}^{\infty} p(y)\, dy = 1 \tag{2.13}$$

The normalization sum or integral is also called the zeroth moment of the probability distribution—as it is the expectation value of y^0. The other two most important expectation values of a distribution are also moments of the distribution.

The *mean* μ_y of a probability distribution is defined as the expectation value of y itself, that is, of y^1. It is the first moment of the distribution.

$$\mu_y = \langle y\rangle \tag{2.14}$$

The mean is a measure of the central value of the distribution.

The sample mean—$\bar{y}$ of Eq. 2.7—is an estimate of the true mean. It becomes a better estimate as N increases and the two become equal as $N \to \infty$. How closely $\bar{y}$ and μ_y should agree with one another for finite N is discussed in Chapter 7. Here we would like to point out a related feature. Taking the expectation value of both sides of Eq. 2.7 and noting $\langle y_i \rangle = \mu_y$ for all N samples gives

$$\langle \bar{y} \rangle = \mu_y \tag{2.15}$$

thereby demonstrating that the expectation value of the sample mean is equal to the true mean.

> Any parameter estimate having an expectation value equal to the parameter it is estimating is said to be an *unbiased estimate*; it will give the true parameter value "on average."

Thus, the sample mean is an unbiased estimate of the true mean.

Defining $y - \mu_y$ as the *deviation* in a random variable's value from its mean, Eq. 2.14 can be rewritten

$$\langle y - \mu_y \rangle = 0 \tag{2.16}$$

showing that for any distribution, by definition, the mean deviation is zero. The sample y-value can be above or below the mean and so deviations can be positive or negative and have a mean of zero. If one is trying to describe the size of typical deviations, the mean deviation is unsuitable as it is always zero.

The *mean absolute deviation* would be one possible choice. Defined as the expectation value $\langle |y - \mu_y| \rangle$, the mean absolute deviation for a random variable y would be nonzero and a reasonable measure of the expected deviations. However, the mean absolute deviation does not arise naturally when formulating the basic statistical procedures considered here, whereas the *mean squared deviation* plays a central role. Consequently, the standard measure of a deviation, i.e., the *standard deviation* σ_y, is taken as the square root of the mean squared deviation.

The mean squared deviation is also called the *variance* and written σ_y^2 for a random variable y. It is the second moment about the mean and defined as the following expectation value

$$\sigma_y^2 = \left\langle (y - \mu_y)^2 \right\rangle \tag{2.17}$$

The variance has units of y^2. Its square root, the standard deviation σ_y has the same units as y and is a measure of the width of the distribution.

Expanding the right side of Eq. 2.17 gives $\sigma_y^2 = \left\langle y^2 - 2y\mu_y + \mu_y^2 \right\rangle$ and then taking expectation values term by term, noting μ_y is a constant and $\langle y \rangle = \mu_y$, gives:

$$\sigma_y^2 = \left\langle y^2 \right\rangle - \mu_y^2 \tag{2.18}$$

This equation is useful for evaluating the variance of a given probability distribution and in the form

$$\left\langle y^2 \right\rangle = \mu_y^2 + \sigma_y^2 \tag{2.19}$$

shows that the expectation value of y^2 (the second moment about the origin) exceeds the square of the mean by the variance.

The *sample variance* would then be given by Eq. 2.6 with $g(y) = (y - \mu_y)^2$. It will be denoted s_y^2 and thus defined by

$$s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \tag{2.20}$$

Taking the expectation value of this equation shows the sample variance is an unbiased estimate of the true variance.

$$\left\langle s_y^2 \right\rangle = \sigma_y^2 \tag{2.21}$$

The proof this time requiring an application of Eq. 2.17 to each term in the sum.

Typically, μ_y is not known and Eq. 2.20 can not be used to get an estimate of an unknown variance. Can the sample mean $\bar{y}$ be used in its place? Yes, but making this substitution requires the following minor modification to Eq. 2.20.

$$s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{2.22}$$

As will be proven later, the denominator must be reduced by one so that this sample variance will also be unbiased.

The sample mean and sample variance are random variables and each follows its own probability distribution. The fact that they are unbiased means that the means of their distributions will be the true mean and true variance, respectively. Other details of these two distributions, such as their widths will be discussed later.

Chapter 3

Probability Distributions

In this section, definitions and properties of a few fundamental probability distributions will be discussed.

The Gaussian distribution

The *Gaussian* or *normal* probability density function has the form

$$p(y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right] \tag{3.1}$$

and is parameterized by two quantities: the mean μ_y and the standard deviation σ_y.

Figure 3.1 shows the Gaussian pdf and gives various integral probabilities. Because of its form, probabilities can always be described relative to the mean and standard deviation. There is a 68% probability that a Gaussian random variable will be within one standard deviation of the mean, 95% probability it will be within two, and a 99.7% probability it will be within three. These "1-sigma," "2-sigma," and "3-sigma" probabilities should be committed to memory. A more complete listing can be found in Table 10.2.

The binomial distribution

The binomial distribution arises when a random event, called a *Bernoulli trial*, can be considered to have only two outcomes. One outcome is termed

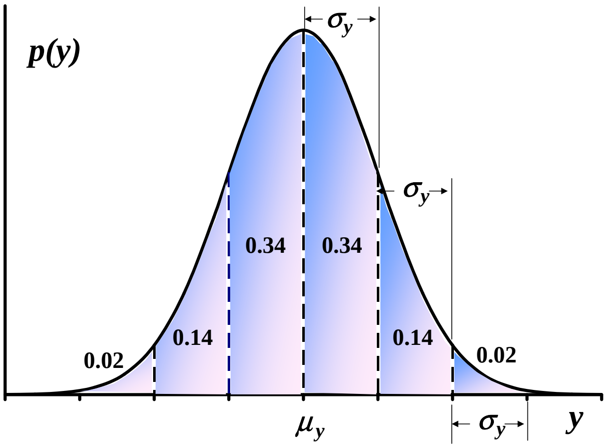


Figure 3.1: The Gaussian distribution labeled with the mean μ_y, the standard deviation σ_y and some areas, i.e., probabilities.

a success and occurs with a probability p. The other, termed a failure, occurs with a probability $1-p$. Then, with N Bernoulli trials, the number of successes n can be any integer from zero (none of the N trials were a success) to N (all trials were successes).

The probability of n successes (and thus $N-n$ failures) is given by the binomial distribution

$$P(n) = \frac{N!}{n!(N-n)!}\, p^n(1-p)^{N-n} \tag{3.2}$$

The probability $p^n(1-p)^{N-n}$ would be the probability that the first n trials were successes and the last $N-n$ were not. Since the n successes and $N-n$ failures can occur in any order and each distinct ordering would occur with this probability, the extra multiplicative factor, called the binomial coefficient, is needed to count the number of distinct orderings.

The most common application of the binomial distribution is associated with the construction of sample frequency distributions. The frequency in each histogram bin is governed by the binomial probability distribution. A particular bin at y_j represents a particular outcome or range of outcomes and has an associated probability $P(y_j)$. Each Bernoulli trial consists of taking one new sample and either sorting it into that bin (a success with a probability $P(y_j)$) or not (a failure with a probability $1-P(y_j)$). After N

samples, the number of successes (the bin frequency) should follow a binomial distribution for that N and $p = P(y_j)$.

The Poisson distribution

Poisson-distributed variables arise in particle and photon counting experiments. For example, under unchanging conditions and averaged over long times, the number of clicks y from a Geiger counter due to natural background radiation might consistently give an average of, say, one tick per second. However, over any 10-second interval while an average of 10 ticks is expected, more or less ticks are also possible.

More specifically, if μ_y is the average number expected in an interval, then values of y around μ_y will be the most likely, but all integers zero or larger are theoretically possible. Values of y can be shown to occur with probabilities governed by the Poisson distribution.

$$P(y) = e^{-\mu}\frac{\mu^y}{y!} \tag{3.3}$$

For the Poisson distribution, one can show that the parent variance satisfies

$$\sigma_y^2 = \mu_y \tag{3.4}$$

For large values of μ_y, the Poisson probability for a given y is very nearly Gaussian—given by Eq. 2.1 with $\Delta y = 1$ and $p(y)$ given by Eq. 3.1 (with $\sigma_y^2 = \mu_y$). That is,

$$P(y) \approx \frac{1}{\sqrt{2\pi\mu_y}} \exp\left[-\frac{(y-\mu_y)^2}{2\mu_y}\right] \tag{3.5}$$

Eqs. 3.4 and 3.5 are the origin of the commonly accepted practice of applying "square root statistics" or "counting statistics," whereby Poisson-distributed variables are treated as Gaussian-distributed variables with a variance chosen to be μ_y or some estimate of μ_y.

One common application of counting statistics arises when a single count is measured from a Poisson distribution of unknown mean and observed to take on a particular value y. With no additional information, that measured y-value becomes an estimate of μ_y and thus it also becomes an estimate of the

	binomial	Poisson	uniform	Gaussian
form	$P(n) = \dfrac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$	$P(n) = e^{-\mu}\dfrac{\mu^n}{n!}$	$p(y) = \dfrac{1}{\lvert b-a\rvert}$	$p(y) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\dfrac{(y-\mu)^2}{2\sigma^2}\right]$
mean	Np	μ	$(a+b)/2$	μ
variance	$Np(1-p)$	μ	$(b-a)^2/12$	σ^2

Table 3.1: Common probability distributions with their means and variances.

variance of its own parent distribution. That is, y is assumed to be governed by a Gaussian distribution with a standard deviation given by

$$\sigma_y = \sqrt{y} \tag{3.6}$$

Counting statistics is a good approximation for large values of y—greater than about 30. Using it for values of y below 10 or so can lead to significant errors in analysis.

The uniform distribution

The uniform probability distribution arises, for example, when using digital metering. One might assume a reading of 3.72 V on a 3-digit, digital voltmeter implies the underlying variable is equally likely to be any value in the range 3.715 to 3.725 V. A variable with a constant probability in the range from a to b has a pdf given by

$$p(y) = \frac{1}{|b-a|} \tag{3.7}$$

Exercise 1 *(a) Use a software package to generate random samples from a Gaussian distribution with a mean $\mu_y = 0.5$ and a standard deviation $\sigma_y = 0.05$. Use a large sample size N and well-chosen bins (make sure one bin is exactly centered at 0.5) to create a reasonably smooth, bell-shaped histogram of the sample frequencies vs. the bin centers.*
(b) Consider the histogramming process with respect to the single bin at the center of the distribution—at μ_y. Explain why the probability for a sample to fall in that bin is approximately $\Delta y/\sqrt{2\pi\sigma_y^2}$, where Δy is the bin size, and

use it with your sample size to predict the mean and standard deviation for that bin's frequency. Compare your actual sample frequency at μ_y with this prediction. Is the difference between them reasonable?

Exercise 2 *Eqs. 2.14 and 2.17 provide the definitions of the mean μ and variance σ^2 with Eqs. 2.10 or 2.11 used for their evaluation. Show that the means and variances of the various probability distributions are as given in Table 3.1. Also show that they satisfy the normalization condition.*

Do not use integral tables. Do the normalization sum or integral first, then the mean, then the variance. The earlier results can often be used in the later calculations.

For the Poisson distribution, evaluation of the mean should thereby demonstrate that the parameter μ appearing in the distribution is, in fact, the mean. For the Gaussian, evaluation of the mean and variance should thereby demonstrate that the parameters μ and σ^2 appearing in the distribution are, in fact, the mean and variance.

Hints: For the binomial distribution you may need the expansion

$$(a+b)^N = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!} a^n b^{N-n} \tag{3.8}$$

For the Poisson distribution you may need the power series expansion

$$e^a = \sum_{n=0}^{\infty} \frac{a^n}{n!} \tag{3.9}$$

For the Gaussian distribution be sure to always start by eliminating the mean (with the substitution $y' = y - \mu_y$). The evaluation of the normalization integral $I = \int_{-\infty}^{\infty} p(y)\,dy$ is most readily done by first evaluating the square of the integral with one of the integrals using the dummy variable x and the other using y. (Both pdfs would use the same μ and σ.) That is, evaluate

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y)\,dx\,dy$$

and then take its square root. To evaluate the double integral, first eliminate the mean and then convert from cartesian coordinates x' and y' to cylindrical coordinates r and θ satisfying $x' = r\cos\theta$, $y' = r\sin\theta$. Convert the area element $dx'\,dy' = r\,dr\,d\theta$, and set the limits of integration for r from 0 to ∞ and for θ from 0 to 2π.

Chapter 4

Measurement Model

This section presents an idealized model for measurements, defining in more detail the ideas behind random and systematic errors.

Central limit theorem

While it would be useful to know the shape of the probability distributions for all random variables occurring in an analysis, taking large enough samples to get such information is not often feasible. The *central limit theorem* asserts that with sufficiently large data sets, detailed information about the shape of the distributions is overkill; the mean and variance are often the only parameters that will survive the analysis.

Specifically, the central limit theorem says that the sum of a sufficiently large number of random variables will follow a Gaussian distribution having a mean equal to the sum of the means of each variable in the sum and having a variance equal to the sum of the variances of each variable in the sum. Moreover, the individual variables can follow just about any probability distribution. They do not have to be Gaussian distributed.

The central limit theorem can be taken a step further. Formulas such as those associated with regression analysis will soon be derived based on the assumption that the input variables are governed by Gaussian distributions. A loose interpretation of the central limit theorem suggests that for data sets that are large enough, these formulas will be valid even if the data are governed by non-Gaussian distributions. The trick is to simply use the standard deviation of the particular non-Gaussian distribution involved for the corresponding standard deviation in the assumed Gaussian distribution.

Exercise 3 *(a) Predict the mean and standard deviation of the sum of 12 uniformly distributed random numbers on the interval* $(0, 1)$*.*
(b) Create 1000 samples of such 12-number sums and submit a histogram of the frequency distribution. Overlay the histogram with a smooth curve giving the predicted frequencies based on the central limit theorem and comment on the comparison.
(c) Evaluate the sample mean (Eq. 2.7) and the sample variance (Eq. 2.22) and comment on their agreement with predictions. The sample mean and the sample variance are random variables. Determining how closely they should match the predictions of the central limit theorem, which only refers to parent distributions and expectation values, requires the probability distributions associated with these random variables. These distributions depend on the sample size and will be discussed shortly. For $N = 1000$*, about 95% of the time, the sample mean should be within* ± 0.06 *of the true mean and the sample variance should be within* ± 0.09 *of the true variance.*

Random errors

A measurement y can be expressed as the sum of the mean of its probability distribution μ_y and a *random error* δ_y that scatters individual measurements both above and below the mean.

$$y = \mu_y + \delta_y \tag{4.1}$$

The quantity $\delta_y = y - \mu_y$ is also called the *deviation.*

Whenever possible, the experimentalist should supply an estimate of the standard deviation. A ± notation is often used. A rod length recorded as 2.64 ± 0.02 cm indicates a sample value $y = 2.64$ cm and a standard deviation $\sigma_y = 0.02$ cm.

One method for estimating standard deviations is to take a large sample for one particular measured variable while experimental conditions remain constant. The resulting sample standard deviation might then be assumed to be the σ_y for all future measurements of the same kind. Or, an estimate of σ_y might be based on instrument scales or other information about the measurement. The experimenter's confidence in the values assigned for σ_y

will determine the confidence that should be placed on later comparisons of that data with theoretical predictions.

Although they are often only approximately known, the σ_y entering into an analysis will be assumed exactly known. Issues associated with uncertainty in σ_y will only be considered after first exploring the results that can be expected when this quantity is completely certain.

Systematic Errors

In contrast to random errors which cause measurement values to differ randomly from the mean of the measurement's parent distribution, systematic errors cause the mean of the parent distribution to differ systematically (nonrandomly) from the true physical quantity the mean is interpreted to represent. With y_t representing this true value and $\delta_{\rm sys}$ the systematic error, this can be expressed

$$\mu_y = y_t + \delta_{\rm sys} \tag{4.2}$$

Sometimes $\delta_{\rm sys}$ is constant as y_t varies. In such cases, it is called an offset or zeroing error and μ_y will always be above or below the true value by the same amount. Sometimes $\delta_{\rm sys}$ is proportional to y_t and it is then referred to as a scaling or gain error. For scaling errors, μ_y will always be above or below the true value by the same fractional amount, e.g., always 10% high. In some cases, $\delta_{\rm sys}$ is a combination of an offset and a scaling error. Or, $\delta_{\rm sys}$ might vary in some arbitrary manner. The procedures to minimize systematic errors are called *calibrations*, and their design requires careful consideration of the particular instrument and its application.

Combining Eqs. 4.1 and 4.2

$$y = y_t + \delta_y + \delta_{\rm sys} \tag{4.3}$$

demonstrates that both random and systematic errors contribute to every measurement. Both can be made smaller but neither can ever be entirely eliminated. *Accuracy* refers to the size of possible systematic errors while *precision* refers to the size of possible random errors.

Statistical analysis procedures deal with the effects of random errors only. Thus, systematic errors are often neglected in the first round of data analysis in which results and their uncertainties are obtained taking into account random error only. Then, one examines how the measurement means might deviate non-randomly from the true physical quantities and one determines how such deviations would change those results. If the changes are found to be small compared to the uncertainties determined in the first round, systematic errors have been demonstrated to be inconsequential. If systematic errors could change results at a level comparable to or larger than the uncertainties determined in the first round, those changes should be reported separately or additional measurements (calibrations) should be made to reduce them.

Chapter 5

Independence and Correlation

Statistical procedures typically involve multiple random variables as input and produce multiple random variables as output. Probabilities associated with multiple random variables depend on whether the variables are statistically independent or not. Correlation describes a situation in which the deviations for two random variables are related. For statistically independent variables there is no expected correlation. The consequences of independence and correlation affect all manner of statistical analysis.

Independence

Two events are statistically independent if knowing the outcome of one has no effect on the outcomes of the other. For example, if you flip two coins, one in each hand, each hand is equally likely to hold a heads or a tails. Knowing that the right hand holds a heads, say, does not change the equal probability for heads or tails in the left hand. The two coin flips are independent.

Two events are statistically dependent if knowing the results of one affects the probabilities for the other. Consider a drawer containing two white socks and two black socks. You reach in without looking and pull out one sock in each hand. Each hand is equally likely to hold a black sock or a white sock. However, if the right hand is known to hold a black sock, say, the left hand is now twice as likely to hold a white sock as it is to hold a black sock. The two sock pulls are dependent.

The *unconditional probability* of event A, expressed by $\Pr(A)$, represents the probability of event A occurring without regard to any other events.

The *conditional probability* of "A given B," expressed $\Pr(A|B)$ represents the probability of event A occurring given that event B has occurred. Two events are statistically independent if and only if

$$\Pr(A|B) = \Pr(A) \tag{5.1}$$

The multiplication rule for joint probabilities follows from Eq. 5.1 and is more useful. The *joint probability* is the probability for both of two events to occur. The *multiplication rule* is that the joint probability for two independent events to occur is the product of the unconditional probability for each to occur.

