Statistical Analysis of Data in the Linear Regime

Robert DeSerio

University of Florida — Department of Physics

Preface

This book was written for students in an upper-division physics laboratory course. It covers propagation of error, regression analysis, and related topics.

The linearity in the book title is between the random and systematic errors that can be expected in the input data and the resulting variations in the output parameters derived from that data. Covariance matrices describe the input and output variations while first-order Taylor expansions and Jacobian matrices describe the linear relationships involved. Linearity is guaranteed when first-order Taylor expansions provide accurate input-output relationships over the range of typical input errors. This is always the case when the relationships are linear and generally the case when input errors are small. Moreover, even when the input errors begin contributing nonlinearly, an analysis based on linearity typically provides a good first approximation.

Microsoft Excel is an excellent platform for demonstrating the material. The final chapter shows how to use it for the procedures presented here and several spreadsheet examples can be found on the lab website. Only Excel's *Regression* and *Solver* programs and its standard matrix and array functions are used. No other add-ins are needed. A few exercises are scattered throughout the book to fill in various steps in the logic or to provide practice with the equations and procedures.

Two noteworthy results are presented in the chapter on regression analysis. One shows how a least squares solution is equivalent to a maximum likelihood solution not only for Gaussian-distributed data, but for Poisson- and binomial-distributed data as well. The other shows how the uncertainty in an instrument calibration propagates to the uncertainty in any results obtained using that instrument. Hopefully, the reader will find these and other topics treated rigorously and clearly, but also practically, with formulas and procedures applicable to many everyday data analysis tasks.

Contents

1 Introduction **9**
Linear Algebra and Taylor Expansions 12
Small Error Approximation 15

2 Random Variables **17**
Law of Large Numbers 18
Sample Averages and Expectation Values 18
Normalization, Mean and Variance 20

3 Probability Distributions **25**
The Gaussian Distribution 25
The Binomial Distribution 25
The Poisson Distribution 27
The Uniform Distribution 29

4 Statistical Dependence **31**
Correlation 35
The Covariance Matrix 38

5 Measurement Model **41**
Random Errors 41
Systematic Errors 42
Correlated Data 43

6 Propagation of Error **45**
Complete Solutions 46
Propagation of Error 49
Correction to the Mean 54

7 Regression Analysis **57**
Principle of Maximum Likelihood 58
Least-Squares Principle 61
Iteratively Reweighted Least Squares 62
Sample Mean and Variance 63
Weighted Mean 66
Linear Regression 67
Equally-Weighted Linear Regression 72
Nonlinear Regression 73
The Gauss-Newton Algorithm 76
Uncertainties in Independent Variables 79
Regression with Correlated y_i 81
Calibrations and Instrument Constants 82
Transforming the Dependent Variables 86

8 Central Limit Theorem **89**
Single Variable Central Limit Theorem 90
Multivariable Central Limit Theorem 91

9 Evaluating a Fit **95**
Graphical Evaluation 95
The χ^2 Distribution 97
The χ^2 Test 101
Uncertainty in the Uncertainty 102
The Reduced χ^2 Distribution 104
Confidence Intervals 106
Student-T Probabilities 108
The $\Delta\chi^2 = 1$ Rule 109

10 Regression with Excel **113**
Excel's Linear Regression Program 116
General Regression with Excel 117
Parameter Variances and Covariances 121
Multicollinearity and Other Problems 122

Probability tables **127**
Gaussian 127
Reduced Chi-Square 128

Student-T 129

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 might then be histogrammed as in Fig. 1.1. Each histogram bin represents a possible value or range of values as indicated by its placement along the horizontal axis. The height of each bar gives the *frequency,* or number of times a measurement falls in that bin.

The measurements are referred to as a *sample*, the number of measurements is the *sample size* and the histogram is the *sample frequency distri-*

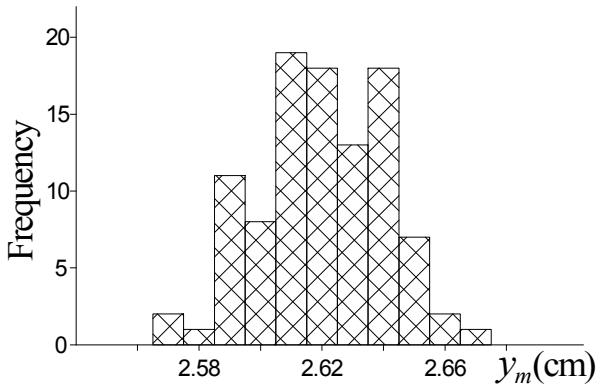


Figure 1.1: A sample frequency distribution for 100 measurements of the length of a rod.

bution. Dividing the frequencies by the sample size yields the fraction of measurements that fall in each bin. A histogram of this kind is called a *sample probability distribution* because it provides an estimate for the probability to fall in each bin. Were new sample sets taken, the randomness of the measurement process would cause each new sample distribution to vary. However, with an ever increasing sample size, the *law of large numbers* states that the sample probability distribution converges to the *parent probability distribution* — a complete statistical description of that particular measurement.

Thus, a single measurement is simply one sample from a parent distribution. It is typically interpreted as the sum of a fixed "signal" component and a random "noise" component. The signal is taken as the mean of the parent distribution and the noise results from stochastic processes that cause individual measurements to deviate randomly from that mean.

The experimenter gives physical meaning to the signal through an understanding of the measuring instrument and its application to a particular apparatus. For example, a thermometer's signal component might be interpreted to be the temperature of the system to which it is attached. Obviously, the interpretation is subject to possible deviations that are distinct from and in addition to the deviations associated with the measurement noise. For example, the thermometer may be out of calibration or it may not be in good thermal contact with the system. Such problems give rise to a deviation between the measurement mean and the true value of the physical quantity being measured.