Whether events are independent or not, the joint probability of "A and B," expressed $\Pr(A \cap B)$, is logically the equivalent of $\Pr(B)$, the unconditional probability of B occurring without regard to A, multiplied by the conditional probability of A given B.

$$\Pr(A \cap B) \quad = \quad \Pr(B)\,\Pr(A|B) \tag{5.2}$$

Then, substituting Eq. 5.1 gives the multiplication rule valid for independent events.

$$\Pr(A \cap B) = \Pr(A)\,\Pr(B) \tag{5.3}$$

And, of course, the roles of A and B can be interchanged in the logic or equations above.

Equation 5.3 states the commonly accepted principle that the probability for multiple independent events to occur is simply the product of the probability for each to occur.

For a random variable, an event can be defined as getting one particular value or getting within some range of values. Consistency with the multiplication rule for independent events then requires a *product rule* for the pdfs or dpfs governing the probabilities of independent random variables.

The *joint probability distribution* for two variables gives the probabilities for both variables to take on specific values. For independent, discrete random variables x and y governed by the dpfs $P_x(x)$ and $P_y(y)$, the joint probability $P(x, y)$ for values of x and y to occur is given by the product of each variable's probability

$$P(x, y) = P_x(x)P_y(y) \tag{5.4}$$

And for independent, continuous random variables x and y governed by the pdfs $p_x(x)$ and $p_y(y)$, the differential joint probability $dP(x,y)$ for x and y to be in the intervals from x to $x+dx$ and y to $y+dy$ is given by the product of each variable's probability

$$dP(x,y) = p_x(x)p_y(y)dx\,dy \tag{5.5}$$

The product rule for independent variables leads to the following important corollary. The expectation value of any function that can be expressed in the form $f_1(y_1)f_2(y_2)$ will satisfy

$$\langle f_1(y_1)f_2(y_2)\rangle = \langle f_1(y_1)\rangle\,\langle f_2(y_2)\rangle \tag{5.6}$$

if y_1 and y_2 are independent.

For discrete random variables the proof proceeds from Eq. 5.4 as follows:

$$\begin{aligned}
&\langle f_1(y_1)f_2(y_2)\rangle \\
&\quad = \sum_{\text{all } y_1,y_2} f_1(y_1)f_2(y_2)\,P(y_1,y_2) \\
&\quad = \sum_{\text{all } y_1}\sum_{\text{all } y_2} f_1(y_1)f_2(y_2)\,P_1(y_1)\,P_2(y_2) \\
&\quad = \sum_{\text{all } y_1} f_1(y_1)P_1(y_1) \sum_{\text{all } y_2} f_2(y_2)P_2(y_2) \\
&\quad = \langle f_1(y_1)\rangle\,\langle f_2(y_2)\rangle
\end{aligned} \tag{5.7}$$

And for continuous random variables it follows from Eq. 5.5:

$$\begin{aligned}
&\langle f_1(y_1)f_2(y_2)\rangle \\
&\quad = \int f_1(y_1)f_2(y_2)\,dP(y_1,y_2) \\
&\quad = \int\int f_1(y_1)f_2(y_2)p_1(y_1)p_2(y_2)\,dy_1\,dy_2 \\
&\quad = \int f_1(y_1)p_1(y_1)\,dy_1 \int f_2(y_2)p_2(y_2)\,dy_2 \\
&\quad = \langle f_1(y_1)\rangle\,\langle f_2(y_2)\rangle
\end{aligned} \tag{5.8}$$

A simple example of Eq. 5.6 is for the expectation value of the product of two independent variables, y_1 and y_2; $\langle y_1y_2\rangle = \langle y_1\rangle\,\langle y_2\rangle = \mu_1\mu_2$. For independent samples y_i and y_j both from the same distribution (having a mean μ_y and standard deviation σ_y) this becomes $\langle y_i y_j\rangle = \mu_y^2$ for $i \neq j$. Coupling this result with Eq. 2.19 for the expectation value of the square of any y-value: $\langle y_i^2\rangle = \mu_y^2 + \sigma_y^2$, gives the following relationship for independent variables from the same distribution

$$\langle y_i y_j\rangle = \mu_y^2 + \sigma_y^2\delta_{ij} \tag{5.9}$$

where δ_{ij} is the Kronecker delta function: equal to 1 if $i = j$ and zero if $i \neq j$.

A related corollary arises from Eq. 5.6 with the substitutions: $f_1(y_1) = y_1 - \mu_1$ and $f_2(y_2) = y_2 - \mu_2$ where y_1 and y_2 are independent random variables

$$\langle (y_1 - \mu_1)(y_2 - \mu_2)\rangle = \langle y_1 - \mu_1\rangle \langle y_2 - \mu_2\rangle \tag{5.10}$$

Here μ_1 and μ_2 are the means of y_1 and y_2, and satisfy $\langle y_i - \mu_i\rangle = 0$. Thus the right-hand side of Eq. 5.10 is the product of two zeros and demonstrates that

$$\langle (y_1 - \mu_1)(y_2 - \mu_2)\rangle = 0 \tag{5.11}$$

for independent variables.

Note that both $y_1 - \mu_1$ and $y_2 - \mu_2$ always have an expectation value of zero whether or not y_1 and y_2 are independent. However, the expectation value of their product is guaranteed to be zero only if y_1 and y_2 are independent.

The product rule can be extended—by repeated multiplication—to any number of independent random variables. The explicit form for the joint probability for a data set y_i, $i = 1...N$ will be useful for our later treatment of regression analysis. This form will depend on the particular probability distributions for the y_i. Most lab data can be modeled on either the Poisson or Gaussian probability distributions and lead to the relatively simple expressions considered next.

For N independent Gaussian random variables, with each y_i having its own mean μ_i and standard deviation σ_i, the joint probability distribution becomes the following product of terms each having the form of Eq. 2.1 with $p(y_i)$ having the Gaussian form of Eq. 3.1.

$$P(\{y\}) = \prod_{i=1}^{N} \frac{\Delta y_i}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right] \tag{5.12}$$

where $\{y\}$ represents the complete set of y_i, $i = 1...N$ and Δy_i represents the size of the least significant digit in y_i, which are assumed small compared to σ_i.

For N independent random variables, each governed by its own Poisson distribution (with mean μ_i), the joint probability distribution becomes the following product of terms each having the form of Eq. 3.3.

$$P(\{y\}) = \prod_{i=1}^{N} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{5.13}$$

The joint probability distributions of Eqs. 5.12 and 5.13 are the basis for regression analysis and produce amazingly similar expressions when applied to that problem.

Correlation

Correlation describes relationships between pairs of random variables that are not statistically independent. Statistically independent random variables are always uncorrelated.

The generic data set under consideration now consists of two random variables, x and y, say—always measured or otherwise determined in unison—so that a single sample consists of an x, y pair. They are sampled repeatedly to make an ordered set x_i, y_i, $i = 1..N$ taken under unchanging experimental conditions so that only random, but perhaps not independent, variations are expected.

Considered as separate sample sets: x_i, $i = 1..N$ and y_i, $i = 1..N$, two sample probability distributions could be created—one for each set. The sample means $\bar{x}$ and $\bar{y}$ and the sample variances s_x^2 and s_y^2 could be calculated and would be best estimates for the means μ_x and μ_y and variances σ_x^2 and σ_y^2 for each variable's parent distribution $p_x(x)$ and $p_y(y)$. These sample and parent distributions would be considered unconditional because they provide probabilities without regard to the other variable's values.

The first look at the variables as pairs is typically with a *scatter plot*, in which the N values of (x_i, y_i) are represented as points on a graph. Figure 5.1 shows five different 1000- point samples of pairs of random variables. The set on the left is uncorrelated and the other four are correlated. The unconditional parent pdfs, $p_x(x)$ and $p_y(y)$, are the same for all five, namely Gaussian distributions having the parameters: $\mu_x = 4$, $\sigma_x = 0.1$ and $\mu_y = 14$, $\sigma_y = 1$. Even though the unconditional pdfs are the same, the scatter plots clearly show that the joint probability distributions are different and depend on the degree and sign of the correlation.

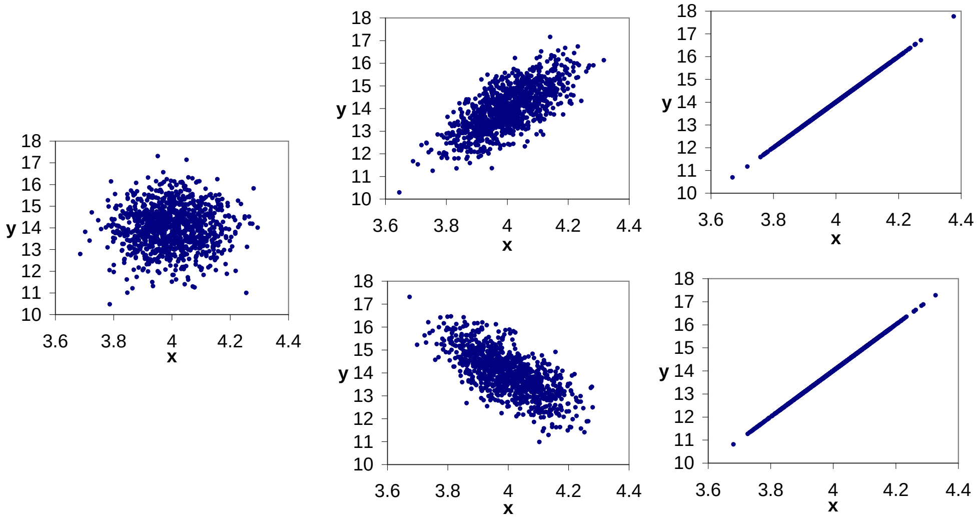


Figure 5.1: The behavior of uncorrelated and correlated Gaussian random variables. The leftmost figure shows uncorrelated variables, the middle two show partial correlation and the two on the right show total correlation. The upper two show positive correlations while the lower two show negative correlations. The Excel spreadsheet Correlated RV.xls on the lab website shows how these correlated random variables were generated.

The leftmost plot shows the case where the variables are independent and thus uncorrelated. The probability for a given x is then independent of the value of y. For example, if only those points within some narrow slice in y, say around $y = 15$, are analyzed—thereby making them conditional on that value of y, the values of x for that slice are, as in the unconditional case, just as likely to be above μ_x as below it.

For the four correlated cases, selecting different slices in one variable will give different conditional probabilities for the other variable. In particular, the conditional mean goes up or down as the slice moves up or down in the other variable. The top two plots show positively correlated variables, the bottom two show negatively correlated variables. For positive correlation, the variables are more likely to be on the same side of their means; when one variable is above (or below) its mean, the other is more likely to be above (or below) its mean. The conditional mean of one variable increases for slices at increasing values for the other variable. For negative correlation, these dependencies reverse. The variables are more likely to be on opposite sides of their means.

The degree of correlation determines the strictness of the dependence between the two variable's random deviations. For no correlation, knowing the value of x gives no information about the value of y. At the other extreme, the variables lie on a perfect line and the value of x completely determines the value of y. In between, the conditional mean of the y-variable is linearly related to the value of the x-variable, but y-values still have random variations of their own—although with a standard deviation that is smaller than for the unconditional case.

The standard measure of correlation between two variables x and y is the *sample covariance* s_{xy}, defined

$$s_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{5.14}$$

The true covariance σ_{xy} is defined as the sample covariance in the limit of infinite sample size

$$\sigma_{xy} = \lim_{N\to\infty}\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{5.15}$$

or equivalently as the expectation value

$$\sigma_{xy} = \langle (x-\mu_x)(y-\mu_y)\rangle \tag{5.16}$$

With positive, negative, or no correlation, σ_{xy} will be positive, negative or zero. To see how the sign of the correlation predicts the sign of the covariance, consider the relative number of x_i, y_i data points that will produce positive vs. negative values for the product $(x_i - \mu_x)(y_i - \mu_y)$. This product is positive when both $x_i - \mu_x$ and $y_i - \mu_y$ have the same sign and it is negative when they have opposite signs. With positive correlation, there are more points with a positive product and thus the covariance is positive. With negative correlation, there are more points with a negative product and the covariance is negative. And with no correlation, there should be equal numbers with either sign and the covariance is zero.

The covariance σ_{xy} is limited by the size of σ_x and σ_y. The Cauchy-Schwarz inequality says it can vary between

$$-\sigma_x\sigma_y \le \sigma_{xy} \le \sigma_x\sigma_y \tag{5.17}$$

Thus, σ_{xy} is also often written

$$\sigma_{xy} = \rho\sigma_x\sigma_y \tag{5.18}$$

where ρ, called the *correlation coefficient,* is between -1 and 1. Correlation coefficients at the two extremes represent perfect correlation where x and y follow a linear relation exactly. The correlation coefficients used to generate Fig. 5.1 were 0, ±0.7 and ±1.

The inequality expressed by Eq. 5.17 is also true for the sample standard deviations and the sample covariance with the substitution of s_x, s_y and s_{xy} for σ_x, σ_y and σ_{xy}. The sample correlation coefficient r is then defined $s_{xy} = rs_xs_y$ and also varies between -1 and 1.

Of course, a sample correlation coefficient from a particular data set is a random variable. Its probability distribution depends on the true correlation coefficient and the sample size and is of interest, for example, when looking for evidence of any correlation, even a weak one, between two variables. A sample covariance near zero may be consistent with the assumption that the variables are uncorrelated. A value too far from zero, however, might be too improbable under this assumption thereby implying a correlation exists. These kinds of probabilities are not commonly needed in physics experiments and will not be discussed.

The covariance matrix

The *covariance matrix* denoted $[\sigma]$ describes all the variances and covariances possible between two or more variables. For a set of 3 variables $\{y\} = y_1, y_2, y_3$, it would be

$$[\sigma_y] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \tag{5.19}$$

with the extension to more variables obvious. Note that $\sigma_{11} = \sigma_1^2$ is the variance of y_1 with similar relations for σ_{22} and σ_{33}.

Thus the covariance matrix for a set of variables is a shorthand way of describing all of the variables' standard deviations (or uncertainties) and the covariances (or correlations) between them.

If all variables are independent, the covariances are zero and the covariance matrix is diagonal and given by

$$[\sigma_y] = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} \tag{5.20}$$

When variables are independent, their joint probability distribution follows the product rule, which leads to Eq. 5.12 when they are all Gaussian. What replaces the product rule for variables that are known to be dependent—that have a covariance matrix with off-diagonal elements? No simple expression exists when the variables follow arbitrary unconditional distributions. However, for the important case where they are all Gaussian, the expression is quite elegant. Eq. 5.12 must be replaced by

$$P(\{y\}) = \frac{\prod_{i=1}^{N} \Delta y_i}{\sqrt{(2\pi)^N \left|[\sigma_y]\right|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{y}^T - \boldsymbol{\mu}^T\right)\left[\sigma_y^{-1}\right](\boldsymbol{y} - \boldsymbol{\mu})\right] \tag{5.21}$$

where $[\sigma_y]$ is the covariance matrix for the N variables, $|[\sigma_y]|$ is its determinant, and $\left[\sigma_y^{-1}\right]$ is its inverse. A vector/matrix notation has been used where $\boldsymbol{y}$ and $\boldsymbol{\mu}$ are column vectors of length N with elements given by y_i and μ_i, respectively, and $\boldsymbol{y}^T$ and $\boldsymbol{\mu}^T$ are transposes of these vectors, i.e., row vectors with the same elements. Normal vector-matrix multiplication rules apply so that the argument of the exponential is a scalar.

Note that, as it must, Eq. 5.21 reduces to Eq. 5.12 if the covariance matrix is diagonal.

Chapter 6

Propagation of Errors

Propagation of errors comes into play when calculating one or more quantities a_k, $k = 1..M$ based on one or more random variables y_i, $i = 1..N$. The a_k are to be determined according to M given functions of the y_i

$$a_k = f_k(y_1, y_2, ..., y_N) \tag{6.1}$$

For example, the random variables might be a measured voltage V across a circuit element and a measured current I passing through it. The calculated quantities might be the element's resistance $R = V/I$ and/or the power dissipated $P = IV$.

In the general case, the joint probability distribution for the input variables transforms to a joint probability distribution for the output variables. In the Box-Müller transformation, for example, y_1 and y_2 are uncorrelated and uniformly distributed on the interval $[0, 1]$. The two calculated quantities

$$\begin{aligned} a_1 &= \sqrt{-2\ln y_1}\sin 2\pi y_2 \\ a_2 &= \sqrt{-2\ln y_1}\cos 2\pi y_2 \end{aligned} \tag{6.2}$$

will then be uncorrelated Gaussian random variables, each with a mean of zero and a variance of one.

Propagation of errors refers to a very restricted case of transformations where the ranges for the input variables are small—small enough that Eq. 6.1 for each a_k would be well represented by a first-order Taylor series expansion about the means of the y_i. This is not the case for the Box-Müller transformation and these more general cases will not be considered further.

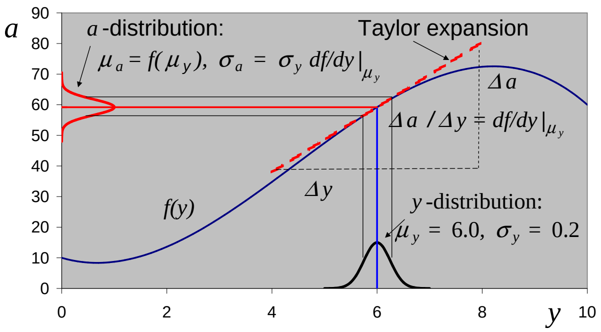


Figure 6.1: Single variable propagation of errors. Only the behavior of $f(y)$ over the region $\mu_y \pm 3\sigma_y$ affects the distribution in a.

To see how small errors lead to simplifications via a Taylor expansion, consider the case where there is only one calculated variable, a, derived from one random variable, y, according to a given function, $a = f(y)$. Figure 6.1 shows the situation where the standard deviation σ_y is small enough that for y-values in the range $\mu_y \pm 3\sigma_y$, $a = f(y)$ is well approximated by a straight line—the first order Taylor expansion of $f(y)$ about μ_y.

$$a = f(\mu_y) + \frac{df}{dy}(y - \mu_y) \tag{6.3}$$

where the derivative is evaluated at μ_y. With a linear relation between a and y, a Gaussian distribution in y will lead to a Gaussian distribution in a with $\mu_a = f(\mu_y)$ and $\sigma_a = \sigma_y|df/dy|$. Any second order term in the Taylor expansion—proportional to $(y - \mu_y)^2$—would lead to an asymmetry in the a-distribution and a bias in its mean. In Fig. 6.1, for example, $a = f(y)$ is always below the tangent line and thus the mean of the a-distribution will be slightly less than $f(\mu_y)$.

In treating the general case where there are several y_i involved in calculating several a_k, the y_i will be assumed to follow a Gaussian joint probability distribution—with or without correlation.

The sample set of y_i values together with their covariance matrix $[\sigma_y]$ are assumed given and will be used to determine the values for the a_k and their covariance matrix. The number of input and output variables is arbitrary. They need not be equal nor does one have to be more or less than the other.

The M sample a_k are easy to determine. They are evaluated according to the M functions expressed by Eq. 6.1 using the sample y_i values. Determining the covariance matrix for the a_k is not so simple. Before presenting the general case it is worth mentioning some special cases that are commonly encountered. Often all input variables are independent and thus terms involving their covariances can be dropped. Often only a single formula is to be evaluated and so there are no output covariances—only the single output variable's variance or standard deviation is sought. Often both conditions hold. These special cases will be considered after the general treatment is presented.

Just as the M functions f_k of Eq. 6.1 are used to provide sample a_k values from a given sample set of y_i, they should also be considered as giving the true values of the calculated quantities, denoted α_k, based on the true means μ_i of the y_i.

$$\alpha_k = f_k(\mu_1, \mu_2, ..., \mu_N) \tag{6.4}$$

Consider next the distribution of values obtained by repeatedly evaluating the a_k, each time from a newly sampled set of y_i. As additional sample sets are taken, the y_i would vary according to their joint probability distribution causing the values for the a_k to vary as well—according to their joint probability distribution. It is this distribution we seek.

Each y_i is in the range $\mu_i \pm 3\sigma_i$ better than 99% of the time. For the propagation of error formulas to be valid for all such y_i will then require that over such ranges, the functions a_k are accurately represented by a first-order Taylor expansions of f_k about the values $\mu_1, \mu_2, ..., \mu_N$.

$$\begin{aligned} a_k &= f_k(y_1, y_2, ..., y_N) \\ &= f_k(\mu_1, \mu_2, ..., \mu_N) + \\ &\quad \frac{\partial f_k}{\partial y_1}(y_1 - \mu_1) + \frac{\partial f_k}{\partial y_2}(y_2 - \mu_2) + \ ... + \frac{\partial f_k}{\partial y_N}(y_N - \mu_N) \\ &= \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i) \end{aligned} \tag{6.5}$$

where Eq. 6.4 has been used in the final step.

The partial derivatives are simply constants that should be evaluated at the expansion point, $\mu_1, \mu_2, ...\mu_N$. However, as these means are typically unknown, the derivatives will have to be evaluated at the measured point $y_1, y_2, ...y_N$ instead. This is not really an issue as all f_k are assumed linear over a range of several σ_i about each μ_i and thus the derivatives must be nearly constant for any y_i in that range.

Within the linear approximation, the means, variances, and covariances for the joint probability distribution for the a_k can easily be determined. The mean for the variable a_k is defined as $\langle a_k \rangle$ and this expectation value is easily evaluated from Eq. 6.5.

$$\begin{aligned} \langle a_k \rangle &= \left\langle \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i) \right\rangle \\ &= \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i} \langle (y_i - \mu_i) \rangle \\ &= \alpha_k \end{aligned} \tag{6.6}$$

where the expectation values $\langle y_i - \mu_i \rangle = 0$ has been used to eliminate all terms in the sum. This demonstrates the important result that the quantity $a_k = f_k(y_1, y_2, ..., y_M)$ will be unbiased estimates of the true α_k.

For notational convenience and to distinguish them from the standard deviations σ_i and covariances σ_{ij} for the input y_i, the standard deviations of the a_k will be denoted ζ_k and their covariances will be denoted ζ_{kl}. The variances defined by Eq. 2.17 and the covariances defined by Eq. 5.16 can be expressed by the single definition:

$$\zeta_{kl} = \langle (a_k - \alpha_k)(a_l - \alpha_l) \rangle \tag{6.7}$$

with the understanding that $\zeta_{kk} = \zeta_k^2 = \left\langle (a_k - \alpha_k)^2 \right\rangle$ is the variance of a_k.

Using Eq. 6.5 for a_k and a_l (with different dummy indexes) gives

$$\zeta_{kl} = \left\langle \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i) \sum_{j=1}^{N} \frac{\partial f_l}{\partial y_j}(y_j - \mu_j) \right\rangle \tag{6.8}$$

Rearranging the sums, distributing the expectation value over the terms in the sum and factoring constants from each term gives

$$\zeta_{kl} = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial f_k}{\partial y_i} \frac{\partial f_l}{\partial y_j} \langle (y_i - \mu_i)(y_j - \mu_j) \rangle \tag{6.9}$$

Now the simultaneous definitions for the input variances and covariances $\sigma_{ij} = \langle (y_i - \mu_i)(y_j - \mu_j) \rangle$ can be used to express ζ_{kl} as

$$\zeta_{kl} = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\partial f_k}{\partial y_i} \frac{\partial f_l}{\partial y_j} \sigma_{ij} \tag{6.10}$$

This equation gives the elements of the covariance matrix $[\zeta_a]$ associated with the full set of a_k. It can be seen to be just one element of the equation for the entire matrix, with the entire matrix elegantly given by:

$$[\zeta_a] = [J^T][\sigma_y][J] \tag{6.11}$$

where standard matrix multiplication rules apply and $[J]$ is the $N \times M$ *Jacobian* matrix with elements given by

$$[J]_{ik} = \frac{\partial f_k}{\partial y_i} \tag{6.12}$$

Here, $i = 1...N$ and $k = 1...M$ label the rows and columns, respectively, and $[J^T]$ is the transpose of $[J]$—an $M \times N$ matrix with the roles of rows and columns interchanged.

$$[J^T]_{ki} = [J]_{ik} \tag{6.13}$$

Now a few special cases are considered.

The first is not really a special case. It is simply a rewrite for a diagonal element of the covariance matrix. The quantity ζ_{kk}, i.e., the variance ζ_k^2 is especially important because its square root is the standard deviation or uncertainty of a_k. The rewrite will separate the terms involving the variances of the y_i (terms in the double sum where $i = j$) from those involving the covariances (cross terms where $i \neq j$).

$$\zeta_k^2 \;=\; \sum_{i=1}^{N} \left(\frac{\partial f_k}{\partial y_i} \right)^2 \sigma_i^2 + 2 \sum_{j>i=1}^{N} \frac{\partial f_k}{\partial y_i} \frac{\partial f_k}{\partial y_j} \sigma_{ij} \tag{6.14}$$

where the second sum is meant to represent a sum over all pairs i, j where $j > i$. The factor of 2 arises because Eq. 6.10 would produce two equivalent cross terms while the sum above is meant to include each cross term only once. Note that whenever correlated variables are used together as input to a calculation, the uncertainty in the calculated quantity will have to take into account the input covariances via this equation.

Now consider the case for the variance ζ_k^2 when all y_i are independent, i.e., their covariances σ_{ij}, $i \neq j$ are all zero. In this case, the formula simplifies to

$$\zeta_k^2 = \sum_{i=1}^{N} \left(\frac{\partial f_k}{\partial y_i} \right)^2 \sigma_i^2 \tag{6.15}$$

This is the typical propagation of error formula given in most references, but it should be understood to apply only to uncorrelated y_i.

Special conditions can lead to uncorrelated output variables. However, in general, the $\zeta_{kl}, k \neq l$ will be non-zero (and thus the a_k will be correlated) for both dependent or independent input y_i. For the special case where all y_i are independent, Eq. 6.10 simplifies to

$$\zeta_{kl} \;\;=\;\; \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i} \frac{\partial f_l}{\partial y_i} \sigma_i^2 \tag{6.16}$$

Exercise 4 *Simulate 1000 pairs of simultaneous measurements of a current I through a circuit element and the voltage V across it. Assume that the current and voltage measurements are independent. Take I-values from a Gaussian with a mean of 76 mA and a standard deviation of 3 mA, and take V-values from a Gaussian with a mean of 12.2 V and a standard deviation of 0.2 V.*

Calculate sample values for the element's resistance $R = V/I$ and power dissipated $P = IV$ for each pair of I and V and submit a scatter plot for the 1000 R, P sample values. Calculate the predicted means (Eq. 6.4) and variances (Eq. 6.15) for the R and P distributions and calculate their predicted covariance (Eq. 6.16). Evaluate the sample means (Eq. 2.7) for the 1000 R and P values, their sample variances (Eq. 2.22), and the sample covariance between R and P (Eq. 5.14).

To compare the predictions and the sample values determined above requires the probability distributions for sample means, sample variances, and sample covariances. These distributions will be discussed shortly but their standard deviations will be given here as an aid to the comparison. The standard deviation of the mean of the sample resistances is predicted to be $\sigma_{\bar{R}} = \sigma_R/\sqrt{N}$. Similarly for the power. Check if your two sample means agree with predictions at the 95% or two-sigma level. The fractional standard deviation—the standard deviation of a quantity divided by the mean of that quantity—for the two variances and the covariance are predicted to *be* $\sqrt{2/(N-1)}$. *For* $N = 1000$, *this is about 4.5% so check if your sample variances and covariance agree with predictions at the 9% or two-sigma level.*

Chapter 7

Principle of Maximum Likelihood

Theoretical models often include parameters that are not known in advance and must be estimated based on the measured data. A common example is a straight-line fit to a set of measured (x_i, y_i) data points predicted to obey a linear relationship: $y = mx + b$. Regression analysis then gives the best estimate of the slope m and the intercept b based on the data.

The *principle of maximum likelihood* says to choose parameter estimates so that they maximize the probability of the data set from which they are derived. Using the principle guarantees that if the experiment and theory are then deemed incompatible, they will be incompatible regardless of the parameter values. Any other values will only make the data less likely. Parameters determined by this principle will be called *best estimates*.

How is the principle of maximum likelihood implemented? The first step is to get an expression for the probability of the whole data set. For now, we assume that all random variables involved can be expressed by the set y_i, $i = 1...N$ and we assume they are all statistically independent so that the product rule applies. Independence is commonly the case and it greatly simplifies the math.

With all the μ_i and σ_i given, the joint probability distributions of Eq. 5.12 (for Gaussian variables) or Eq. 5.13 (for Poisson variables) can be considered as providing the probabilities for any set of y_i—larger for sets that are more likely and smaller for sets that are less likely.

For the purposes of estimating theoretical parameters, the joint probability takes on a second purpose. Theoretical models typically make some prediction about the means μ_i and/or the standard deviations σ_i appearing in the joint probability. The principle of maximum likelihood simply states that any adjustable theory parameters involved in those predictions should be chosen to maximize the joint probability.

Any function $f(a, b, ...)$ has an extremum with respect to variable a where its derivative with respect to a is zero: $\partial f/\partial a = 0$. For the purpose of finding the maximum probability, it is common to first take the natural logarithm of $f(a, b, ...)$ and maximize that. This works because $\partial(\ln f)/\partial a = (1/f)\partial f/\partial a$ and thus where one derivative is zero, so is the other.

The natural logarithm of the probability is called the *log likelihood* and simplifies the math because it transforms the products into sums which are easier to differentiate.

For a Gaussian data set (Eq. 5.12) it is

$$\ln(P) \quad = \quad -\frac{N}{2}\ln 2\pi + \sum_{i=1}^{N} \ln\left(\frac{\Delta y_i}{\sigma_i}\right) - \frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \mu_i)^2}{\sigma_i^2} \tag{7.1}$$

And for a Poisson data set (Eq. 5.13) this log likelihood becomes:

$$\ln(P) = \sum_{i=1}^{N} -\mu_i + y_i \ln \mu_i - \ln y_i! \tag{7.2}$$

Keep in mind that each increase (or decrease) of one in the log likelihood increases (or decreases) the data set probability by a factor of e.

Sample mean and variance

As a first application, the principal of maximum likelihood will be used to obtain the best estimate of μ_y from N samples: y_i, $i = 1...N$, all from the same Poisson parent distribution. For this case, all μ_i in Eq. 7.2 are the same ($\mu_i = \mu_y$) and the log likelihood function simplifies to

$$\ln(P) = -N\mu_y + \ln \mu_y \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \ln y_i! \tag{7.3}$$

The value of μ_y that maximizes $\ln(P)$ for any particular data set is not expected to be the true mean. It will be a estimate of that quantity and will be given a different name. The value of μ_y where $\partial \ln(P)/\partial\mu_y = 0$ will be designated $\bar{y}$ for reasons that will become clear shortly. This equation is written

$$0 = \left.\frac{\partial \ln(P)}{\partial \mu_y}\right|_{\mu_y=\bar{y}} \tag{7.4}$$

And its solution proceeds from Eq. 7.3 as follows:

$$\begin{aligned} 0 &= \left. -N + \frac{1}{\mu_y}\sum_{i=1}^{N} y_i \right|_{\mu_y=\bar{y}} \\ &= -N + \frac{1}{\bar{y}}\sum_{i=1}^{N} y_i \end{aligned} \tag{7.5}$$

Solving for $\bar{y}$ gives Eq. 2.7—the standard equation for the sample mean.

As a second application, best estimates of both μ_y and σ_y will be obtained from N samples—all from the same Gaussian parent. That is, all samples have $\Delta y_i = \Delta y$, $\mu_i = \mu_y$ and $\sigma_i = \sigma_y$ and the log likelihood function simplifies to

$$\ln(P) = -\frac{N}{2}\ln 2\pi + N \ln\left(\frac{\Delta y}{\sigma_y}\right) - \frac{1}{2\sigma_y^2}\sum_{i=1}^{N}(y_i-\mu_y)^2 \tag{7.6}$$

Again, $\bar{y}$ will be used to represent that value of μ_y where this log likelihood function maximizes giving

$$\begin{aligned} 0 &= \left.\frac{\partial \ln(P)}{\partial \mu_y}\right|_{\mu_y=\bar{y}} \\ &= \left. -\frac{1}{2\sigma_y^2}\sum_{i=1}^{N} 2(y_i-\mu_y)(-1)\right|_{\mu_y=\bar{y}} \\ &= \sum_{i=1}^{N}(y_i-\bar{y}) \\ &= -N\bar{y} + \sum_{i=1}^{N} y_i \end{aligned} \tag{7.7}$$

And, again, solving for $\bar{y}$ gives Eq. 2.7.

Thus, the sample mean of Eq. 2.7 has now been proven to be a best estimate of the distribution mean for variables governed by either a Poisson or a Gaussian distribution. The sample mean $\bar{y}$ has also been proven to be an unbiased estimate of μ_y. Thus, for these two cases, the best estimate is an unbiased estimate. Using the principle of maximum likelihood does not guarantee that a parameter obtained from it will be unbiased. Checking that estimates are unbiased is an important part of statistical analysis procedures.

Exercise 5 *The variance of $\bar{y}$ is most easily determined from Eq. 2.18—as the difference between the second moment and the square of the first—in this case: $\sigma_{\bar{y}}^2 = \langle \bar{y}^2 \rangle - \mu_y^2$. Evaluate the right side of this equation to show that*

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{N} \tag{7.8}$$

Hint 1: Re-express $\bar{y}^2$ using

$$\begin{aligned} \bar{y}^2 &= \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)\left(\frac{1}{N}\sum_{j=1}^{N} y_j\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \end{aligned} \tag{7.9}$$

before taking the expectation value. Note how each $\bar{y}$ must use its own private dummy index to clearly enumerate terms with $i = j$ and $i \neq j$ for use with Eq. 5.9.

Eq. 7.8 indicates that the standard deviation of the mean of N samples is $\sqrt{N}$ times smaller than the standard deviation of a single sample, i.e., the average of 100 samples is 10 times more precise an estimate of the true mean than is a single sample. This is an important fact, but it does not provide an estimate of σ_y from the sample set.

To get this estimate, first assume that μ_y is known.

Exercise 6 *Show that the best estimate of σ_y^2 is the sample variance s_y^2 as given by Eq. 2.20. Hint: Let s_y be that value of σ_y where the derivative of Eq. 7.6 with respect to σ_y is equal to zero.*

The fact that $\bar{y}$ satisfies the principle of maximum likelihood helps explain why the divisor N in Eq. 2.20 was changed to $N-1$ in Eq. 2.22. By choosing $\bar{y}$ such that it maximizes the log-likelihood of Eq. 7.6, the sum in Eq. 2.22 (using $\bar{y}$) is guaranteed to be a minimum and is thus always smaller than the sum in Eq. 2.20 (using μ_y). As you will prove in the next exercise, changing N to $N-1$ corrects for the average reduction so that the s_y^2 of Eq. 2.22 is still unbiased.

Exercise 7 *Show that Eq. 2.22 also satisfies Eq. 2.21. Hint 1: Explain why each of the N terms in Eq. 2.22 has the same expectation value and use this fact to get rid of the sum over i—replacing it with a factor of N times the expectation value of one term (say $i = 1$). Hint 2: Expand $(y_1 - \bar{y})^2$ before taking the expectation value term by term. Then use Eqs. 2.7 and 5.9 and/or results from Exercise 5 as needed for the individual terms.*

Using the best estimate s_y in place of σ_y in Eq. 7.8 then gives the *sample standard deviation of the mean*

$$s_{\bar{y}} = \frac{s_y}{\sqrt{N}} \tag{7.10}$$

which is a best estimate of $\sigma_{\bar{y}}$.

Chapter 8

Regression Analysis

Regression analysis refers to techniques involving data sets with one or more dependent variables measured as a function of one or more independent variables with the goal to compare that data with a theoretical model and extract model parameters and their uncertainties.

Here, the terms dependent and independent variable refer to experimental relationships unrelated to statistical dependence or independence. For example, the dependent variables are often statistically independent from one another. The naming largely arises because the dependent variables can be considered a function of (depend on) the values of the independent variables.

The dependent variables will be denoted y_i, $i = 1...N$ and each y_i will be modeled as an independent sample from either a Gaussian or Poisson probability distribution. For each y_i, the theoretical model predicts the value of the distribution mean μ_i via a fitting function F_i that depends on (1) the values of the independent variables associated with that point and (2) a set of M fitting parameters α_k, $k = 1...M$

$$\mu_i = F_i(\alpha_1, \alpha_2, ..., \alpha_M) \tag{8.1}$$

where the subscript i in $F_i(\{\alpha\})$ denotes the independent variables. The lack of any explicit independent variables in Eq. 8.1 is intentional. In a regression analysis they merely serve to distinguish the point by point dependences of the predicted μ_i on the fitting parameters.

Independent variables can be random variables, but accounting for their random error can be difficult and will only be addressed after an initial treatment assuming they are known exactly.

All input random variables in a regression analysis are typically assumed to be statistically independent. If correlations exist, this assumption would have to be abandoned and covariances between variables would have to be taken into account. The required modifications depend on the details of the correlations, but when the only correlations are between dependent variables, they are relatively simple and are addressed later in this chapter. Until then, all input variables will be considered statistically independent.