Measurement error refers to the difference between a measurement and the true value and thus consists of two components. The deviation between an individual measurement and the mean of its distribution is called the *random error.* The deviation between the distribution mean and the true value is called the *systematic error. Measurement uncertainty* refers to the experimenter's inability to provide specific values for either error in any particular measurement. However, estimates of reasonably likely deviations will be needed for data analysis. Indeed, the term measurement uncertainty often refers to such quantitative estimates.

Theoretical models provide relationships among physical variables. For example, the temperature, pressure, and volume of a quantity of gas might be measured to test various equations of state such as the ideal gas law or the Van der Waals model, which predict specific relationships among those variables.

Broadly summarized, statistical analysis often amounts to a compatibility test between the measurements and the theory as specified by the following two hypotheses:

Experimental: The measurement uncertainties are well characterized.

Theoretical: The underlying true 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 they are not compatible, at least one of the two hypotheses must be rejected. The experimental hypothesis is usually first on the chopping block because compatibility depends on how the random and systematic errors are modeled and quantified. Only after careful assessment of both sources of error can one conclude that predictions are the problem. However, 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.

The goal of this book is to present the statistical models, formulas, and procedures needed to accomplish the compatibility test for a range of experimental situations commonly encountered in the physical sciences.

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 sample averages and expectation values.

Four of the most common measurement probability distributions are described in Chapter 3. Chapter 4 introduces the joint probability distribution for multiple random variables and the related topics of statistical independence, correlation, and the covariance matrix. Chapter 5 discusses systematic errors and other measurement issues.

Chapter 6 provides propagation of error formulas for determining the uncertainty in variables defined from other variables. Chapter 7 discusses regression analysis based on the principle of maximum likelihood. Chapter 7 also treats special situations such as uncertainty in the independent variable and systematic errors that can be modeled with instrument calibrations.

Chapter 8 demonstrates the central limit theorem with one- and two-dimensional examples. Chapter 9 discusses the evaluation of regression results and the chi-square random variable and Chapter 10 provides a guide to using Excel for regression analysis.

Linear Algebra and Taylor Expansions

Linear algebra is an indispensable tool for data analysis. It condenses equation sets into single vector equations and replaces summation symbols with the implied sums of linear algebra multiplication rules. The notation used is as follows.

Column and row vectors will be displayed in bold face type. For example, the main input data to an analysis will typically be represented by the set y_i, $i = 1...N$. It will be referred to as the data set $\{y\}$ (or $\{y_i\}$ if the index name is to be identified), by the expression "the y_i" or by the column vector

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{1.1}$$

The transpose of a column vector — signified with a superscript T — is a row vector with the same elements in the same order. The transpose of $\boldsymbol{y}$ is

$$\boldsymbol{y}^T = (y_1, y_2, ..., y_N) \tag{1.2}$$

The product $\boldsymbol{y}^T\boldsymbol{y}$ is an inner product — a scalar given by

$$\begin{aligned} \boldsymbol{y}^T\boldsymbol{y} &= (y_1, y_2, ..., y_N) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \\ &= \sum_{i=1}^{N} y_i^2 \end{aligned} \tag{1.3}$$

The product $\boldsymbol{y}\boldsymbol{y}^T$ is an outer product — the $N \times N$ matrix given by

$$\begin{aligned} \boldsymbol{y}\boldsymbol{y}^T &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} (y_1, y_2, ..., y_N) \\ &= \begin{pmatrix} y_1^2 & y_1y_2 & \cdots & y_1y_N \\ y_2y_1 & y_2^2 & \cdots & y_2y_N \\ \vdots & \vdots & \ddots & \vdots \\ y_Ny_1 & y_Ny_2 & \cdots & y_N^2 \end{pmatrix} \end{aligned} \tag{1.4}$$

The output or results of an analysis will typically be represented by the set $\{a\}$ of size M, i.e., a_k, $k = 1...M$, or by the column vector $\boldsymbol{a}$. The N input y_i are typically measurements and each comes with some uncertainty. Because of this, there will be some uncertainty in the M values of a_k derived from them. The analysis must not only provide the a_k from the y_i, but also the uncertainty in the a_k arising from the uncertainty in the y_i.

The relationship between the input and output uncertainties is largely determined by the $M \times N$ (M rows by N columns) *Jacobian* matrix $[J_y^a]$ giving the partial derivatives of each a_k with respect to each y_i.

$$[J_y^a]_{ki} = \frac{\partial a_k}{\partial y_i} \tag{1.5}$$

Matrices such as $[J_y^a]$ will be represented by a descriptive name in regular math fonts surrounded by square brackets. Double subscripts outside the square brackets (k and i above) label the row and column, respectively, of the identified element. For Jacobians, a subscript and superscript inside the square brackets identify the variable sets involved and are a reminder of the units for the elements; $[J_y^a]_{ki}$ has the units of a_k/y_i.

Variances, covariances, and the covariance matrix are discussed more fully in Chapter 4. Briefly stated, the covariance matrix for a data set describes the fluctuations that can be expected in the set elements if the procedure to obtain the data set were repeated over and over. Diagonal elements provide information (variances) about the size of the fluctuations that can be expected for each variable. Off-diagonal elements provide information (covariances) describing correlations that can be expected between the fluctuations of any two variables.

Covariance matrices for the input y_i and output a_k will be written $[\sigma_y^2]$ and $[\sigma_a^2]$, respectively, with the subscripts y and a providing the data sets involved. The superscript 2 in a covariance matrix such as $[\sigma_y^2]$ is a reminder that the ith diagonal element has the units of y_i^2 and off-diagonal elements have the units of $y_i y_j$.

The transpose of a matrix, signified by a superscript T outside the square brackets, has the same matrix elements but interchanges the rows of the original matrix for the columns of its transpose and vice versa. Thus, the transpose of $[J_y^a]$ is the $N \times M$ (N rows by M columns) matrix $[J_y^a]^T$ with

elements given by

$$\begin{aligned} [J_y^a]_{ik}^T &= [J_y^a]_{ki} \\ &= \frac{\partial a_k}{\partial y_i} \end{aligned} \tag{1.6}$$

Matrix inversion will be signified by a superscript -1 outside the square brackets and is only valid for certain square matrices. The inverse $[X]^{-1}$ of an $N \times N$ invertible matrix $[X]$ satisfies

$$[X][X]^{-1} = [X]^{-1}[X] = [1] \tag{1.7}$$

where $[1]$ is the $N \times N$ *unit matrix* with ones along the diagonal and zeros elsewhere. It is also called the *identity matrix* because it leaves any appropriately sized vector or matrix unchanged under multiplication from the left or right.