Adjustable parameters in a regression analysis are used only in predicting the set of μ_i appearing in the joint probability distribution for the entire data set. These parameters will be chosen to maximize the data set probability by maximizing the log likelihood functions of either Eq. 7.1 or Eq. 7.2 depending, respectively, on whether the y_i follow Gaussian or Poisson distributions.

For Gaussian-distributed y_i, only the last term in Eq. 7.1, the one with a sum of squared deviations, has any dependence on the μ_i and thus the best-fit solution must only maximize this quantity. As this last term has a negative prefactor, it is traditional to consider the maximum likelihood process as a minimization of the following *sum of squares.*

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \tag{8.2}$$

That is why regression analysis is often referred to as a *least-squares* procedure despite its correct origin as a maximum likelihood method.

The χ^2 given by Eq. 8.2 has a form very similar to the chi-square random variable used for determining "goodness of fit"—a topic for the next chapter. At this point, χ^2 should only be considered as providing a quantity to be made as small as possible in order to make the probability of the data set as big as possible. Its importance here is in how it depends on the data and the true means.

For Poisson-distributed y_i, only the first two terms in the log likelihood function of Eq. 7.2 have any dependence on the μ_i. Consequently, the best-fit solution must only maximize this part of the log likelihood function. As we will see shortly, regression formulas for Poisson-distributed data are quite similar to those for Gaussian- distributed data. In anticipation of this similarity, we define a chi-square-like variable for fits to Poisson-distributed data that is, as with the χ^2 for Gaussian-distributed data, -2 times the

μ_i-dependent terms in the log likelihood function.

$$\chi^2_P = 2\sum_{i=1}^N \mu_i - y_i \ln \mu_i \tag{8.3}$$

Maximizing the data set probability will then require minimizing this quantity.

Notice that both χ^2 and χ^2_P are unitless quantities and must decrease by two to increase the log likelihood function by one and increase the data set probability by a factor of e.

According to the principle of maximum likelihood, the best-fit parameters, which will be denoted a_k, when used in place of the true α_k must maximize the probability, and thus minimize the chi-square, for that particular data set. Consequently, the a_k are associated with one particular data set and are random variables.

These ideas can be summarized in $N + 1 + M$ equations. First, when substituted for the α_k in the fitting functions, the a_k are said to give the *best-fit* y-values, which will be denoted $y_i^{\rm fit}$, $i = 1...N$. These N equations are

$$y_i^{\rm fit} = F_i(\{a\}) \tag{8.4}$$

where $\{a\}$ is the set of a_k, $k = 1...M$.

Second, using the $y_i^{\rm fit}$ in place of the μ_i in Eq. 8.2 for Gaussian variables, or in Eq. 8.3 for Poisson variables, gives the minimized chi-square. This single equation is either

$$\chi^2 \quad = \quad \sum_{i=1}^N \frac{\left(y_i - y_i^{\rm fit}\right)^2}{\sigma_i^2} \tag{8.5}$$

for Gaussian y_i, or

$$\chi^2_P \quad = \quad 2\sum_{i=1}^N y_i^{\rm fit} - y_i \ln y_i^{\rm fit} \tag{8.6}$$

for Poisson y_i.

Third, the condition that the χ^2 or χ^2_P be a minimum with respect to all best-fit parameters is that its partial derivatives with respect to each a_k must be zero. Keeping in mind that both χ^2 and χ^2_P depend on the a_k only via the dependence of $y_i^{\rm fit}$ on these quantities (Eq. 8.4), chain rule differentiation gives for Gaussian-distributed variables:

$$\begin{aligned}
0 &= \frac{\partial\chi^2}{\partial a_k} \\
&= \frac{\partial}{\partial a_k}\sum_{i=1}^{N}\frac{(y_i-y_i^{\rm fit})^2}{\sigma_i^2} \\
&= \sum_{i=1}^{N}\frac{2(y_i-y_i^{\rm fit})(-1)}{\sigma_i^2}\frac{\partial y_i^{\rm fit}}{\partial a_k} \\
&= \sum_{i=1}^{N}\frac{(y_i-y_i^{\rm fit})}{\sigma_i^2}\frac{\partial y_i^{\rm fit}}{\partial a_k} \qquad (8.7)
\end{aligned}$$

and for Poisson-distributed variables it gives

$$\begin{aligned}
0 &= \frac{\partial\chi_P^2}{\partial a_k} \\
&= \frac{\partial}{\partial a_k}\sum_{i=1}^{N} y_i^{\rm fit} - y_i\ln y_i^{\rm fit} \\
&= \sum_{i=1}^{N}\left(1-\frac{y_i}{y_i^{\rm fit}}\right)\frac{\partial y_i^{\rm fit}}{\partial a_k} \\
&= \sum_{i=1}^{N}\frac{(y_i^{\rm fit}-y_i)}{y_i^{\rm fit}}\frac{\partial y_i^{\rm fit}}{\partial a_k} \\
&= \sum_{i=1}^{N}\frac{(y_i-y_i^{\rm fit})}{y_i^{\rm fit}}\frac{\partial y_i^{\rm fit}}{\partial a_k} \qquad (8.8)
\end{aligned}$$

Equation 8.7 or 8.8 should be considered a set of M equations, one for each k, which can then be solved for the M unknown a_k.

Note how this development has generated another example of counting statistics for Poisson-distributed variables. Equation 8.8 can be considered a special case of Eq. 8.7—one with with $\sigma_i^2 = y_i^{\rm fit}$. However, using $\sigma_i^2 = y_i^{\rm fit}$ has a caveat and it presents a minor complication. The caveat is that solutions will then be minimizing the χ_P^2 of Eq. 8.6 not the χ^2 of Eq. 8.5. The complication arises because the σ_i are inputs for finding the best-fit parameters, and because the $y_i^{\rm fit}$ are not known until the best-fit parameters are known, an iterative approach will be needed. One might start with a guess for the fitting parameters to get an initial set of $y_i^{\rm fit}$, which would then be used for the σ_i^2. The results from this first fit can then be used to get a better set of fitting parameters, a more accurate set of $y_i^{\rm fit}$, and thus more accurate values of σ_i^2 to use in a second iteration. Additional iterations can be performed until the a_k and the $y_i^{\rm fit}$ converge.

Thus, with the understanding that $\sigma_i^2 = y_i^{\rm fit}$ for Poisson-distributed variables, Eqs. 8.7 and 8.8 can both be represented

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \frac{\partial y_i^{\rm fit}}{\partial a_k} = \sum_{i=1}^{N} \frac{y_i^{\rm fit}}{\sigma_i^2} \frac{\partial y_i^{\rm fit}}{\partial a_k} \tag{8.9}$$

Don't forget this equation should be considered a set of M simultaneous equations, one for each a_k.

Linear algebra is typically used to solve Eq. 8.9. With quantities represented by vectors or matrices, the resulting expressions simplify. Sets of equations become vector equations and summation symbols disappear—replaced by summations implied by standard vector/matrix multiplication rules. All regression formulas using linear algebra techniques are derived in the *Regression Analysis Addendum* and demonstrated for a quadratic fit using array formulas in Excel in *Linear Regression.xls* available on the lab web site.

The parameters α_k and their best estimates a_k are represented by column vectors of length M.

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix} \tag{8.10}$$

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{pmatrix} \tag{8.11}$$

The data y_i and the best fit $y_i^{\rm fit}$ are represented by column vectors of length N.

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{8.12}$$

and

$$\boldsymbol{y}^{\rm fit} = \begin{pmatrix} y_1^{\rm fit} \\ y_2^{\rm fit} \\ \vdots \\ y_N^{\rm fit} \end{pmatrix} \tag{8.13}$$

The standard deviations of the y_i are represented by their $N \times N$ diagonal covariance matrix

$$[\sigma_y] = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots \\ 0 & \sigma_2^2 & 0 & \cdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix} \tag{8.14}$$

or by its inverse, called the *weighting matrix*

$$\left[\sigma_y^{-1}\right] = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \cdots \\ 0 & 1/\sigma_2^2 & 0 & \cdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_N^2 \end{pmatrix} \tag{8.15}$$

Finally, defining the $N \times M$ Jacobian matrix with elements

$$[J]_{ik} = \frac{\partial y_i^{\rm fit}}{\partial a_k} \tag{8.16}$$

the M equations of Eq. 8.9 become the M components of the following vector equation

$$\left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y} = \left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y}^{\rm fit} \tag{8.17}$$

where $\left[J^T\right]$ is the transpose of $[J]$, i.e., is an $M \times N$ matrix with the rows and columns of $[J]$ interchanged. The elements of the transpose satisfy

$$\left[J^T\right]_{ki} = [J]_{ik} \tag{8.18}$$

It is recommended that the reader check Eq. 8.17 and verify that it is, indeed, a vector equation having M elements with the kth element reproducing Eq. 8.9 including the proper summation over the i index.

Linear Regression

Linear regression is used when the fitting function is linear in the fitting parameters. For a fitting function with a single independent variable x_i, each $y_i^{\rm fit}$ would be of the form

$$\begin{aligned} y_i^{\rm fit} &= a_1 f_1(x_i) + a_2 f_2(x_i) + ... + a_M f_M(x_i) \\ &= \sum_{k=1}^{M} a_k f_k(x_i) \end{aligned} \tag{8.19}$$

where the $f_k(x)$ are linearly independent *basis functions* of x with no unknown parameters. For example, a data set for the position y_i vs. time t_i of a cart rolling on an inclined track might be checked against a predicted quadratic based on motion at constant acceleration:

$$y_i^{\rm fit} = a_1 + a_2 t_i + a_3 t_i^2 \tag{8.20}$$

This is linear in the three parameters with basis functions: $f_1(t_i) = 1$, $f_2(t_i) = t_i$, and $f_3(t_i) = t_i^2$.

Equation 8.16 for the Jacobian is then the $N \times M$ matrix given by

$$[J] = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_M(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_M(x_N) \end{pmatrix} \tag{8.21}$$

The set of N equations for the $y_i^{\rm fit}$ (Eq. 8.19) can then be expressed by the single vector equation (with N elements)

$$\boldsymbol{y}^{\rm fit} = [J]\,\boldsymbol{a} \tag{8.22}$$

and Eq. 8.17 becomes

$$\left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y} = \left[J^T\right]\left[\sigma_y^{-1}\right][J]\,\boldsymbol{a} \tag{8.23}$$

The combination $\left[J^T\right]\left[\sigma_y^{-1}\right][J]$ is an $M \times M$ matrix that will become prominent in the discussion of the parameter covariance matrix. It will be given its own symbol $[X]$

$$[X] = \left[J^T\right]\left[\sigma_y^{-1}\right][J] \tag{8.24}$$

so that Eq. 8.23 becomes

$$\left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y}=[X]\boldsymbol{a} \tag{8.25}$$

$[X]$ is called the *curvature matrix* because it determines the parabolic shape of chi-square values as a function of the M parameter values. This issue is discussed in the section on nonlinear regression.

Equation 8.25 is solved for the best-fit parameter vector $\boldsymbol{a}$ by determining the inverse $[X^{-1}]$ of the curvature matrix and multiplying by it on the left of both sides of Eq. 8.25

$$\boldsymbol{a}=\left[X^{-1}\right]\left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y} \tag{8.26}$$

or

$$\boldsymbol{a}=\left[J^{\dagger}\right]\boldsymbol{y} \tag{8.27}$$

where

$$\left[J^{\dagger}\right]=\left[X^{-1}\right]\left[J^T\right]\left[\sigma_y^{-1}\right] \tag{8.28}$$

$\left[J^{\dagger}\right]$ is an $M\times N$ matrix called the (weighted) *Moore-Penrose* pseudo-inverse of $[J]$.

The a_k are random variables. If new sets of y_i were taken repeatedly, the a_k calculated for each set would vary. If the y_i are Gaussian distributed, or even if they are not and the data set N is large enough, the a_k will follow Gaussian or near-Gaussian distributions and their variations with each new data set will be correlated. What can be expected for the means, variances and covariances for the a_k?

The a_k will be unbiased. That is, the distribution for a_k will have a mean of α_k. The vector equation

$$\langle\boldsymbol{a}\rangle=\boldsymbol{\alpha} \tag{8.29}$$

is provably true. The proof can be found in the *Regression Analysis Addendum.* This addendum also shows that the parameter covariance matrix is given by

$$[\sigma_a]=\left[X^{-1}\right] \tag{8.30}$$

In other words, each diagonal element of $[X^{-1}]$ gives the variance and thus its square root gives the standard deviation for the corresponding parameter. Furthermore, fitting parameters are usually correlated and the off-diagonal elements give their covariances. The *Slope-Intercept Correlation* Excel workbook shows an example of predicted and sample values for the variances and covariance of the slope and intercept for simulated data sets following a straight-line relationship.

It is important to point out that the parameter covariance matrix is not a random variable. It does not depend on the y_i, which are the only random variables in the analysis. To the extent that the σ_i and the Jacobian $[J]$ can be predetermined, Eqs. 8.30 and 8.24 show that the parameter covariance matrix can be predetermined as well.

Weighted mean

The simplest linear regression problem is a fit to a constant. It arises when averaging sample values of the same physical quantity obtained from a variety of sources so that each sample value has a different uncertainty. A straight mean (Eq. 2.7) would be correct if all the uncertainties were the same. In the more general case, however, precise sample values with small uncertainties must be weighted more heavily than less precise values with larger uncertainties. The correct result will be a *weighted mean* that properly takes into account the uncertainty of each value.

The data set consists of y_i, $i = 1...N$ with varying standard deviations σ_i. The y_i are assumed independent so the $N \times N$ covariance and weighting matrix are as given in Eqs. 8.14 and 8.15.

The prediction is that the mean of the distribution associated with each y_i is the same, $\mu_i = \mu_y$, where μ_y is the (unknown) true value of the physical quantity. The best estimate of that quantity, let's call it m_y, is sought. This is a one-parameter, linear fit to a constant. In the notation of the linear regression formulas, $M = 1$, $a_1 = m_y$ and $f_1 = 1$ for all i.

The parameter vector $\boldsymbol{a}$ is of single-element form.

$$\boldsymbol{a} = (m_y) \tag{8.31}$$

and the Jacobian matrix is an $N \times 1$ matrix given by

$$[J] = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{8.32}$$

The curvature matrix $[X] = \left[J^T\right]\left[\sigma_y^{-1}\right][J]$ is then easily shown to be the

1×1 matrix given by

$$[X] = \left[\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right] \tag{8.33}$$

and the covariance matrix for $\boldsymbol{a}$—the inverse of $[X]$—is simply the 1×1 matrix given by

$$[X^{-1}] = \left[\left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1} \right] \tag{8.34}$$

Using $\boldsymbol{a} = [X^{-1}] \left[J^T\right] \left[\sigma_y^{-1}\right] \boldsymbol{y}$ to solve for the best-fit $\boldsymbol{a}$ gives, after a bit of simple vector-matrix manipulations, its only element

$$m_y = \frac{\displaystyle\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}}{\displaystyle\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \tag{8.35}$$

Eq. 8.35 is called a weighted average:

$$m_y = \frac{w_1 y_1 + w_2 y_2 + ... + w_N y_N}{w_1 + w_2 + ... + w_N} \tag{8.36}$$

where each weight is the inverse of the variance:

$$w_i = \frac{1}{\sigma_i^2} \tag{8.37}$$

Note that larger standard deviations indicate less precisely known measurements and, appropriately, smaller weights in the average.

The weighting effect of data point uncertainties is most obvious for a fit to a constant, but persists in all regression problems. The larger the σ_i, the smaller the weighting of that point in its effect on the fitting parameters.

The diagonal (and only) element of the covariance matrix $[X^{-1}]$ is the variance of the only element m_y of the parameter vector $\boldsymbol{a}$. That is,

$$\sigma_{m_y}^2 = \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \right)^{-1} \tag{8.38}$$

which is often expressed in a form somewhat easier to remember

$$\frac{1}{\sigma_{m_y}^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{8.39}$$

Effectively, Eqs. 8.35 and 8.39 are a prescription for turning a group of independent samples (with known standard deviations) into a single sample m_y with a reduced standard deviation σ_{m_y}.

Equally-weighted linear regression

Occasionally, all y_i are obtained using the same technique and have the same uncertainty. Or, the uncertainties are not known very well and assuming they are equal is an appropriate starting point. This case is fairly common and will be the second linear regression example to consider. For a data set where the standard deviations are the same for all y_i, $\sigma_i = \sigma_y$, the regression equations and results are then called *equally-weighted.*

The equations simplify because the covariance and weighting matrix are proportional to the identity matrix. $[\sigma_y] = \sigma_y^2[I]$ and $\left[\sigma_y^{-1}\right] = (1/\sigma_y^2)[I]$. With this substitution, Eq. 8.24 becomes

$$[X] = \frac{1}{\sigma_y^2}[X_u] \tag{8.40}$$

where now $[X_u]$ is the $[X]$ matrix without the weighting matrix

$$[X_u] = \left[J^T\right][J] \tag{8.41}$$

and is thus independent of σ_y. The inverse of Eq. 8.40 is then $[X^{-1}] = \sigma_y^2[X_u^{-1}]$ where $[X_u^{-1}]$, the inverse of $[X_u]$, is also independent of σ_y.

By Eq. 8.30, $[X^{-1}]$ is the parameter covariance matrix. That is,

$$[\sigma_a] \quad = \quad \sigma_y^2[X_u^{-1}] \tag{8.42}$$

demonstrating that every element of the parameter covariance matrix is proportional to σ_y^2. Equation 8.42 thus further implies that the standard deviation of every parameter (square root of the corresponding diagonal element) is proportional to the standard deviation (σ_y) of the y-values for that data set.

Equation 8.26 for the parameter values becomes

$$\boldsymbol{a} = \left[X_u^{-1}\right]\left[J^T\right]\boldsymbol{y} \tag{8.43}$$

showing that the σ_y^2 has canceled and thus that the parameter values themselves do not depend on its value. Moreover, the chi-square of Eq. 8.5 becomes

$$\chi^2 = \frac{1}{\sigma_y^2}\sum_{i=1}^{N}\left(y_i - y_i^{\text{fit}}\right)^2 \tag{8.44}$$

The σ_y factors from the sum. The best-fit parameters minimize the χ^2 and for this case minimize the same sum of squared deviations for any σ_y.

Nonlinear Regression

Linear regression techniques can only be used when the Jacobian is independent of the entire set of fitting parameters. Fitting functions that are nonlinear in the fitting parameters do not satisfy this requirement. For example, consider a short-lived radioactive sample positioned in front of a Geiger-Müller tube. The counts y_i are measured over a ten-second interval as a function of the time t_i since starting the experiment. A nonlinear model predicting exponential decay would be represented

$$y_i^{\text{fit}} = a_1 e^{-t_i/a_2} + a_3 \tag{8.45}$$

where a_1 is proportional to the initial sample activity, a_2 is the mean lifetime, and a_3 is a constant arising from natural background radiation.