When two matrices, two vectors, or a vector and a matrix are multiplied, their sizes must be compatible and their ordering matters. The adjacent indices in a multiplication will be summed over in forming the result and must be of the same size. For example, $[J_y^a][J_y^a]^T$ is an $M \times M$ matrix with elements given by

$$[[J_y^a][J_y^a]^T]_{kj} = \sum_{i=1}^{N} [J_y^a]_{ki} [J_y^a]_{ij}^T \tag{1.8}$$

The elements of the Jacobian are defined as for any partial derivative. After performing an analysis, thereby finding the a_k from the y_i, change one y_i to y_i' — that is, change it by the small amount $\Delta y_i = y_i' - y_i$. Redo the analysis. Each a_k will change by Δa_k (to $a_k' = a_k + \Delta a_k$). Make the change Δy_i smaller and smaller until the Δa_k are proportional to Δy_i. For Δy_i in this linear regime, the elements of the Jacobian are given by: $[J_y^a]_{ki} = \Delta a_k / \Delta y_i$. Of course, elements of the Jacobian can also be obtained by analytic differentiation (Eq. 1.5) when the functional form: $a_k = f_k(\{y_i\})$ is known.

With the Jacobian in hand, if all the y_i are then simultaneously allowed to change — within the linear regime — to a new set $\{y_i'\}$, a first-order Taylor expansion gives the new a_k' for this set

$$a_k' = a_k + \sum_{i=1}^{N} \frac{\partial a_k}{\partial y_i} (y_i' - y_i) \tag{1.9}$$

or

$$\Delta a_k = \sum_{i=1}^{N} [J_y^a]_{ki} \Delta y_i \tag{1.10}$$

which is just the kth row of the vector equation

$$\Delta \boldsymbol{a} = [J_y^a] \Delta \boldsymbol{y} \tag{1.11}$$

The Jacobian evaluated at one set of y_i describes how all the a_k will change when any or all of the y_i change by small amounts.

To similarly express the row vector, $\Delta \boldsymbol{a}^T$, recall the rules for the transpose of a matrix-matrix or vector-matrix multiplication: The transpose of a product of terms is the product of the transpose of each term with the terms' ordering reversed: $[[A][B]]^T = [B]^T[A]^T$. Thus, the equivalent transpose to Eq. 1.11 is

$$\Delta \boldsymbol{a}^T = \Delta \boldsymbol{y}^T [J_y^a]^T \tag{1.12}$$

Small Error Approximation

The first-order Taylor expansion expressed by Eqs. 1.9-1.12 describes how changes to the input variables will propagate to changes in the output variables. It is the basis for propagation of error and regression analysis — topics covered in Chapters 6 and 7. However, one issue is worthy of a brief discussion here.

In order for the first-order Taylor expansion to be valid, the effects of higher-order derivatives must be kept small. This happens for any size Δy_i when higher-order derivatives are absent, i.e., when the relationships between the a_k and the y_i are linear. When the relationships are nonlinear, it requires keeping the range of possible Δy_i small enough that throughout that range, the higher-order terms in the Taylor expansions would be small compared to the first-order terms. The range of possible Δy_i is determined by the measurement uncertainty and will be assumed small enough to satisfy this requirement.

Chapter 9 presents calculations that can be used to check for nonlinear effects. However, analysis of data having large uncertainties that could take Δy_i outside the linear regime will only be treated for a few special cases. If

the uncertainties are too big to trust treatments based on first-order Taylor expansions, reducing them into the linear regime serves two purposes. It not only makes the uncertainty calculations more trustworthy, it also lowers the uncertainties in the final results.

Chapter 2

Random Variables

The experimental model treats each input or measured y_i as a *random sample* — a quantity whose value varies randomly as the procedure used to obtain it is repeated. Possible values occur randomly but with fixed probabilities as described next.

When the possible y_i form a discrete set, the quantity $P(y_i)$ gives the probability for one particular y_i to occur. The complete set of probabilities for all y_i is called a *discrete probability function* (or dpf).

When the possible values cover a continuous interval, their probabilities are described by a *probability density function* (or pdf). With the pdf $p(y)$ specified for all possible values of y, 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 range are obtained by integration. The probability of an outcome between y_1 and y_2 is given by

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

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 $\pm 1/2$ of the size, Δy, of the least significant digit.

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

Note how the values of $P(y)$ for a complete set of nonoverlapping 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

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

Both discrete probability functions and probability density functions are referred to as probability distributions. The $P(y_i)$, being probabilities, must be between zero and one and are unitless. And because $p(y)\Delta y$ is a probability, $p(y)$ must be a "probability per unit y" and thus it must be nonnegative with units inverse to those of y.

Before discussing important properties of a distribution such as its mean and standard deviation, the related subject of sampling is addressed more generally.

Law of Large Numbers

$P(y)$ for an unknown distribution can be determined by acquiring and histogramming a sample of sufficient size.

For a discrete probability distribution, the histogram bins should be labeled by the allowed values y_i. For a continuous probability distribution, the bins should be labeled by their midpoints y_i 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 $f(y_i)$ for each bin.

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

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

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 finite N, the sample average of any (nonconstant) function $g(y)$ is a random variable; taking a new sample set of y_i would likely produce a different sample average. However, in the limit of infinite sample size, the law of large numbers implies that the sample average converges to a well defined constant depending only on the parent probability distribution and the particular function $g(y)$. This constant is called the *expectation value* of $g(y)$ and will be denoted by putting angle brackets around the function.

$$\langle g(y)\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} g(y_i) \tag{2.7}$$

To obtain analytical expressions for expectation values that do not require an infinite sample, Eq. 2.7 can be cast into a form suitable for use with a given probability distribution as follows. Consider a large sample of size N that 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 the bin widths Δy are 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. 2.7.