Because the Jacobian depends on the parameters, solving for the best-fit parameters of a nonlinear function will require techniques that differ from those of linear regression while retaining significant similarities.

Unlike linear regression, which can find the best-fit parameters from a single evaluation of the appropriate formulas, a nonlinear regression program must find the solution iteratively. The user must provide initial guesses for the fitting parameters that will be used as a starting point. From these initial guesses, the program must test other nearby parameter sets—evaluating the chi-square value each set produces. Each time the program finds that the chi-square has decreased, it uses those improved parameter values as a new starting point as it tries to find its way to the best-fit parameters which produce the minimum chi-square and maximize the data set probability.

The following discussion describes nonlinear regression for Gaussian-distributed y_i. The two minor adjustments necessary for treating Poisson-distributed data will not always be mentioned but are always the same: (1) Use the most current $y_i^{\rm fit}$ for σ_i^2 in the input covariance or weighting matrix, and (2) Use the χ_P^2 of Eq. 8.6 for the chi-square when checking for an improved fit.

The foundation for nonlinear regression is the *Gauss-Newton* algorithm. It is also the quickest way to find the best-fit parameters when the starting parameters are already sufficiently close to the best-fit parameters. Its big drawback is that it tends to fail if the starting point is not close enough. The *gradient-search* algorithm is better at improving the fit parameters when the starting parameters are farther from the best-fit parameters. Its big drawback is that it tends to take a long time to find the best fit. The *Levenberg-Marquardt* algorithm elegantly addresses the shortcomings of both algorithms. In effect, it uses the gradient-search algorithm when the starting points are far off and switches to the Gauss-Newton algorithm as the starting points get closer to the best fit.

All algorithms start by evaluating the Jacobian (Eq. 8.16) at the starting parameter values—either by numerical differentiation or from formulas supplied by the user.

If the starting point is near the best-fit values, the chi-square will be near its true minimum and will have a quadratic dependence on the M fitting parameters—it will be an M-dimensional parabola. The Gauss-Newton algorithm uses the Jacobian at the present starting point to determine this M-dimensional parabola and having done so can jump directly to the minimum. For linear fitting functions, the parabolic shape is guaranteed to be accurate—even when the starting parameters are far from the minimum. In effect, this is why linear regression formulas can find the best fit parameters in a single step. For nonlinear fitting functions, if the starting point is close enough to the best fit parameters, the parabolic shape is again guaranteed and the Gauss-Newton algorithm would jump directly to the correct best-fit parameters in one try.

However, if the starting point is too far from the true minimum, the local parabolic shape may not predict where the true χ^2 minimum will be. When the Gauss-Newton algorithm jumps to the predicted best-fit parameters, it may find the χ^2 has decreased, but not to its minimum. In this case, it can simply start another iteration from there. All too often though, it finds that the χ^2 has actually increased and it would then have no recourse to improve the fit.

The gradient search algorithm always works and is the method of choice from starting points where the Gauss-Newton algorithm fails. The gradient search ignores the curvature of the chi-square parabola and uses the Jacobian only to determine the gradient of chi-square at the starting point. It then moves the parameters from the starting point (takes a step) in a direction opposite the gradient—along the direction of steepest descent for the chi-square. Because it ignores the curvature, however, this algorithm can't be sure how big a step to take. So it takes a step and if the chi-square does not decrease at the new point, it goes back to the starting point, decreases the step size by some factor, and tries again. Sooner or later the step size will be small enough that the chi-square will decrease—leading to a new improved starting point from which to continue the search.

The problem with the gradient algorithm is that near the chi-square minimum, the gradient is not steep at all, the steps become very small and the algorithm proceeds only slowly toward the best-fit solution.

The elegance of the Levenberg-Marquardt algorithm is in how it monitors the χ^2 at each new trial point and uses a single scalar parameter to smoothly adjust the step size and switch between the two algorithms.

The Gauss-Newton algorithm

The Gauss-Newton algorithm is presented in some detail because it illustrates the similarities between the nonlinear regression formulas and their linear regression counterparts, and it gives the parameter covariance matrix.

The treatment will require distinguishing between the best fit parameters a_k, $k = 1...M$ and another set—nearby, but otherwise arbitrary. These nearby fitting parameters will be labeled $a_k^{\rm ini}$, $k = 1...M$, for which the fitting function gives a set of fitted y-values which will be labeled $y_i^{\rm ini}$, $i = 1...N$

$$y_i^{\rm ini} = F_i(\{a^{\rm ini}\}) \tag{8.46}$$

This nearby or almost-there solution must be close enough to the best fit that for every point i in the data set, a first-order Taylor series expansion about $y_i^{\rm ini}$ must accurately reproduce the best-fit $y_i^{\rm fit}$.

$$y_i^{\rm fit} = y_i^{\rm ini} + \sum_{k=1}^{M} \frac{\partial y_i^{\rm ini}}{\partial a_k^{\rm ini}}(a_k - a_k^{\rm ini}) \tag{8.47}$$

Where this expansion is accurate, the χ^2 surface is parabolic.

Differentiating this Taylor expansion gives the elements of the Jacobian, Eq. 8.16, $[J]_{ik} = \partial y_i^{\rm fit}/\partial a_k$ as

$$[J]_{ik} = \frac{\partial y_i^{\rm ini}}{\partial a_k^{\rm ini}} \tag{8.48}$$

For any reasonable fitting function, the Taylor expansion is guaranteed accurate for values of $a_k^{\rm ini}$ that are sufficiently close to the a_k. If the Taylor expansion remains valid for a wider range of $a_k^{\rm ini}$, so too will the Gauss-Newton algorithm find the best-fit a_k from those more distant starting $a_k^{\rm ini}$. Moreover, the Gauss-Newton treatment also provides the parameter covariance matrix based on this expansion. Consequently, the range of validity for the Taylor expansion will have important implications for parameter uncertainties.

To see how the Taylor series expansion will lead to linear-regression-like formulas, several new, modified quantities are defined. First, a modified input Δy_i and a modified best-fit $\Delta y_i^{\rm fit}$ are defined as relative to the almost-there fit values $y_i^{\rm ini}$.

$$\Delta y_i = y_i - y_i^{\rm ini} \tag{8.49}$$

$$\Delta y_i^{\rm fit} = y_i^{\rm fit} - y_i^{\rm ini} \tag{8.50}$$

In vector notation these last two equations (for all i) can be written $\Delta\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{y}^{\rm ini}$ and $\Delta\boldsymbol{y}^{\rm fit} = \boldsymbol{y}^{\rm fit} - \boldsymbol{y}^{\rm ini}$.

Subtracting $\left[J^T\right]\left[\sigma_y^{-1}\right]\boldsymbol{y}^{\rm ini}$ from both sides of the defining equation for the maximum likelihood solution, namely, Eq. 8.17, then gives

$$\left[J^T\right]\left[\sigma_y^{-1}\right]\Delta\boldsymbol{y} = \left[J^T\right]\left[\sigma_y^{-1}\right]\Delta\boldsymbol{y}^{\rm fit} \tag{8.51}$$

Next, the modified best-fit parameters are defined as the difference between the actual best-fit parameters and the almost-there parameters.

$$\Delta a_k = a_k - a_k^{\rm ini} \tag{8.52}$$

With these definitions, Eq. 8.47 can be written

$$\Delta y_i^{\rm fit} = \sum_{k=1}^{M} [J]_{ik}\,\Delta a_k \tag{8.53}$$

or in vector notation

$$\Delta\boldsymbol{y}^{\rm fit} = [J]\,\Delta\boldsymbol{a} \tag{8.54}$$

which is now linear in the modified fit parameters.

Using Eq. 8.54 in Eq. 8.51 gives the linear regression-like result

$$\left[J^T\right]\left[\sigma_y^{-1}\right]\Delta\boldsymbol{y} = \left[J^T\right]\left[\sigma_y^{-1}\right][J]\,\Delta\boldsymbol{a} \tag{8.55}$$

This equation is now in the linear regression form analogous to Eq. 8.23 and the solution for the best fit Δa_k is analogous to Eq. 8.27.

$$\Delta\boldsymbol{a} = \left[J^\dagger\right]\Delta\boldsymbol{y} \tag{8.56}$$

where $\left[J^\dagger\right] = \left[X^{-1}\right][J]\left[\sigma_y^{-1}\right]$ (Eq. 8.28) with $[X] = \left[J^T\right]\left[\sigma_y^{-1}\right][J]$ (Eq. 8.24), but with the elements of $[J]$ now given by Eq. 8.48, i.e., the Jacobian derivatives are evaluated using the starting parameters.

Note that there are no changes to the weighting matrix $\left[\sigma_y^{-1}\right]$. For example, for a fit to y-values governed by a Poisson distribution, $\sigma_i^2 = y_i^{\rm fit}$ still applies.

After Eq. 8.56 is applied to determine the best-fit $\Delta\boldsymbol{a}$, Eq. 8.52 must then be applied to each element to find the best-fit a_k

$$a_k \;=\; a_k^{\rm ini} + \Delta a_k \tag{8.57}$$

The underlying linear regression formulation implies $\left[X^{-1}\right]$ is the covariance matrix for the Δa_k. Because of the constant offset relation between Δa_k and a_k expressed by Eq. 8.57, $\left[X^{-1}\right]$ would then also be the covariance matrix for the parameters a_k themselves.

The final values of a_k from this analysis should then be used as a new initial point $a_k^{\rm ini}$ for another iteration of the process. The Jacobian, and perhaps the weighting matrix as well, should be reevaluated there and the Gauss-Newton algorithm reiterated. Iterations should be continued until no significant changes in the a_k result, i.e., until Eq. 8.56 (with the Jacobian evaluated at the best fit) gives $\Delta\boldsymbol{a} = 0$.

Determining when to stop the iteration process is not a simple matter. There is often a termination parameter giving the smallest change or fractional change in χ^2 from one iteration to the next for which another iteration will be performed. There will also be a maximum number of iterations to ensure the program stops eventually—even if the χ^2 termination condition is never met. Always make sure a nonlinear regression program has not encountered a problem and has found the true χ^2 minimum.

The Gauss-Newton, gradient search, and Levenberg-Marquardt algorithms are demonstrated for (simulated) exponential decay data in the two Excel spreadsheets *Nonlinear Regression* and *Nonlinear Regression Poisson* for a Gaussian- and Poisson-distributed variable, respectively. Note, in particular, how increasing or deceasing the Levenberg-Marquardt parameter (λ in the spreadsheet) changes the weighting of the diagonal element of the (modified) curvature matrix, thereby changing from a near-gradient-search algorithm with small step size when $\lambda \gg 1$ to a near-Gauss-Newton algorithm when $\lambda \ll 1$.

The parameter covariance matrix $[\sigma_a]$ should always be obtained using an unmodified curvature matrix ($\lambda = 0$). It should also be evaluated with the Jacobian $[J]$ and the input covariance matrix $[\sigma_y]$ evaluated at the best-fit.

The $\Delta\chi^2 = 1$ rule

The use of Excel's Solver program to perform nonlinear regression is described in Chapter 10. Solver finds the best-fit solution by minimizing the χ^2 (or χ^2_P for Poisson-distributed y_i) but it does not provide the fitting parameter uncertainties. Nonetheless, the parameter covariance matrix $[\sigma_a]$ can be determined with the Solver, but it takes a few extra steps to do so. The principle needed is called the $\Delta\chi^2 = 1$ rule and is described next. As an added benefit, the rule can also provide a check on the range of validity of the first order Taylor expansion.

To use the $\Delta\chi^2 = 1$ rule, the χ^2 of Eq. 8.5 should be used even if the χ^2_P of Eq. 8.6 is minimized to find the best fit. The χ^2 must be evaluated at the best fit and then again using fitting parameters that are offset from their best-fit values. This χ^2 is a minimum when evaluated with the best-fit parameters and increases when evaluated with any other parameter set. Where the Taylor expansion (Eq. 8.47) is valid, the increase, $\Delta\chi^2$, will be a generalized quadratic function of the deviations of the parameters from their best-fit values. This M-dimensional parabola can be expressed

$$\Delta\chi^2 = \Delta\boldsymbol{a}^T \left[\sigma_a^{-1}\right] \Delta\boldsymbol{a} \tag{8.58}$$

where $\Delta\boldsymbol{a} = \boldsymbol{a}' - \boldsymbol{a}$ gives the deviations from the best-fit values $\boldsymbol{a}$ and $[\sigma_a^{-1}] = [X]$ is the inverse of the parameter covariance matrix.

The $\Delta\chi^2 = 1$ rule turns Eq. 8.58 inside out—finding and the parameter uncertainties from variations in χ^2 as follows.

> If a fitting parameter is offset from its best-fit value by its standard deviation e.g., from a_k to $a_k + \sigma_k$ and then fixed there while

> all other fitting parameters are readjusted to minimize the χ^2, the new χ^2 will be one higher than its best-fit minimum.

Where the Taylor expansion is valid, so is Eq. 8.58 and so is the following equation—a more general form of the $\Delta\chi^2 = 1$ rule showing explicitly the expected quadratic dependence of $\Delta\chi^2$ on the change in a_k.

$$\sigma_k^2 = \frac{(a'_k - a_k)^2}{\Delta\chi^2} \tag{8.59}$$

Here, a'_k is the changed value of the parameter whose best-fit value is a_k and $\Delta\chi^2$ is the increase in χ^2 after re-fitting the other parameters for a minimum χ^2 at the offset value a'_k.

The covariances between a_k and the other parameters can also be determined by keeping track of the parameter changes after the re-fit. If some other parameter with a best-fit value of a_m goes to a'_m after the re-fit, the covariance between a_k and a_m, including its sign, is given by

$$\sigma_{km} = \frac{(a'_k - a_k)(a'_m - a_m)}{\Delta\chi^2} \tag{8.60}$$

To check the range of validity of the Taylor expansion, one can check that the variances and covariances obtained with Eq. 8.59 and 8.60 give the same result for any a'_k in the range $a_k \pm 3\sigma_k$. The check should be performed for each parameter individually—varying a'_k by small amounts both above and below the best-fit value and sized so that $\Delta\chi^2$ values are around 0.1 and around 10. This tests that the quadratic scaling behavior is valid for variations from about $a'_k - a_k \approx \pm\sigma_k/3$ ($\Delta\chi^2 \approx 1/9$) to $a'_k - a_k \approx \pm 3\sigma_k$ ($\Delta\chi^2 \approx 9$). If σ_k is roughly constant for all four cases (say within 10%), one can be reasonably assured that confidence intervals $a_k \pm z\sigma_k$ should have close to Gaussian probabilities up to $z = 3$.

Equation 8.58 is valid—even for Poisson-distributed variables—as long as the σ_i^2 in the χ^2 denominator are held fixed. Thus, for Poisson-distributed y_i, it is the change in the χ^2 (not χ^2_P) that should be used in the $\Delta\chi^2 = 1$ rule with the χ^2 denominator be fixed at $\sigma_i^2 = y_i^{\rm fit}$. It should not be allowed to vary during the re-fit.

The guidelines of the previous paragraph are important in theory. In practice, allowing the σ_i^2 in the χ^2 denominator to vary during the re-fit is unlikely to significantly affect the resulting calculations of the parameter variances or covariances. Indeed, the variations in χ^2_P are normally similar enough to the variations in χ^2, that a $\Delta\chi^2_P = 1$ rule, with χ^2_P replacing the normal χ^2, will generally give good results.

Uncertainties in independent variables

Up to now, only the dependent variable had uncertainty; only the y_i were random variables. What can be done when there are uncertainties in the independent variable; when the x_i are also random variables? There is no rigorous treatment for the general case. However, if the x_i are statistically independent and have uncertainties that are small enough, a simple modification to the data point weightings can accurately model this complexity.

Only a single independent variable x will be considered here, i.e., where

$$y_i^{\rm fit} = F(\{a\}; x_i) \tag{8.61}$$

but the extension to additional independent variables should be obvious. Letting σ_{x_i} represent the standard deviation of x_i and letting μ_{x_i} represent its mean, $F(\{a\}; x_i)$ must be nearly linear throughout the range $\mu_{x_i} \pm 3\sigma_{x_i}$. That is, each $y_i^{\rm fit}$ should be well described by a first order Taylor expansion

$$y_i^{\rm fit} = F(\{a\}; \mu_{x_i}) + \frac{\partial F(\{a\}; x_i)}{\partial x_i}(x_i - \mu_{x_i}) \tag{8.62}$$

for any x_i in this range.

Under these conditions, propagation of error implies that random variations in x_i with a standard deviation of σ_{x_i} would cause random variations in $y_i^{\rm fit}$ with a standard deviation

$$\sigma_{y_i^{\rm fit}} = \sigma_{x_i}\frac{\partial F(\{a\}; x_i)}{\partial x_i} \tag{8.63}$$

If x_i is statistically independent from y_i, the variations in $y_i^{\rm fit}$ will be uncorrelated with the variations in y_i and propagation of errors implies that the quantity $y_i - y_i^{\rm fit}$ will have variations with a variance given by

$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{\partial F(\{a\}; x_i)}{\partial x_i}\right)^2 \sigma_{x_i}^2 \tag{8.64}$$

where now σ_{y_i} is the standard deviation associated with y_i.

To account for uncertainty in an independent variable, simply replace the σ_i^2 appearing in the regression formulas with the modified values of Eq. 8.64. These values will give the proper weighting matrix with the correct dependences on the σ_{x_i} and σ_{y_i}. Most importantly, the adjusted σ_i^2 will give the correct covariance matrix for the fitting parameters, and when used in the denominator of the chi-square sum of Eq. 8.5, will maintain its proper expectation value. This is a critical aspect of the chi-square test and will be discussed shortly.

There is one minor problem to address with this technique. The fitting parameters of $F(\{a\};x_i)$ would need to be known in order to evaluate the partial derivative in Eq. 8.64. Thus an iterative approach would be necessary, perhaps starting with an equally-weighted fit or a fit neglecting uncertainties in the independent variables. After an initial parameter set is obtained, σ_i^2 can then be calculated from Eq. 8.64 and the fit would be repeated. If necessary, additional iterations could be performed to be sure the solution has converged.

Data sets with a calibration

An instrument calibration typically involves making measurements on one or more *standards*—samples or sources with known values for the quantity measured by the instrument. For example, our transmission grating spectrometer is used to determine the wavelengths of light emitted from a source. Calibrating this spectrometer involves measuring the diffraction angles for spectral lines from standard sources of known wavelengths and fitting them to a calibration equation involving the grating groove spacing and other apparatus parameters.

A calibration fitting function (with one independent variable x) might be expressed

$$y_i^{\text{fit}} = F(\{a\};x_i) \tag{8.65}$$

where $\{a\}$ represents its M fitting parameters.