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. 2.7 can be grouped by bins and becomes

$$\langle g(y)\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{\text{all } y_j} g(y_j) f(y_j) \tag{2.8}$$

Note the change from a sum over all samples in Eq. 2.7 to a sum over all histogram bins in Eq. 2.8.

Moving the limit and factor of $1/N$ inside the sum, the law of large numbers (Eq. 2.5) can be used giving

$$\langle g(y)\rangle = \sum_{\text{all } y_j} g(y_j)\, P(y_j) \tag{2.9}$$

Note that any reference to a sample is gone. Only the range of possible y-values, the probability distribution, and the arbitrary function $g(y)$ are involved. Equation 2.9 is called a weighted average of $g(y)$; each value of $g(y_j)$ in the sum is weighted by the probability of its occurrence $P(y_j)$.

For a continuous probability density function, substitute $P(y_j) = p(y_j)\Delta y$ in Eq. 2.9 and take the limit as $\Delta y \to 0$. This converts the sum to the integral

$$\langle g(y) \rangle = \int_{-\infty}^{\infty} g(y)\, p(y)\, dy \tag{2.10}$$

Eq. 2.10 is a weighted integral with each $g(y)$ weighted by its occurrence probability $p(y)\, dy$.

Some frequently used properties of expectation values are given below. Justifications are given based on simple substitutions for $g(y)$ in Eqs. 2.9 or 2.10 or based on the operational definition of an expectation value as an average for an effectively infinite data set (Eq. 2.7).

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

2. Constants can be factored out of expectation value brackets: $\langle c\, u(y) \rangle = c\, \langle u(y) \rangle$. Substitute $g(y) = c\, u(y)$, where c is a constant. Guaranteed by the distributive property of multiplication over addition for the terms involved in the average.

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 average.

But also keep in mind that 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 $\langle u(y)v(y) \rangle = \langle u(y) \rangle\, \langle v(y) \rangle$.

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 becomes

$$\sum_{\text{all } y_j} P(y_j) = 1 \tag{2.11}$$

and for a continuous probability distribution it becomes

$$\int_{-\infty}^{\infty} p(y)\, dy = 1 \tag{2.12}$$

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. It is the first moment of the distribution.

$$\mu_y = \langle y \rangle \tag{2.13}$$

If $P(y)$ or $p(y)$ is specified, μ_y could be evaluated using $g(y) = y$ in Eq. 2.9 or 2.10, respectively.

The mean is a measure of the central value of the distribution. It is a point at the "center of probability" in analogy to a center of mass. Were mass distributed along the y-axis in proportion to $P(y)$ (point masses) or in proportion to $p(y)$ (a mass distribution), μ_y would be the center of mass. The *median* is also a quantitative and common measure of the center of a distribution. It is that value of y where there is equal probability above as below. The mean is the only measure that will be considered further.

The quantity μ_y is sometimes called the true mean to distinguish it from the sample mean. The *sample mean* of a set of y_i is simply the sample average of y — defined by Eq. 2.6 with $g(y) = y$.

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{2.14}$$

The sample mean is often used as an estimate of the true mean because, by definition, it becomes exact as $N \to \infty$. In addition, the sample mean

satisfies another important property for any good estimate. Taking the expectation value of both sides of Eq. 2.14 and noting that $\langle y_i \rangle = \mu_y$ for all N samples (Eq. 2.13) gives

$$\begin{aligned} \langle \bar{y} \rangle &= \left\langle \frac{1}{N} \sum_N y_i \right\rangle \\ &= \frac{1}{N} \sum_{i=1}^{N} \langle y_i \rangle \\ &= \frac{1}{N} \sum_{i=1}^{N} \mu_y \\ &= \frac{1}{N} N \mu_y \\ &= \mu_y \end{aligned} \tag{2.15}$$

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.

After the mean, the next most important descriptor of a probability distribution is its standard deviation — a measure of the how far away from the mean individual sample values are likely to be. The quantity

$$\delta y = y - \mu_y \tag{2.16}$$

is called the *deviation* — the signed difference between a sample value and the mean of its parent distribution. One of its properties, true for any distribution, can be obtained by rewriting Eq. 2.13 in the form

$$\langle y - \mu_y \rangle = 0 \tag{2.17}$$

Deviations are signed quantities and for any distribution, by definition, the mean deviation is always zero.

The *mean absolute deviation* is the expectation value of the absolute value of the deviation: $\langle |y - \mu_y| \rangle$. This quantity would be nonzero and a reasonable measure of the expected magnitude of typical deviations. However, the mean absolute deviation does not arise naturally when formulating the basic

statistical procedures considered here. The *mean squared deviation*, on the other hand, plays a central role and so 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 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

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

For a given probability distribution, the variance could then be evaluated with $g(y) = (y - \mu_y)^2$ in Eq. 2.9 or 2.10.

The standard deviation σ_y is the square root of the variance — the square root of the mean-squared deviation. Thus, it is often referred to as the rms or root-mean-square deviation — particularly when trying to emphasize its role as the typical size of a deviation.

The variance has units of y^2 while the standard deviation has the same units as y. The standard deviation is the most common measure of the width of a distribution and the only one that will be considered further.

Expanding the right side of Eq. 2.18 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

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

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

$$\left\langle y^2 \right\rangle = \mu_y^2 + \sigma_y^2 \tag{2.20}$$

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

$$s_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2 \tag{2.21}$$

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

$$\left\langle s_y^2 \right\rangle = \sigma_y^2 \tag{2.22}$$

The proof is similar to that of Eq. 2.15, this time requiring an application of Eq. 2.18 to each term in the sum.