The dependent y_i for the fit are the known values for the standards and together with their corresponding x_i determine the best-fit a_k using a standard regression analysis. The y_i are normally of such high accuracy that uncertainties in the x_i play a dominant role in the regression analysis. Consequently, the uncertainties in the x_i must be small enough to perform a valid fit as described in the previous section. This regression analysis then determines the calibration parameters a_k and their covariance matrix $[\sigma_a]$.

With the calibration results in hand, the instrument is then used again on some new system of interest—one where the y-values are not known in advance. To this end, a second set of x-measurements, x'_i, are made and used with the previously fitted calibration function to determine a set of y'_i-values, for the new system

$$y'_i = F(\{a\}; x'_i) \tag{8.66}$$

The y'_i will now be considered "measured" y-values for the new system and will be used in a separate analysis. For our spectrometer example, these measured wavelengths are from a hydrogen source and are used in a fit to the Balmer formula. The second fit will then need as input the covariance matrix for the y'_i. Assuming there are N data points in this second fit, this $N \times N$ covariance matrix depends on the results of the calibration and the uncertainties in the measured x'_i.

If the x'_i were known exactly, the covariance matrix $[\sigma'_y]$ for the y'_i would be given by the propagation of error formula (Eq. 6.11)

$$[\sigma'_y] = [J][\sigma_a][J^T] \tag{8.67}$$

where $[\sigma_a]$ is the $M \times M$ covariance matrix for the calibration parameters determined in the calibration step and the $N \times M$ Jacobian matrix $[J]$ has elements

$$[J]_{ik} = \frac{\partial y'_i}{\partial a_k} \tag{8.68}$$

If the x'_i are statistically independent, their standard deviations, $\sigma_{x'_i}$, would add $(\partial y'_i/\partial x'_i)^2\sigma^2_{x'_i}$ to the diagonal elements of $[\sigma'_y]$. The elements of $[\sigma'_y]$ then become

$$[\sigma'_y]_{ij} \quad = \quad \sum_{k=1}^{M}\sum_{l=1}^{M}[J]_{ik}[\sigma_a]_{kl}[J^T]_{lj} + \delta_{ij}\left(\frac{\partial y'_i}{\partial x'_i}\right)^2 \sigma^2_{x'_i} \tag{8.69}$$

The first term is simply an explicit expression for the ijth element of Eq. 8.67 and the second term is the additional diagonal contribution from $\sigma_{x'_i}$.

Now, when the y'_i and their covariance matrix $[\sigma'_y]$ are used as input for the second fit, there will be a new twist—$[\sigma'_y]$ is not diagonal and thus the y_i will not be independent. This issue is discussed next.

Regression with correlated $\boldsymbol{y_i}$

Performing a fit to a set of y_i having a non-diagonal covariance matrix $[\sigma_y]$ is relatively simple. Assuming the joint probability distribution for the y_i is reasonably well-described by the correlated Gaussian of Eq. 5.21, the regression formulas already presented remain valid without modification. One need only substitute the non-diagonal covariance matrix and its inverse for the diagonal versions assumed up to now.

This simple substitution works because the log likelihood for the correlated joint probability of Eq. 5.21 when multiplied by -2 still depends on the μ_i only via a χ^2 of the form

$$\chi^2 = \left(\boldsymbol{y}^T - \boldsymbol{\mu}^T\right)\left[\sigma_y^{-1}\right]\left(\boldsymbol{y} - \boldsymbol{\mu}\right) \tag{8.70}$$

As with a diagonal $[\sigma_y]$, when the best-fit parameters a_k are found, they will determine the best-fit $y_i^{\rm fit}$ via the fitting function, which when used for the μ_i in Eq. 8.70 produce a minimum χ^2.

$$\chi^2 = \left(\boldsymbol{y}^T - \boldsymbol{y}^{{\rm fit}\,T}\right)\left[\sigma_y^{-1}\right]\left(\boldsymbol{y} - \boldsymbol{y}^{\rm fit}\right) \tag{8.71}$$

That this χ^2 is a minimum with respect to all fitting parameters, implies that its derivative with respect to every a_k is zero. Performing this chain-rule differentiation then gives

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_k}\left(\boldsymbol{y}^T - \boldsymbol{y}^{{\rm fit}\,T}\right)\left[\sigma_y^{-1}\right]\left(\boldsymbol{y} - \boldsymbol{y}^{\rm fit}\right) \\ &= -\left(\boldsymbol{y}^T - \boldsymbol{y}^{{\rm fit}\,T}\right)\left[\sigma_y^{-1}\right]\frac{\partial \boldsymbol{y}_i^{\rm fit}}{\partial a_k} - \frac{\partial \boldsymbol{y}^{{\rm fit}\,T}}{\partial a_k}\left[\sigma_y^{-1}\right]\left(\boldsymbol{y} - \boldsymbol{y}^{\rm fit}\right) \end{aligned} \tag{8.72}$$

The two terms in this last equation are scalars. In fact, they are the exact same scalar, just formed from expression that are transposes of one another. Thus each must be zero at the best fit and choosing the second of these gives

$$\frac{\partial \boldsymbol{y}^{{\rm fit}\,T}}{\partial a_k}\left[\sigma_y^{-1}\right]\boldsymbol{y} = \frac{\partial \boldsymbol{y}^{{\rm fit}\,T}}{\partial a_k}\left[\sigma_y^{-1}\right]\boldsymbol{y}^{\rm fit} \tag{8.73}$$

This scalar equation must be true for each of the M fitting parameters a_k and with the definition of the Jacobian (Eq. 8.16), all M can be rewritten in the vector form of Eq. 8.17. Because Eq. 8.17 was the starting point for the derivation of the regression results already presented, and because the derivation does not rely on the diagonality of $[\sigma_y]$, the equations for the best-fit parameters and their covariance matrix do not change when $[\sigma_y]$ is non-diagonal.

Chapter 9

Evaluating a Fit

Evaluating the agreement between a fitting function and a data set should begin with a graph.

The main graph should show the fitting function as a smooth curve without markers for a set of x-values (not necessarily only the x_i of the data points) that give a good representation of the best fit curve throughout the fitting region. The x_i, y_i data points should not have connecting lines but should include error bars—vertical line segments extending one standard deviation above and below each point. If there are x-uncertainties, horizontal error bars should also be placed on each point.

Figure 9.1 shows a case where the error bars would be too small to show clearly on the main graph, The fix is shown below the main graph—a plot of *residuals* or raw deviations, $y_i - y_i^{\rm fit}$, with error bars. If the σ_i vary too widely to all show clearly on a residual plot, logarithmic or other nonlinear y-axis scaling may fix the problem. Or, normalized deviations $(y_i - y_i^{\rm fit})/\sigma_i$ (without error bars) could be used.

The purpose of these graphs is to make it easy to see each data point's deviation relative to its standard deviation and to assess the entire set of deviations for their expected randomness. If the sample size is large enough, deviations or normalized deviations can be histogrammed and checked against the expected distribution.

Specifically look for the following "bad fit" problems and possible causes.

- Deviations are non-random and show some kind of trend. For example, data points are mostly above or mostly below the fit, or mostly above at one end and mostly below at the other. Deviations should be random.

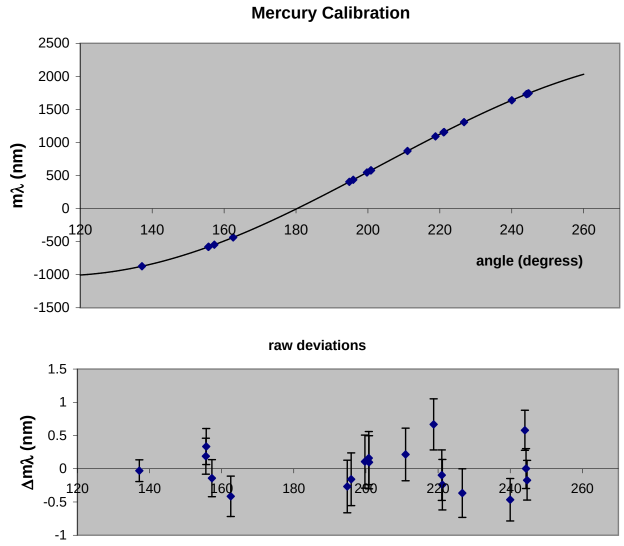


Figure 9.1: Top: main graph for a fit to a calibration function for data from our visible spectrometer. Bottom: corresponding residual plot.

In particular, positive and negative deviations are equally likely and should be present in roughly equal numbers. Systematic deviations may indicate a problem with the fitting program or a fitting model that is incomplete or wrong.

- Too many error bars miss the fitted curve. Approximately two-thirds of the error bars should cross the fit. If the deviations are random and simply appear larger than predicted, the σ_i may be underestimated.

- Outliers—points missing the fit by three or more σ_i. These should be very rare and may indicate data entry mistakes, incorrect assignment of x_i, or other problems.

- The fit goes through most data points near the middle of the error bars. On average, the y_i should miss the fit by one error bar and about one-third of the error bars should miss the fit entirely. This should probably be called a "good fit" problem. It is not all that unlikely if there are only a few data points, but with a sufficient number of data points, it indicates the σ_i have been overestimated—the measurements are more precise than expected.

The chi-square distribution

The χ^2 value after the best-fit has been found can be used with the *chi-square test* to quantitatively assess the size of the residuals in relation to the assumed standard deviations. Keep in mind, however, that the chi-square for a fit is an aggregate statistic and a statistical test of its value does not replace the point-by-point assessment of a graphical analysis.

Whether χ^2 or χ^2_P was minimized in the fit, only the standard χ^2 calculated with Eq. 8.5 (or Eq. 8.70 if the weighting matrix is not diagonal) has the properties under consideration now. If the y_i are Poisson-distributed, use $\sigma_i^2 = y^{\text{fit}}$ for calculating this χ^2. The χ^2 is a random variable—were a new set of y_i acquired and analyzed, a different χ^2 value should be expected. To determine whether or not a particular χ^2 value is reasonable, its probability distribution must be known.

The χ^2 probability distribution depends on the particular probability distributions governing the y_i as well as the number of data points N and the number of fitting parameters M. The quantity $N - M$ is called the chi-square's *degrees of freedom* and plays a central role in describing the distribution. Each data point adds one to the degrees of freedom and each fitted parameter subtracts one. For example, there are no degrees of freedom for a straight-line fit to two data points. The fit can always be made to pass exactly through both points and the χ^2 will always be zero.

The mean of the χ^2 distribution is most important and its evaluation begins with the definition, Eq. 2.17, for the variance of each y_i—rewritten in the form

$$\left\langle \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right\rangle = 1 \tag{9.1}$$

Summing over all N data points gives

$$\left\langle \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right\rangle = N \tag{9.2}$$

The sum in this expression is the χ^2 of Eq. 8.2. However, the true μ_i are normally unknown and the χ^2 must be evaluated with Eq. 8.5 using the y_i^{fit} instead. Is Eq. 9.2 valid if y_i^{fit} is substituted for μ_i? It turns out that the chi-square calculated with the y_i^{fit} satisfies

$$\left\langle \sum_{i=1}^{N} \frac{(y_i - y_i^{\text{fit}})^2}{\sigma_i^2} \right\rangle = N - M \tag{9.3}$$

The deviations from $y_i^{\rm fit}$ are smaller than those from μ_i and the expected average χ^2 is reduced by the number of fitting parameters.

The expectation value of each of the N terms in Eq. 9.3 is the same; each has an expectation value of $(N - M)/N$. Thus for each data point:

$$\left\langle \left(y_i - y_i^{\rm fit}\right)^2 \right\rangle = \left(1 - \frac{M}{N}\right) \sigma_i^2 \tag{9.4}$$

Contrast this with Eq. 2.17, here

$$\left\langle (y_i - \mu_i)^2 \right\rangle = \sigma_i^2 \tag{9.5}$$

which defines the mean squared deviation from the true mean as the true variance σ_i^2. Equation 9.4 then implies that the mean squared deviation from the best fit is reduced somewhat from the true variance—by the factor $1 - M/N$.

The proof of Eq. 9.3 is given in the *Regression Analysis Addendum.* It is still based on Eq. 9.1 and thus valid for any distribution for the y_i. In effect, it was demonstrated in Exercise 7 for the case of the constant fitting function with equally weighted y-values.

It is easy to understand that some reduction in the χ^2 should be expected when $y_i^{\rm fit}$ replaces μ_i. Simply consider an iterative fitting program with an initial guess for the fitting parameters equal to the true parameters. The fitting function then starts with the true means μ_i and the starting χ^2 is as calculated from Eq. 8.2. The expected average starting χ^2 would then be given by Eq. 9.2 and thus equal to N.

When the fitting program is then run from this initial guess, either no further improvement in the fit will be achieved and the fitting parameters and the χ^2 will not change, or, as will usually be the case, some decrease in the χ^2 can be achieved by adjusting the parameters from the starting, true values. As the χ^2 after the best fit can only be equal to or less than the starting χ^2, the expected average χ^2 using the best-fit $y_i^{\rm fit}$ must be less than N. Thus, the reduction is a non-negative random variable. For Gaussian-distributed y_i, the reduction is a chi-square random variable with M degrees of freedom.

According to Eq. 9.3, the mean of the chi-square distribution is $N - M$. How high above (or below) the expected mean does the χ^2 value have to be before one must conclude that it is too big (or too small) to be reasonable? That question calls into play the width or variance of the χ^2 distribution.

Unlike the mean of the χ^2 distribution, its variance depends on the probability distribution for the y_i. For the case where no fitting is involved and the χ^2 can be evaluated using Eq. 8.2, the variance is readily predicted starting from Eq. 2.18.

$$\begin{aligned}
\sigma^2_{\chi^2} &= \left\langle \left(\chi^2\right)^2 \right\rangle - \left(\left\langle \chi^2 \right\rangle\right)^2 \\
&= \left\langle \left(\sum_{i=1}^{N} \frac{(y_i-\mu_i)^2}{\sigma_i^2} \right) \left(\sum_{j=1}^{N} \frac{(y_j-\mu_j)^2}{\sigma_j^2} \right) \right\rangle - N^2 \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \left(\frac{(y_i-\mu_i)^2}{\sigma_i^2} \right) \left(\frac{(y_j-\mu_j)^2}{\sigma_j^2} \right) \right\rangle - N^2 \qquad (9.6)
\end{aligned}$$

where Eq. 9.2 has been used for the expectation value of χ^2.

In the double sum, there are $N^2 - N$ terms where $i \neq j$ and N terms where $i = j$. Assuming all y_i are statistically independent, Eq. 5.6 applies and thus each of the terms with $i \neq j$ has an expectation value of one—equal to the product of the expectation value of its two factors (one by Eq. 9.1). The N terms with $i = j$ become a single sum of terms of the form: $\langle (y_i - \mu_i)^4/\sigma_i^4 \rangle$. Making these substitutions in Eq. 9.6 gives

$$\sigma^2_{\chi^2} = \sum_{i=1}^{N} \left\langle \left(\frac{(y_i-\mu_i)^4}{\sigma_i^4} \right) \right\rangle - N \qquad (9.7)$$

For y_i governed by a Gaussian distribution, $\langle (y_i - \mu_i)^4/\sigma_i^4 \rangle = 3$, giving

$$\sigma^2_{\chi^2} = 2N \qquad (9.8)$$

For y_i governed by a uniform distribution, $\langle (y_i - \mu_i)^4/\sigma_i^4 \rangle = 1.8$ giving

$$\sigma^2_{\chi^2} = 0.8N \qquad (9.9)$$

For y_i governed by a Poisson distribution (where $\sigma_i^2 = \mu_i$), $\langle (y_i - \mu_i)^4/\sigma_i^4 \rangle = 3 + 1/\mu_i$ giving

$$\sigma^2_{\chi^2} = 2N + \sum_{i=1}^{N} \frac{1}{\mu_i} \qquad (9.10)$$

For y_i governed by a binomial distribution, $\langle (y_i - \mu_i)^4/\sigma_i^4\rangle = 3 + (1 - 5p_i + p_i^2/(1-p_i))/\mu_i$ giving

$$\sigma^2_{\chi^2} \;=\; 2N + \sum_{i=1}^{N} \frac{1 - 5p_i + p_i^2/(1-p_i)}{\mu_i} \tag{9.11}$$

Of course, the mean and variance do not provide a complete description of the chi-square distribution. The detailed shape can be found in references and statistics software, but only for the standard chi-square distribution where the y_i follow a Gaussian distribution. Chi-square distributions for y-variables following non-Gaussian distributions are not usually discussed in the literature.

Equations 9.8-9.11 assume the μ_i are known so that the χ^2 could be calculated with Eq. 8.2. How does the chi-square distribution change when fitting is involved and the χ^2 must be evaluated with Eq. 8.5 using the $y_i^{\rm fit}$ instead? For the case of Gaussian-distributed y_i, the chi-square value follows a standard chi-square distribution with $N - M$ degrees of freedom; the mean decreases to $N - M$ and the variance decreases to $2(N - M)$. However, for y-values governed by non-Gaussian distributions, exactly how the shape of the distribution changes has not been studied in detail. *Chi-Square distributions.xls* on the lab website has simulations for a fit to a mean ($M = 1$) showing the chi-square variance decreased by 2 for Gaussian-distributed data, increased slightly for uniformly-distributed data, and decreased by more than 2 for Poisson- and binomial-distributed data.

The standard chi-square distribution will be assumed appropriate in the following discussions, but if evaluating χ^2 probabilities is an important aspect of the analysis, keep this assumption in mind. For example, it would directly affect the chi-square test, discussed next.

The chi-square test

The chi-square test uses the χ^2 distribution to decide whether a χ^2 value from a fit is too large or too small to be reasonably probable.

One first compares the χ^2 from the fit to $N - M$; the degrees of freedom and the mean of its distribution. For an example, suppose $N - M = 50$. The standard deviation of the χ^2 distribution will be $\sigma_{\chi^2} = \sqrt{2(N-M)} = 10$.

While the χ^2 distribution is not exactly Gaussian, χ^2 values outside the two-sigma range from 30-70 might be cause for concern.

So suppose the actual χ^2 value from the fit is significantly above $N-M$ and the analysis must decide if it is too big. To decide the issue, the chi-square distribution is used to determine the probability of getting a value as large or larger than the actual χ^2 value from the fit. If this probability is too small to be accepted as a chance occurrence, one must conclude that the χ^2 is unreasonably large.

In rare cases, the χ^2 value from the fit may come out too small—well under the expected value of $N-M$. To check if it is too low, the chi-square distribution should be used to find the probability of getting a value that small or smaller. If this probability is too low to be accepted as a chance occurrence, one must conclude that the χ^2 is unreasonably small. Keep in mind, however, that there should be at least three degrees of freedom to test for an undersized χ^2. With only one or two degrees of freedom, the χ^2 pdf is non-zero at $\chi^2 = 0$ and decreases monotonically as χ^2 increases. Thus, for these two case, smaller values are always more likely than larger values.