Typically, the true mean μ_y is not known and Eq. 2.21 can not be used to determine s_y^2. Can the sample mean $\bar{y}$ be used in place of μ_y? Yes, but making this substitution requires the following modification to Eq. 2.21.

$$s_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{2.23}$$

As will be proven later, the denominator is reduced by one so that this definition of the sample variance will also be unbiased, i.e., will still satisfy Eq. 2.22.

The sample mean and the sample variance are random variables and each follows its own probability distribution. They are unbiased; the means of their distributions will be the true mean and true variance, respectively. The standard deviations of these distributions will be discussed in Chapters 7 and 9.

Chapter 3

Probability Distributions

In this section, definitions and properties of a few fundamental probability distributions are presented.

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. Gaussian probabilities are described relative to the mean and standard deviation. There is a 68% probability that a sample from a Gaussian distribution 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 results when an experiment, called a *Bernoulli trial*, is repeated a fixed number of times. A Bernoulli trial can have only

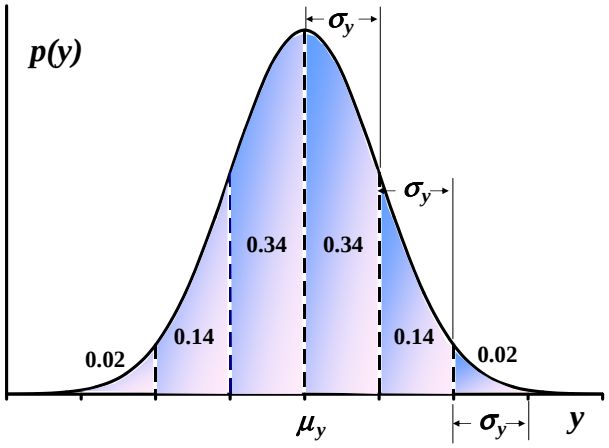


Figure 3.1: The Gaussian distribution labeled with the mean μ_y, the standard deviation σ_y and some areas, i.e., probabilities.

two outcomes. One outcome is termed a success and occurs with a probability p. The other, termed a failure, occurs with a probability $1-p$. Then, with $\mathcal{N}$ Bernoulli trials, the number of successes y can be any integer from zero (none of the $\mathcal{N}$ trials were a success) to $\mathcal{N}$ (all trials were successes).

The probability of y successes (and thus $\mathcal{N}-y$ failures) is given by

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

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

The binomial distribution has a mean

$$\mu_y = \mathcal{N}p \tag{3.3}$$

and a variance

$$\sigma_y^2 = \mathcal{N}p(1-p) \tag{3.4}$$

It will prove useful to rewrite the distribution and the variance in terms of

$\mathcal{N}$ and μ_y rather than $\mathcal{N}$ and p. Substituting $\mu_y/\mathcal{N}$ for p, the results become

$$P(y) = \frac{\mathcal{N}!}{y!(\mathcal{N}-y)!}\frac{1}{\mathcal{N}^{\mathcal{N}}}\ (\mu_y)^y(\mathcal{N}-\mu_y)^{\mathcal{N}-y} \tag{3.5}$$

$$\sigma_y^2 = \mu_y\left(1-\frac{\mu_y}{\mathcal{N}}\right) \tag{3.6}$$

The binomial distribution arises, for example, when histogramming sample frequency distributions. Consider $\mathcal{N}$ samples from a given probability distribution for a random variable x. A particular bin at x_j represents a particular outcome or range of outcomes and the parent distribution would determine the associated probability $P(x_j)$ for a result in that bin. While any distribution might give rise to the $P(x_j)$, the frequency in that particular histogram bin would be governed by the binomial distribution. Each Bernoulli trial consists of taking one new sample and, according to its value, either sorting it into that bin — a success with a probability $P(x_j)$, or not sorting it in that bin — a failure with a probability $1-P(x_j)$. After $\mathcal{N}$ samples, the number of successes (the bin frequency y) is a binomial random variable with that $\mathcal{N}$ and $\mu_y = \mathcal{N}P(x_j)$.

The Poisson Distribution

Poisson-distributed variables arise, for example, in particle and photon counting experiments. Under unchanging experimental conditions and averaged over long times, "counts" or "clicks" from a particle or photon detector might be occurring at an average rate of, say, one per second. Over many ten-second intervals, ten counts would be the average, but the actual number in any particular interval will often be higher or lower with probabilities governed by the Poisson distribution.

More specifically, if μ_y is the average number of counts expected in an interval (which need not be integer valued), then the counts y actually measured in any such interval (which can only be zero or a positive integer) will occur randomly with probabilities governed by the Poisson distribution

$$P(y) = e^{-\mu_y}\frac{(\mu_y)^y}{y!} \tag{3.7}$$

Not all counts are Poisson distributed. The *Poisson Variables* addendum on the lab website describes conditions that guarantee a count will be Poisson

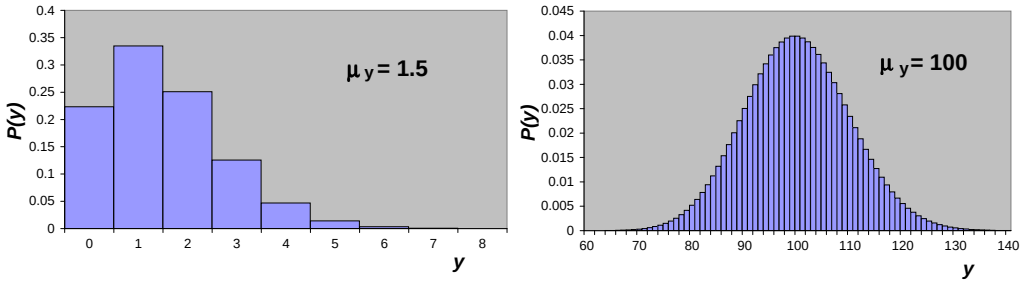


Figure 3.2: Poisson probability distributions for means of 1.5 and 100.

distributed. It also gives a derivation of the Poisson distribution and the related exponential distribution based on the assumption that the probability per unit time for an event to occur is constant. Poisson probability distributions for $\mu_y = 1.5$ and $\mu_y = 100$ are shown in Fig. 3.2.