If the χ^2 is unacceptably large or small, the deviations are not in accord with predictions and the experimental model and theoretical model are incompatible. The same problems mentioned earlier for a graphical assessment may be applicable to an unacceptable χ^2.

When the σ_i are unknown

It is not uncommon for an experimentalist to be unsure of the correct σ_i to use in a regression analysis. If the σ_i are unknown, the chi-square variable can not be calculated and the chi-square test can not be directly applied. Instead, unknown σ_i can be estimated by choosing values that make the χ^2 equal to its expectation value of $N-M$. This can be a useful procedure because the σ_i determine the weighting matrix $\left[\sigma_y^{-1}\right]$ (Eq. 8.15) and via Eq. 8.30 with Eq. 8.24 they also determine the parameter covariance matrix $[\sigma_a]$. If the σ_i are unreliable or wrong, the parameter uncertainties would be unreliable or wrong as well. Forcing $\chi^2 = N-M$ is a valid method for adjusting uncertain σ_i to achieve a predictable level of confidence in the parameter uncertainties.

Basically, the technique will use the scatter of the data about the best fit to set the scale of the σ_i. If the scatter is small the σ_i will be small, and vice versa. The technique is particularly suited for large sample sizes where the actual scatter should accurately predict the true σ_i. However, for small degrees of freedom, the situation is not so clear. When $N - M$ is small, significantly high or low sample deviations are more likely and would, by chance, lead to significantly high or low parameter uncertainties. This issue is addressed in the section on the Student-T probabilities.

For an equally-weighted data set, the technique is straight-forward. Find that one value for σ_y that gives $\chi^2 = N - M$. Equation 8.44 shows this would happen if the following sample variance were used for σ_y^2:

$$s_y^2 = \frac{1}{N-M}\sum_{i=1}^{N}\left(y_i - y_i^{\text{fit}}\right)^2 \tag{9.12}$$

Using s_y^2 for σ_y^2 in Eq. 8.42 for the parameter covariance matrix shows $[\sigma_a]$ would then be proportional to this sample variance.

Note how Eq. 9.12 is a generalized version of Eq. 2.22. (The latter gives the sample variance for a sample set all from the same distribution.) For example, Exercise 7 and Eq. 9.4 imply that both sample variances are unbiased estimates of the true variance. Equation 9.12 is the general case where the N means are individually estimated by the y_i^{fit}, which are based on the M best-fit parameters of the fitting function. Equation 2.22 is just a specific case where all means are the same and estimated by the single best-fit parameter $\bar{y}$.

If the σ_i vary from point to point, their relative sizes must be known in advance. Forcing $\chi^2 = N - M$ would then determine a single overall scale factor for all σ_i. Relatively-sized σ_i might occur when measuring wide-ranging quantities with instruments having multiple scales or measurement ranges. In such cases, the measurement uncertainty typically scales with the instrument range used for the measurement.

Initial values for the σ_i would be set in accordance with the known ratios and then a single multiplier would be determined to achieve a chi-square value of $N - M$. Scaling all the σ_i by a common factor κ scales the χ^2 by a factor of $1/\kappa^2$. The data covariance matrix $[\sigma_y]$ would scale by κ^2 and Eq. 8.30 with Eq. 8.24 implies that the parameter covariance matrix $[\sigma_a]$ would scale by κ^2 as well. On the other hand, Eq. 8.27 with the equations for $\left[J^\dagger\right]$ show that the fit parameters are unaffected by the scale factor. This is because scaling all σ_i together does not change the relative weighting of the data points.

When the σ_i are known, the normal randomness in the data set deviations leads to a randomness in the s_y^2 or in the χ^2 value for the fit. When the σ_i are scaled to achieve $\chi^2 = N - M$ that randomness is transferred to the parameter covariance matrix, which then becomes a random variable. It becomes a sample covariance matrix. Parameter variances obtained from its diagonal elements should then be considered sample variances and their square roots should be considered sample standard deviations. As discussed next, confidence intervals constructed with sample standard deviations have different confidence levels than those constructed with true standard deviations.

Should the σ_i always be scaled to give $\chi^2 = N - M$? No. It is only appropriate if the σ_i are experimentally uncertain. And even then, a coarse estimate of the σ_i based on experimental considerations should still be made. Scaling these σ_i by factor of two to achieve the expected chi-square value may be deemed acceptable, but scaling them by a factor of ten might not.

The reduced chi-square distribution

Dividing a chi-square random variable by its degrees of freedom $N-M$ gives another random variable called the *reduced chi-square* random variable.

$$\chi^2_\nu = \frac{\chi^2}{N-M} \tag{9.13}$$

Reduced chi-square distributions for various degrees of freedom are shown in Fig. 9.2. A table of reduced chi-square values/probabilities is given in Table 10.3. The table can also be used for determining χ^2 probabilities using the scaling above. For example, with 100 degrees of freedom, the probability a χ^2 will exceed 120, is the same as the probability that a χ^2_ν (with 100 degrees of freedom) will exceed 1.2, which is about 8 percent.

Dividing any random variable by a constant will divide its distribution mean by that constant and its variance by the square of that constant. Thus, the reduced chi-square distribution will have a mean of one for all degrees of freedom and a variance equal to $2/(N-M)$.

Dividing both sides of Eq. 9.12 by σ_y^2 and eliminating the sum using Eq. 8.44 gives

$$\frac{s_y^2}{\sigma_y^2} = \frac{\chi^2}{N-M} \tag{9.14}$$

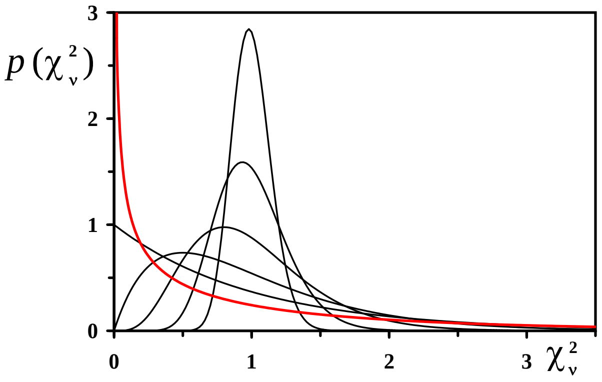


Figure 9.2: The reduced chi-square pdfs $p(\chi^2_\nu)$ for degrees of freedom (dof) $1, 2, 4, 10, 30, 100$. The tall distribution peaking at $\chi^2_\nu = 1$ is for dof $= 100$. The curves get broader and lower as the dof decrease. For dof $= 1$, the distribution (red) is singular at zero.

and shows that the ratio of the sample variance to the true variance is a reduced chi-square variable with $N - M$ degrees of freedom.

The fact that the variance of χ^2_ν decreases as the sample size N increases, implies that the reduced chi-square distribution becomes more sharply peaked around its expectation value of one. This makes good sense as it predicts values of s_y are more likely to be near σ_y as N increases, i.e., the s_y determined from the data becomes a more precise estimate of the true standard deviation σ_y.

For large $N - M$, the chi-square and the reduced chi-square distributions are approximately Gaussian—the former with a mean of $N - M$ and standard deviation of $\sqrt{2(N - M)}$, the latter with a mean of one and a standard deviation of $\sqrt{2/(N - M)}$. This approximation is used in the next exercise.

Exercise 8 *It is often stated that uncertainties should be expressed with only one significant figure. Some say two figures should be kept if the first digit is 1. Roughly speaking, this suggests uncertainties are only good to about 10%. Suppose you take a sample set of y_i and evaluate the sample mean $\bar{y}$. For the uncertainty, you use the sample standard deviation of the mean $s_{\bar{y}}$. Show that it takes around 200 samples if one is to be about 95% confident that $s_{\bar{y}}$ is within 10% of $\sigma_{\bar{y}}$. Hint: s_y will also have to be within 10% of σ_y. Thus, you want to find the value of N such that the probability $P(0.9\sigma_y < s_y <$*

$1.1\sigma_y) \approx 0.95$. *Convert this to a probability on χ^2_ν and use near-Gaussian limiting behavior appropriate for large sample sizes. Then show you can use Table 10.3 and check your results.*

Student-T probabilities

Consider a fitting parameter value a_k and its sample standard deviation s_k obtained by forcing $\chi^2 = N - M$. These quantities are best estimates of the true mean α_k and its true standard deviation σ_k. Now suppose one wishes to find a 68% or 95% confidence interval for the unknown mean α_k.

If a random variable y follows a Gaussian pdf, y-values in the range $\mu_y \pm \sigma_y$ occurs 68% of the time and a y value in the range $\mu_y \pm 2\sigma_y$ occurs 95% of the time. This logic is invertible and one can construct confidence intervals of the form

$$y \pm z\sigma_y$$

for any value of z and the probability such an interval will include the true mean μ_y will likewise follow from the Gaussian pdf; 68% for $z = 1$, 95% for $z = 2$, etc. Such confidence intervals and associated probabilities are seldom reported because they are well known and completely specified once y and σ_y are given.

When using a fitting parameter and its sample standard deviation, the situation changes. One can again express a confidence interval in the form

$$a_k \pm z s_k$$

However, now that the interval is constructed with an estimate rather than a true standard deviation, $z = 1$ (or $z = 2$) are not necessarily 68% (or 95%) likely to include the true value. William Sealy Gosset, publishing around 1900 under the pseudonym "Student" was the first to determine these "Student-T" probabilities.

A difference arises because s_k might, by chance, come out larger or smaller than σ_k. Recall its size will be related to the random scatter of the data about the best fit. When the probabilities for all possible values of s_k are properly taken into account, the confidence level for any z is always smaller than would be predicted based on the Gaussian pdf.

In effect, the uncertainty in how well s_k estimates σ_k decreases the confidence level for a given z. Because the uncertainty in s_k depends on the degrees of freedom, the Student-T confidence intervals also depend on the degrees of freedom. The larger the degrees of freedom, the better the estimate $s_{\bar{y}}$ becomes and the closer the Student-T intervals will be to the corresponding Gaussian intervals.

Table 10.4 at the end of this paper gives some Student-T probabilities. As an example of its use, consider five sample values from which are obtained $\bar{y}$ and $s_{\bar{y}}$. There are four degrees of freedom for an $s_{\bar{y}}$ calculated from five samples. Looking at the row for $\nu = 4$, the entry in the second column indicates a 95% probability that the interval $\bar{y} \pm 2.78 s_{\bar{y}}$ will include the true mean μ_y. If one were ignorant of the Student-T probabilities one might have assumed that a 95% confidence interval would be, as for a Gaussian, $\bar{y} \pm 2\, s_{\bar{y}}$.

Exercise 9 *Three sample values from a Gaussian pdf are 1.20, 1.24, and 1.19. (a) Find the sample mean, sample standard deviation, and sample standard deviation of the mean and give the 68% and 95% confidence intervals for the true mean based on this data alone. (b) Now assume those three sample values are known to come from a pdf with a standard deviation* $\sigma_y = 0.02$*. With this assumption, what are the 68% and 95% confidence intervals? Determine the reduced chi-square and give the probability it would be this big or bigger.*

Chapter 10

Regression with Excel

Excel is more business software than scientific software. Nonetheless, together with Visual Basic for Applications, which is always available within Excel, it is a reasonable platform for most statistical analysis tasks. Other scientific/mathematical software can offer better tools for graphing and analysis, but often come with a steep learning curve. Excel has the advantage that most students already know enough to get started immediately with data entry, formula evaluations, and graphing and a familiarity with these topics is assumed. The following sections provide guidance only on aspects of using Excel for regression.

Excel has *array formulas* for working with vectors (one-dimensional arrays) and matrices (two-dimensional arrays). To use an array formula, select the appropriate rectangular block of cells (one- or two-dimensional, or even a single cell), click in the edit box, enter the array formula and end the process with the three-key combination Ctrl|Shift|Enter. The entire block of cells is evaluated according to the formula and the results are placed in the block. Various odd behaviors or errors result if the selected block is not of the appropriate size.

The built-in array functions needed for linear regression formulas are:

TRANSPOSE(array): Returns the transpose of a vector or matrix.

MMULT(array1, array2): Returns the result of vector and matrix array multiplication.

MINVERSE(array): Returns the inverse of a square matrix.

The *StatsMats.xla* Excel add-in can be found on the lab web site and adds the following array formulas specifically designed for regression:

CovarMat(array) Takes a one-dimensional input array (row or column form) of σ_i values and returns the corresponding square diagonal covariance matrix for independent variables.

WeightMat(array) Like CovarMat, but returns the diagonal weighting matrix.

SDFromCovar(array) Takes a two-dimensional covariance matrix and returns a one-dimensional array of standard deviations from its diagonal elements.

See the *Linear Regression.xls* spreadsheet on the lab web site for an application of array functions to linear regression formulas.

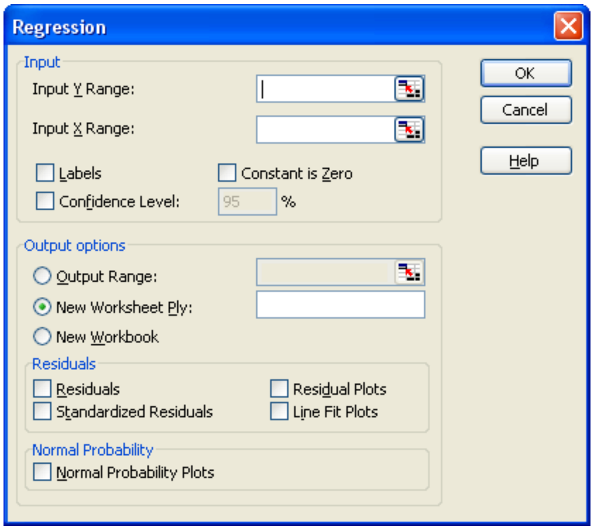
Linear Regression with Excel

Excel's linear regression program is only for equally-weighted fits. For non-equally-weighted data sets, you can program the regression formulas yourself or use nonlinear regression. Excel's linear regression program does not allow the user to provide σ_y and uses the s_y of Eq. 9.12 for σ_y in determining the parameter covariance matrix. That is, it forces $\chi^2 = N - M$.

The starting point is to set up columns for the y_i and for the $f_k(x_i)$. The steps will be illustrated for a quadratic fit: $y_i^{\text{fit}} = a_1 + a_2 x_i + a_3 x_i^2$. Thus, $f_1(x_i) = 1$, $f_2(x_i) = x_i$, and $f_3(x_i) = x_i^2$. In turns out to be unnecessary to have a column for the constant term; Excel can handle the constant function without using a column. A column for y_i and side-by-side columns x_i and x_i^2 must be constructed.

Excel's linear regression program is found in the Tools|Data Analysis|Regression menu. The dialog box for this procedure appears as in Fig. 10.1.

Select the column containing the y_i-values for the Input Y-Range. Select what must be a contiguous rectangular block containing all $f_k(x_i)$ values (two columns for the quadratic fit). Leave the Constant is Zero box unchecked. (It would be checked if the fitting function did not include a constant term.) Leave the Labels box unchecked unless your x- and y-ranges include labels at the top of the columns. If you would like, check the Confidence Level box and supply a probability for a Student-T interval next to it. (Intervals for the



Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Figure 10.1: Excel's linear regression dialog box

95% confidence level are provided automatically.) Select the New Worksheet Ply radio button or the Output Range. For the latter, also specify (in the edit box to its right) the upper left corner of an empty spreadsheet area for the results. Select any of the various options in the lower part of the dialog box and then click OK.

The upper *Regression Statistics* area contains the sample standard deviation s_y (from Eq. 9.12 and referred to as the Standard Error). The lower area contains information about the constant (labeled *Intercept*) and the fitting parameters a_k (labeled X Variable k). Next to the best fit values (labeled *Coefficients*) are the parameter sample standard deviations s_k (labeled *Standard Error*). Then there are columns for the t *Stat* and *P-value* for each parameter. And lastly two double columns for the lower and upper limits of intervals at confidence levels of 95% and the user specified percentage.

Exercise 10 *Fit the data in Table 10.1 to a quadratic formula* ($y_i^{fit} = a + bx_i + cx_i^2$) *using the Excel linear regression program and submit a printout of the appropriate worksheets. Save this spreadsheet. It will be used again for Exercises 11 and 12. (a) What does Excel use as a best estimate for the value of* σ_i *appearing in the linear regression equations? Describe and give the formula for this quantity. Circle and label its value on the worksheet.*

x_i	y_i
2	2.4
3	6.7
5	27.8
6	43.2
8	80.7
9	104.5

Table 10.1: Data for Exercise 10

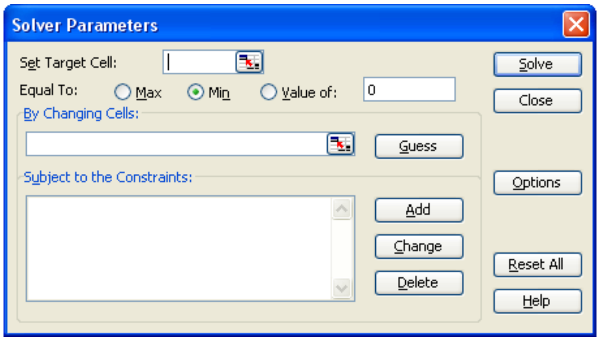
(b) The parameter standard deviations given on the worksheet are sample standard deviations s_k. Circle and label the parameters a_k and their sample standard deviations s_k. (c) Circle the program's 95% confidence interval (for the quadratic coefficient only) and show how it can be obtained from a_k, s_k and the Student-T table.

Exercise 11 *Open the spreadsheet from the previous exercise. (a) Add a column for y^{fit} based on the fit parameters, another for the deviations $(y_i - y_i^{\mathrm{fit}})$, and another for their squares $(y_i - y_i^{\mathrm{fit}})^2$. Use this last column to evaluate s_y and show that it agrees with Excel's value. (b) Suppose the uncertainties in the y_i values for that exercise are known to be $\sigma_y = 0.5$. Use the scaling rule to determine the parameter uncertainties in this case. Give the 95% confidence interval for the quadratic coefficient. Should you use Student-T or Gaussian probabilities? (c) Submit a graph of y_i vs. x_i with error bars and overlay the fit as a smooth curve.*

Nonlinear regression with Excel

Excel's Solver can be used for nonlinear regression. It can be used to find the best fit parameters for linear or nonlinear functions and can take into account varying uncertainties in the input data. It does not provide the covariance matrix directly, but can do so indirectly through additional procedural steps.

The Solver does not let the user specify the Jacobian matrix $[J]$. Solver determines it numerically by evaluating the y_i^{fit} while varying the a_k fit parameters in small steps. When creating worksheets for the Solver program, keep in mind all cells contributing to the calculation of the y^{fit} and χ^2 will



Equal To: ○ Max ● Min ○ Value of: 0

Figure 10.2: Excel's Solver dialog box

be evaluated repeatedly during the iterations. If execution speed is an issue, keep the calculations as efficient as possible.