One can show (see Exercise 1) that the variance of a Poisson distribution is the mean.

$$\sigma_y^2 = \mu_y \tag{3.8}$$

For large values of μ_y, the Poisson probability for a given y is very nearly Gaussian — given by Eq. 2.4 with $\Delta y = 1$ and $p(y)$ given by Eq. 3.1 (with $\sigma_y^2 = \mu_y$). That is,

$$P(y) \approx \frac{1}{\sqrt{2\pi\mu_y}} \exp\left[-\frac{(y-\mu_y)^2}{2\mu_y}\right] \tag{3.9}$$

Eqs. 3.8 and 3.9 are the origin of the commonly accepted practice of applying "square root statistics" or "counting statistics," whereby a Poisson-distributed variable is treated as a Gaussian-distributed variable with the same mean and 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 variance of its own parent distribution. That is, y is assumed to be governed by a Gaussian distribution with a standard deviation given by

$$\sigma_y = \sqrt{y} \tag{3.10}$$

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

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

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 is often used for digital meters. A reading of 3.72 V on a 3-digit voltmeter might imply that 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 (and zero probability outside this range) has a pdf given by

$$p(y) = \frac{1}{|b-a|} \tag{3.11}$$

Exercise 1 *Eqs. 2.13 and 2.18 provide the definitions of the mean μ_y and variance σ_y^2 with Eqs. 2.9 or 2.10 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 or the Γ function. 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_y 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_y and σ_y^2 appearing in the distribution are, in fact, the mean and variance.

Hints: For the binomial distribution you may need the expansion

$$(a+b)^{\mathcal{N}} = \sum_{y=0}^{\mathcal{N}} \frac{\mathcal{N}!}{y!(\mathcal{N}-y)!} a^y b^{\mathcal{N}-y} \tag{3.12}$$

For the Poisson distribution you may need the power series expansion

$$e^a = \sum_{y=0}^{\infty} \frac{a^y}{y!} \tag{3.13}$$

For the Gaussian distribution be sure to start by eliminating the mean (with the substitution $y' = y - \mu$). 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 from $dx'\,dy'$ to $r\,dr\,d\theta$ and set the limits of integration for r from 0 to ∞ and for θ from 0 to 2π.

Exercise 2 *(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 P 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 that probability with your sample size to predict the mean and standard deviation for that bin's frequency. Compare your actual bin frequency at μ_y with this prediction. Is the difference between them reasonable? Hint: the bin frequency follows a binomial distribution, which has a mean of NP and a standard deviation equal to $\sqrt{NP(1-P)}$.

Chapter 4

Statistical Dependence

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. Statistically independent variables show no relationships among their natural random deviations. Statistically dependent variables can show correlated deviations.

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 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, 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 $\Pr(A)$, 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

$$\Pr(A|B) = \Pr(A) \tag{4.1}$$

Whether events are independent or not, the *joint probability* of "A and B" both occurring — 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.

$$\Pr(A \cap B) \quad = \quad \Pr(B) \Pr(A|B) \tag{4.2}$$

Then, substituting Eq. 4.1 gives the *product rule* valid for independent events.

$$\Pr(A \cap B) = \Pr(A) \Pr(B) \tag{4.3}$$

Equation 4.3 states another common definition of independence — that the probability for two independent events to occur is simply the product of the probabilities for each to occur.

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

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

$$P(x, y) = P_x(x) P_y(y) \tag{4.4}$$

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 from y to $y + dy$ is given by the product of each variable's probability

$$dP(x, y) = p_x(x) p_y(y) dx\, dy \tag{4.5}$$

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

$$\langle f_1(y_1) f_2(y_2) \rangle = \langle f_1(y_1) \rangle \langle f_2(y_2) \rangle \tag{4.6}$$

if y_1 and y_2 are independent.

For discrete random variables, the proof starts from the definition of the expectation value for a function of two discrete random variables followed by an application of Eq. 4.4 as follows:

$$\begin{aligned}
\langle f_1(y_1) f_2(y_2) \rangle &= \sum_{\text{all } y_1, y_2} f_1(y_1) f_2(y_2)\, P(y_1, y_2) \\
&= \sum_{\text{all } y_1} \sum_{\text{all } y_2} f_1(y_1) f_2(y_2)\, P_1(y_1)\, P_2(y_2) \\
&= \sum_{\text{all } y_1} f_1(y_1) P_1(y_1) \sum_{\text{all } y_2} f_2(y_2) P_2(y_2) \\
&= \langle f_1(y_1) \rangle \langle f_2(y_2) \rangle
\end{aligned} \tag{4.7}$$

Similarly for continuous random variables, the proof starts from the definition of the expectation value for a function of two continuous random variables followed by an application of Eq. 4.5 as follows:

$$\begin{aligned}
\langle f_1(y_1) f_2(y_2) \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(y_1) f_2(y_2)\, dP(y_1, y_2) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(y_1) f_2(y_2) p_1(y_1) p_2(y_2)\, dy_1\, dy_2 \\
&= \int_{-\infty}^{\infty} f_1(y_1) p_1(y_1)\, dy_1 \int_{-\infty}^{\infty} f_2(y_2) p_2(y_2)\, dy_2 \\
&= \langle f_1(y_1) \rangle \langle f_2(y_2) \rangle
\end{aligned} \tag{4.8}$$

A simple example of Eq. 4.6 is for the expectation value of the product of two independent random variables, y_1 and y_2; $\langle y_1 y_2 \rangle = \langle y_1 \rangle \langle y_2 \rangle = \mu_1 \mu_2$. For the special case where the independent samples y_i and y_j come 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. 2.20 ($\langle y_i^2 \rangle = \mu_y^2 + \sigma_y^2$) for the expectation value of the square of any y-value gives the following relationship for independent samples from the same distribution

$$\langle y_i y_j \rangle = \mu_y^2 + \sigma_y^2 \delta_{ij} \tag{4.9}$$

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. 4.6 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 samples from the same or from different distributions.

$$\langle (y_1 - \mu_1)(y_2 - \mu_2) \rangle = \langle y_1 - \mu_1 \rangle \langle y_2 - \mu_2 \rangle \tag{4.10}$$

Here μ_1 and μ_2 are the means of the distributions for y_1 and y_2 and satisfy $\langle y_i - \mu_i \rangle = 0$. Thus, the right-hand side of Eq. 4.10 is the product of two zeros and demonstrates that

$$\langle (y_1 - \mu_1)(y_2 - \mu_2) \rangle = 0 \tag{4.11}$$

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. Nonzero values for this quantity are possible if y_1 and y_2 are not independent. This issue will be addressed shortly.