To use Solver:

1. Set up an area of the spreadsheet for the fitting parameters. Using Solver will be somewhat easier if parameters are confined to a single block of cells. Enter initial guesses for the values of each fitting parameter.

2. Enter your data in columns, including x_i and y_i. Depending on whether one or both of the x_i or y_i have uncertainties and whether or not they are the same for all data points, various additional cells and/or columns for the raw uncertainties and/or the final σ_i will have to be constructed.

3. Create a column for y_i^{fit} based on the fitting function using the addresses for the x_i and the a_k fitting parameters.

4. Create the main graph including the (x_i, y_i) data points with error bars and a smooth curve for y_i^{fit} vs. x_i. Adjust the fitting parameters to get the data and the fit close enough for the Solver program to work properly. If desired, make a separate plot of the deviations or normalized deviations versus x_i.

5. Make a column for $(y_i - y_i^{\text{fit}})^2/\sigma_i^2$. Sum this column in a separate cell to provide the χ^2 value needed for the fitting procedure.

6. Invoke the Solver from the Tools menu. The dialog box is shown in Fig. 10.2.

7. Provide the cell address of the χ^2 in the Set Target Cell: box. Click the Min radio button next to Equal To: so that the χ^2 will be minimized (as opposed to maximized or made to take on some particular value, which are also available choices).

8. Provide the cell addresses for the fitting parameters in the By Changing Cells: box.

9. Click on the Solve button. The solver's algorithms then start with your initial fitting parameters, varying them to find those values which minimize the χ^2.

10. Accept the final, changed, fitting parameters.

Parameter variances and covariances

Solver does not provide parameter variances or covariances. However, these quantities are easily calculated with the following procedures based on the $\Delta\chi^2 = 1$ rule.

11. Write down or save to a separate area the optimized values of all fitting parameters and the final χ^2. Use the spreadsheet Copy and Paste Special ... Values functions.

12. Change the value in the cell containing one of the fitting parameters, say a_k, by a bit—try to change it by what you suspect will be its uncertainty. Call this new (unoptimized) value a'_k. The χ^2 will increase because it was originally at a minimum.

13. Remove the cell containing a_k from the list of adjustable parameters and rerun the Solver. The other parameters might change a bit and the χ^2 might go down a bit, but it will still be larger than the original χ^2 for the fit.

14. If χ^2 changed by one, then the amount that a_k was changed is its standard deviation: $\sigma_k = |a'_k - a_k|$. If the change in $\Delta\chi^2$ is more (or less) than unity, the tested change in a_k is larger (or smaller) than σ_k. Equation 8.59 can then be used with the parameter change $a'_k - a_k$ and the resulting chi-square change $\Delta\chi^2$ to determine σ_k.

15. The covariances between a_k and the other parameters can also be determined by keeping track of the magnitude and sign of the changes in the other variables after the re-optimization. Using these parameter changes with the change in a_k and Eq. 8.60 then provides the covariances.

16. Check that these procedures gives roughly the same σ_k and σ_{jk} using a'_k values both above and below a_k and such that $\Delta\chi^2$ values are around 0.1 and around 10. If σ_k is roughly constant for all four cases (say within 10%), one can be reasonably assured that confidence intervals $a_k \pm z\sigma_k$ should be close to the Gaussian predictions up to $z = 3$.

Cautions

The Solver may fail to find the best fit if the initial parameter guesses are not close enough to the best-fit values. If necessary, readjust them to get the Solver to proceed to the solution. It may also help to have the magnitude of all parameters near unity. Thus, if the amplitude of an exponential decay is of order 10^6, the decay constant is of order 10^{-3}, and the background is of order 10^3, rather than perform the fit directly to Eq. 8.45 it would be wiser to fit to

$$y_i^{\rm fit} = 10^6 a_1 e^{-10^{-3} a_2 x_i} + 10^3 a_3$$

so that all fitting parameters are of order unity.

While linear regression can only be used for linear models, nonlinear techniques can be used for both linear and nonlinear models. In the next exercise, you are to use the nonlinear regression techniques just described to fit the data of Table 1 to a quadratic formula. This is the same data and model used for Exercises 10 and 11 and you will be asked to show that the same results are obtained.

Exercise 12 *Start with the spreadsheet from Exercise 11. Add a cell for σ_y and reference this cell in the spreadsheet (anywhere σ_i is needed) so you can change the value in that cell and have it updated throughout the spreadsheet. Then go through the steps to do a fit using Solver.*

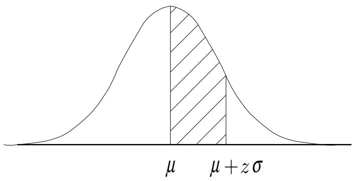
Start by setting σ_y to 2.0 and then to 3.0 running Solver for each value. (a) Demonstrate that the a_k do not depend on the value used for σ_y; that in both cases the optimized a_k are the same as those from Exercise 10. Explain why the algorithm should be insensitive to the value of σ_y.

(b) Now, assume σ_y is unknown and use the value of s_y for σ_y. Recall, this is what the Excel linear regression program does. What value of χ^2 does this produce? What value should be expected? Use the $\Delta\chi^2 = 1$ rule to determine the standard deviation of the quadratic coefficient only. Show it is the same as that obtained by Excel's linear regression.

(c) Now assume $\sigma_y = 0.5$. This value was used in Exercise 11, but in that exercise you were asked to use scaling arguments to determine how this affected the uncertainty in a_3. Here you will show that that scaling was correct. Use the $\Delta\chi^2 = 1$ rule to redetermine the uncertainty of a_3 with $\sigma_y = 0.5$. Now you have two values for the uncertainty in a_3, one using $\sigma_y = s_y$ and one using $\sigma_y = 0.5$. Show that they scale in proportion to the value used for σ_y.

Exercise 13 *(a) What is the χ^2_ν for the fit of Exercise 12 assuming $\sigma_y = 0.5$? What is the probability it would have come out this big or bigger? (b) Suppose σ_y was not known. How small would it have to be before the deviations from the fit would have to be deemed (at the 99% level) too big to be in agreement with the quadratic fitting formula? (c) How big would σ_y have to be before one would have to conclude it is too big to be reasonable (at the 99% level).*

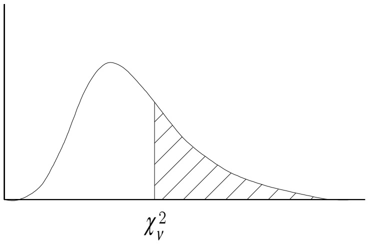
Gaussian probabilities



z	0.00	0.02	0.04	0.06	0.08
0.00	0.00000	0.00798	0.01595	0.02392	0.03188
0.10	0.03983	0.04776	0.05567	0.06356	0.07142
0.20	0.07926	0.08706	0.09483	0.10257	0.11026
0.30	0.11791	0.12552	0.13307	0.14058	0.14803
0.40	0.15542	0.16276	0.17003	0.17724	0.18439
0.50	0.19146	0.19847	0.20540	0.21226	0.21904
0.60	0.22575	0.23237	0.23891	0.24537	0.25175
0.70	0.25804	0.26424	0.27035	0.27637	0.28230
0.80	0.28814	0.29389	0.29955	0.30511	0.31057
0.90	0.31594	0.32121	0.32639	0.33147	0.33646
1.00	0.34134	0.34614	0.35083	0.35543	0.35993
1.10	0.36433	0.36864	0.37286	0.37698	0.38100
1.20	0.38493	0.38877	0.39251	0.39617	0.39973
1.30	0.40320	0.40658	0.40988	0.41308	0.41621
1.40	0.41924	0.42220	0.42507	0.42785	0.43056
1.50	0.43319	0.43574	0.43822	0.44062	0.44295
1.60	0.44520	0.44738	0.44950	0.45154	0.45352
1.70	0.45543	0.45728	0.45907	0.46080	0.46246
1.80	0.46407	0.46562	0.46712	0.46856	0.46995
1.90	0.47128	0.47257	0.47381	0.47500	0.47615
2.00	0.47725	0.47831	0.47932	0.48030	0.48124
2.10	0.48214	0.48300	0.48382	0.48461	0.48537
2.20	0.48610	0.48679	0.48745	0.48809	0.48870
2.30	0.48928	0.48983	0.49036	0.49086	0.49134
2.40	0.49180	0.49224	0.49266	0.49305	0.49343
2.50	0.49379	0.49413	0.49446	0.49477	0.49506
2.60	0.49534	0.49560	0.49585	0.49609	0.49632
2.70	0.49653	0.49674	0.49693	0.49711	0.49728
2.80	0.49744	0.49760	0.49774	0.49788	0.49801
2.90	0.49813	0.49825	0.49836	0.49846	0.49856
3.00	0.49865	0.49874	0.49882	0.49889	0.49896

Table 10.2: Half-sided integral of the Gaussian probability density function. The body of the table gives the integral probability $P(\mu < y < \mu + z\sigma)$ for values of z specified by the first column and row.

Reduced chi-square probabilities



P ν	0.99	0.98	0.95	0.9	0.8	0.2	0.1	0.05	0.02	0.01	0.001
1	0.000	0.001	0.004	0.016	0.064	1.642	2.706	3.841	5.412	6.635	10.827
2	0.010	0.020	0.051	0.105	0.223	1.609	2.303	2.996	3.912	4.605	6.908
3	0.038	0.062	0.117	0.195	0.335	1.547	2.084	2.605	3.279	3.782	5.422
4	0.074	0.107	0.178	0.266	0.412	1.497	1.945	2.372	2.917	3.319	4.617
5	0.111	0.150	0.229	0.322	0.469	1.458	1.847	2.214	2.678	3.017	4.103
6	0.145	0.189	0.273	0.367	0.512	1.426	1.774	2.099	2.506	2.802	3.743
7	0.177	0.223	0.310	0.405	0.546	1.400	1.717	2.010	2.375	2.639	3.474
8	0.206	0.254	0.342	0.436	0.574	1.379	1.670	1.938	2.271	2.511	3.265
9	0.232	0.281	0.369	0.463	0.598	1.360	1.632	1.880	2.187	2.407	3.097
10	0.256	0.306	0.394	0.487	0.618	1.344	1.599	1.831	2.116	2.321	2.959
11	0.278	0.328	0.416	0.507	0.635	1.330	1.570	1.789	2.056	2.248	2.842
12	0.298	0.348	0.436	0.525	0.651	1.318	1.546	1.752	2.004	2.185	2.742
13	0.316	0.367	0.453	0.542	0.664	1.307	1.524	1.720	1.959	2.130	2.656
14	0.333	0.383	0.469	0.556	0.676	1.296	1.505	1.692	1.919	2.082	2.580
15	0.349	0.399	0.484	0.570	0.687	1.287	1.487	1.666	1.884	2.039	2.513
16	0.363	0.413	0.498	0.582	0.697	1.279	1.471	1.644	1.852	2.000	2.453
17	0.377	0.427	0.510	0.593	0.706	1.271	1.457	1.623	1.823	1.965	2.399
18	0.390	0.439	0.522	0.604	0.714	1.264	1.444	1.604	1.797	1.934	2.351
19	0.402	0.451	0.532	0.613	0.722	1.258	1.432	1.587	1.773	1.905	2.306
20	0.413	0.462	0.543	0.622	0.729	1.252	1.421	1.571	1.751	1.878	2.266
22	0.434	0.482	0.561	0.638	0.742	1.241	1.401	1.542	1.712	1.831	2.194
24	0.452	0.500	0.577	0.652	0.753	1.231	1.383	1.517	1.678	1.791	2.132
26	0.469	0.516	0.592	0.665	0.762	1.223	1.368	1.496	1.648	1.755	2.079
28	0.484	0.530	0.605	0.676	0.771	1.215	1.354	1.476	1.622	1.724	2.032
30	0.498	0.544	0.616	0.687	0.779	1.208	1.342	1.459	1.599	1.696	1.990
32	0.511	0.556	0.627	0.696	0.786	1.202	1.331	1.444	1.578	1.671	1.953
34	0.523	0.567	0.637	0.704	0.792	1.196	1.321	1.429	1.559	1.649	1.919
36	0.534	0.577	0.646	0.712	0.798	1.191	1.311	1.417	1.541	1.628	1.888
38	0.545	0.587	0.655	0.720	0.804	1.186	1.303	1.405	1.525	1.610	1.861
40	0.554	0.596	0.663	0.726	0.809	1.182	1.295	1.394	1.511	1.592	1.835
42	0.563	0.604	0.670	0.733	0.813	1.178	1.288	1.384	1.497	1.576	1.812
44	0.572	0.612	0.677	0.738	0.818	1.174	1.281	1.375	1.485	1.562	1.790
46	0.580	0.620	0.683	0.744	0.822	1.170	1.275	1.366	1.473	1.548	1.770
48	0.587	0.627	0.690	0.749	0.825	1.167	1.269	1.358	1.462	1.535	1.751
50	0.594	0.633	0.695	0.754	0.829	1.163	1.263	1.350	1.452	1.523	1.733
60	0.625	0.662	0.720	0.774	0.844	1.150	1.240	1.318	1.410	1.473	1.660
70	0.649	0.684	0.739	0.790	0.856	1.139	1.222	1.293	1.377	1.435	1.605
80	0.669	0.703	0.755	0.803	0.865	1.130	1.207	1.273	1.351	1.404	1.560
90	0.686	0.718	0.768	0.814	0.873	1.123	1.195	1.257	1.329	1.379	1.525
100	0.701	0.731	0.779	0.824	0.879	1.117	1.185	1.243	1.311	1.358	1.494
120	0.724	0.753	0.798	0.839	0.890	1.107	1.169	1.221	1.283	1.325	1.447
140	0.743	0.770	0.812	0.850	0.898	1.099	1.156	1.204	1.261	1.299	1.410
160	0.758	0.784	0.823	0.860	0.905	1.093	1.146	1.191	1.243	1.278	1.381
180	0.771	0.796	0.833	0.868	0.910	1.087	1.137	1.179	1.228	1.261	1.358
200	0.782	0.806	0.841	0.874	0.915	1.083	1.130	1.170	1.216	1.247	1.338

Table 10.3: Integral of the χ^2_ν probability density function for various values of the number of degrees of freedom ν. The body of the table contains values of χ^2_ν, such that the probability P of exceeding this value is given at the top of the column.

Student-T probabilities

P	0.99	0.95	0.90	0.80	0.70	0.68	0.60	0.50
ν								
1	63.6559	12.70615	6.31375	3.07768	1.96261	1.81899	1.37638	1.00000
2	9.92499	4.30266	2.91999	1.88562	1.38621	1.31158	1.06066	0.81650
3	5.84085	3.18245	2.35336	1.63775	1.24978	1.18893	0.97847	0.76489
4	4.60408	2.77645	2.13185	1.53321	1.18957	1.13440	0.94096	0.74070
5	4.03212	2.57058	2.01505	1.47588	1.15577	1.10367	0.91954	0.72669
6	3.70743	2.44691	1.94318	1.43976	1.13416	1.08398	0.90570	0.71756
7	3.49948	2.36462	1.89458	1.41492	1.11916	1.07029	0.89603	0.71114
8	3.35538	2.30601	1.85955	1.39682	1.10815	1.06022	0.88889	0.70639
9	3.24984	2.26216	1.83311	1.38303	1.09972	1.05252	0.88340	0.70272
10	3.16926	2.22814	1.81246	1.37218	1.09306	1.04642	0.87906	0.69981
11	3.10582	2.20099	1.79588	1.36343	1.08767	1.04149	0.87553	0.69744
12	3.05454	2.17881	1.78229	1.35622	1.08321	1.03740	0.87261	0.69548
13	3.01228	2.16037	1.77093	1.35017	1.07947	1.03398	0.87015	0.69383
14	2.97685	2.14479	1.76131	1.34503	1.07628	1.03105	0.86805	0.69242
15	2.94673	2.13145	1.75305	1.34061	1.07353	1.02853	0.86624	0.69120
16	2.92079	2.11990	1.74588	1.33676	1.07114	1.02634	0.86467	0.69013
17	2.89823	2.10982	1.73961	1.33338	1.06903	1.02441	0.86328	0.68919
18	2.87844	2.10092	1.73406	1.33039	1.06717	1.02270	0.86205	0.68836
19	2.86094	2.09302	1.72913	1.32773	1.06551	1.02117	0.86095	0.68762
20	2.84534	2.08596	1.72472	1.32534	1.06402	1.01980	0.85996	0.68695
21	2.83137	2.07961	1.72074	1.32319	1.06267	1.01857	0.85907	0.68635
22	2.81876	2.07388	1.71714	1.32124	1.06145	1.01745	0.85827	0.68581
23	2.80734	2.06865	1.71387	1.31946	1.06034	1.01643	0.85753	0.68531
24	2.79695	2.06390	1.71088	1.31784	1.05932	1.01549	0.85686	0.68485
25	2.78744	2.05954	1.70814	1.31635	1.05838	1.01463	0.85624	0.68443
26	2.77872	2.05553	1.70562	1.31497	1.05752	1.01384	0.85567	0.68404
27	2.77068	2.05183	1.70329	1.31370	1.05673	1.01311	0.85514	0.68369
28	2.76326	2.04841	1.70113	1.31253	1.05599	1.01243	0.85465	0.68335
29	2.75639	2.04523	1.69913	1.31143	1.05530	1.01180	0.85419	0.68304
30	2.74998	2.04227	1.69726	1.31042	1.05466	1.01122	0.85377	0.68276
31	2.74404	2.03951	1.69552	1.30946	1.05406	1.01067	0.85337	0.68249
32	2.73849	2.03693	1.69389	1.30857	1.05350	1.01015	0.85300	0.68223
33	2.73329	2.03452	1.69236	1.30774	1.05298	1.00967	0.85265	0.68200
34	2.72839	2.03224	1.69092	1.30695	1.05249	1.00922	0.85232	0.68177
35	2.72381	2.03011	1.68957	1.30621	1.05202	1.00879	0.85201	0.68156
36	2.71948	2.02809	1.68830	1.30551	1.05158	1.00838	0.85172	0.68137
37	2.71541	2.02619	1.68709	1.30485	1.05116	1.00800	0.85144	0.68118
38	2.71157	2.02439	1.68595	1.30423	1.05077	1.00764	0.85118	0.68100
39	2.70791	2.02269	1.68488	1.30364	1.05040	1.00730	0.85093	0.68083
40	2.70446	2.02107	1.68385	1.30308	1.05005	1.00697	0.85070	0.68067
∞	2.57583	1.95996	1.64485	1.28155	1.03643	0.99446	0.84162	0.67449

Table 10.4: Student-T probabilities for various values of the number of degrees of freedom ν. The body of the table contains values of z, such that the probability P that the interval $y \pm zs_y$ will include the mean μ_y is given at the top of the column.