The product rule (Eqs. 4.4 and 4.5) can be extended — by repeated multiplication — to any number of independent random variables. The explicit form for the joint probability for an entire data set y_i, $i = 1...N$ will be useful for our later treatment of regression analysis. This form depends on the particular probability distributions for the y_i. Often, all y_i come from the same kind of distribution: either Gaussian, Poisson or binomial. These kinds of data sets lead to the joint probability distributions considered next.

For N independent Gaussian random variables, with the distribution for 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. 2.4 with $p(y_i)$ having the Gaussian form of Eq. 3.1.

$$P(\{y\}) = \prod_{i=1}^{N} \frac{\Delta y_i}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma_i^2} \right] \tag{4.12}$$

where Δy_i represents the size of the least significant digit in y_i, which are all assumed to be small compared to the σ_i.

For N independent Poisson random variables, with the distribution for each y_i having its own mean μ_i, the joint probability distribution becomes the following product of terms — each having the Poisson form of Eq. 3.7.

$$P(\{y\}) = \prod_{i=1}^{N} \frac{e^{-\mu_i} (\mu_i)^{y_i}}{y_i!} \tag{4.13}$$

For N independent binomial random variables, with the distribution for each y_i having its own mean μ_i and number of trials $\mathcal{N}_i$, the joint probability distribution becomes the following product of terms — each having the binomial form of Eq. 3.5.

$$P(\{y\}) = \prod_{i=1}^{N} \frac{\mathcal{N}_i!}{(\mathcal{N}_i - y_i)!y_i!} \frac{1}{\mathcal{N}_i^{\mathcal{N}_i}} (\mu_i)^{y_i} (\mathcal{N}_i - \mu_i)^{\mathcal{N}_i - y_i} \tag{4.14}$$

The joint probability distributions of Eqs. 4.12-4.14 are the basis for regression analysis and, as shown in Chapter 7, produce amazingly similar expressions when applied to that problem.

Correlation

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

The generic data set now under consideration consists of pairs of 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 a set of pairs, x_i, y_i, $i = 1...N$, taken under unchanging conditions so that only random, but perhaps not independent, variations are expected.

Considered separately, each variable varies randomly according to an underlying probability distribution. Treated as two separate sample sets: x_i, $i = 1...N$ and y_i, $i = 1...N$, two different 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 estimates for the true means μ_x and μ_y and true 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 in the xy-plane. Figure 4.1 shows scatter plots for five different 1000-point samples of pairs of random variables. For all five sample sets, the unconditional parent pdfs, $p_x(x)$ and $p_y(y)$, are exactly the same, namely Gaussian distributions having

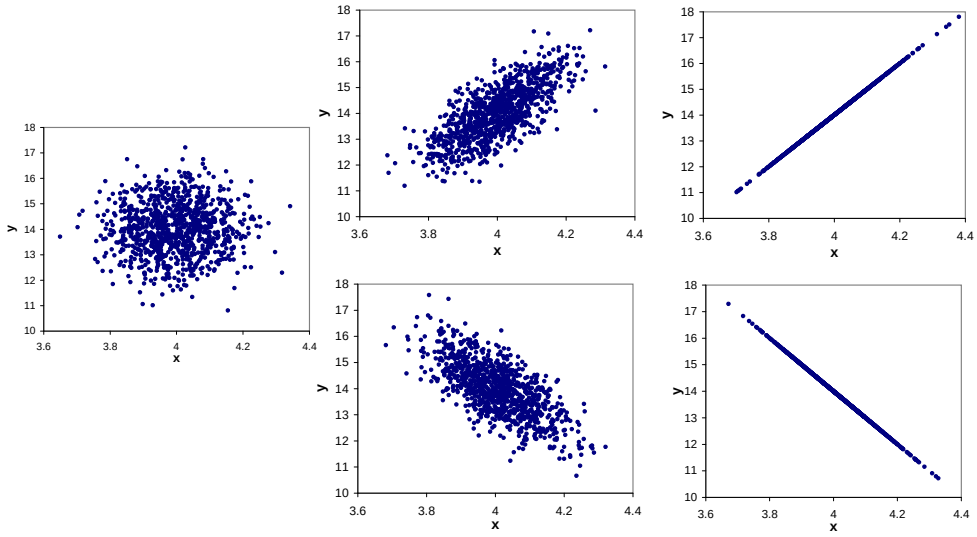


Figure 4.1: The behavior of uncorrelated and correlated Gaussian random variables. The leftmost figure shows uncorrelated variables, the middle two show partial correlation and the two on the right show total correlation. The upper two show positive correlations while the lower two show negative correlations.

the following parameters: $\mu_x = 4$, $\sigma_x = 0.1$ and $\mu_y = 14$, $\sigma_y = 1$. Even though the unconditional pdfs are all the same, the scatter plots clearly show that the joint probability distributions are different. The set on the left is uncorrelated and the other four are correlated.

For the uncorrelated case on the left, the probability for a given y is independent of the value of x. For example, if only those points within some narrow slice in x, say around $x = 4.1$, are analyzed — thereby making them conditional on that value of x, the values of y for that slice have the same probabilities as for the unconditional case — for example, there is still an equal probability for a y-value above the mean of 14 as below it.

The other four cases show correlation. Selecting different slices in one variable will give different conditional probabilities for the other variable. In particular, the conditional mean for one variable 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 conditional mean of one variable increases for slices at increasing values for the other

variable. When one variable is above (or below) its mean, the other is more likely to be above (or below) its mean. The product $(x - \mu_x)(y - \mu_y)$ is positive more often than it is negative and its expectation value is positive. For negative correlation, these dependencies reverse — the variables are more likely to be on opposite sides of their means and the expectation value of $(x - \mu_x)(y - \mu_y)$ is negative. For independent variables, $(x - \mu_x)(y - \mu_y)$ has an expectation value of zero.

One measure of correlation is just this expectation value. The *covariance* σ_{xy} between two variables x and y is defined as the expectation value

$$\sigma_{xy} = \langle (x - \mu_x)(y - \mu_y) \rangle \tag{4.15}$$

It is limited by the size of σ_x and σ_y. The *Cauchy-Schwarz inequality* states that σ_{xy} can vary from $-\sigma_x\sigma_y$ to $\sigma_x\sigma_y$.

$$-\sigma_x\sigma_y \leq \sigma_{xy} \leq \sigma_x\sigma_y \tag{4.16}$$

Thus, σ_{xy} is often expressed

$$\sigma_{xy} = \rho\, \sigma_x\sigma_y \tag{4.17}$$

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 relationship exactly. The correlation coefficients used to generate Fig. 4.1 were 0, ± 0.7 and ± 1.

The *sample covariance* of a data set is defined by

$$s_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \tag{4.18}$$

and is an unbiased estimate of the true covariance σ_{xy}, converging to it in the limit of infinite sample size. The inequality expressed by Eq. 4.16 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}, respectively. The sample correlation coefficient r is then defined by $s_{xy} = r s_x s_y$ and also varies between -1 and 1.

It is informative to see one method for generating two correlated random variables having a given correlation coefficient. Let $R_1(0,1)$ and $R_2(0,1)$ represent two independent random samples from any distributions with a

mean of zero and standard deviation of one. It is not hard to show that random variables x and y generated by

$$\begin{aligned} x &= \mu_x + \sigma_x R_1(0,1) \\ y &= \mu_y + \sigma_y \left(\rho R_1(0,1) + \sqrt{1-\rho^2}\, R_2(0,1) \right) \end{aligned} \tag{4.19}$$

will have means μ_x and μ_y, standard deviations σ_x and σ_y, and correlation coefficient ρ.

Equation set 4.19 shows that while the deviations in x arise from R_1 only, the deviations in y arise from one component proportional to R_1 plus a second independent component proportional to R_2. The required correlation is achieved by setting the relative amplitude of those two components in proportion to ρ and $\sqrt{1-\rho^2}$, respectively. The *Correlated RV.xls* spreadsheet uses these equations to generate correlated random variables with Gaussian or uniform distributions for R_1 and R_2.

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. This distribution is of interest, for example, when testing for evidence of any correlation — even a weak one — between two variables. A sample correlation coefficient 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.

The Covariance Matrix

The *covariance matrix* describes all the variances and covariances possible between two or more random variables. For the set: y_1, y_2, and y_3, the covariance matrix $[\sigma_y^2]$ would be

$$[\sigma_y^2] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \tag{4.20}$$

where the variances and covariances are the elements of $[\sigma_y^2]$:

$$\begin{aligned} [\sigma_y^2]_{ij} &= \sigma_{ij} \\ &= \langle (y_i - \mu_i)\,(y_j - \mu_j) \rangle \end{aligned} \tag{4.21}$$

and μ_i is the mean of y_i. Note how Eq. 4.21 properly defines both the off-diagonal elements as the covariances and the diagonal elements $\sigma_{ii} = \sigma_i^2 = \langle (y_i - \mu_i)^2 \rangle$ as the variances. It also shows that the covariance matrix is symmetric about the diagonal with $[\sigma_y^2]_{ij} = [\sigma_y^2]_{ji}$ and thus is its own transpose. In linear algebra notation, the entire matrix can be written as the expectation value of an outer product:

$$[\sigma_y^2] \;=\; \left\langle (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y}^T - \boldsymbol{\mu}^T) \right\rangle \tag{4.22}$$

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

$$[\sigma_y^2] = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} \tag{4.23}$$

When variables are independent, their joint probability distribution follows the product rule — leading to Eq. 4.12, for example, 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 for the general case. However, the Gaussian joint probability distribution (for N variables) having means μ_i and having variances and covariances satisfying Eq. 4.21 would be expressed

$$P(\{y\}) = \frac{\left(\prod_{i=1}^{N} \Delta y_i\right)}{\sqrt{(2\pi)^N \left|[\sigma_y^2]\right|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right)\right] \tag{4.24}$$

where $\left|[\sigma_y^2]\right|$ is the determinant of $[\sigma_y^2]$ and $\left[\sigma_y^2\right]^{-1}$ is its inverse. Normal vector-matrix multiplication rules apply so that the argument of the exponential is a scalar.

Note that Eq. 4.24 is the general form for a Gaussian joint pdf and reduces to the special case of Eq. 4.12 for independent variables, i.e., for a diagonal covariance matrix.

Chapter 5

Measurement Model

This short chapter presents an idealized measurement model appropriate for the treatments presented. It also briefly addresses analysis of data with systematic errors and correlations.

Random Errors

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

$$y = \mu_y + \delta_y \tag{5.1}$$

For most measurements the true mean is unknown and thus the actual random error (or deviation) $\delta_y = y - \mu_y$ cannot be determined. Whenever possible, however, the experimentalist supplies an estimate of the standard deviation σ_y to set the scale for the size of typical deviations that can be expected. The variance of each y_i in a data set should be consistent with its definition as the mean squared deviation and the covariances between each pair of y_i should be consistent with their definition — Eq. 4.21. As will be discussed in Chapter 8, the values, variances, and covariances of a data set are often the only quantities that will affect the results derived from that data.

One method for estimating variances and covariances for a measurement set is to take a large sample of such sets while all experimental conditions remain constant. The resulting sample variances and covariances might then be calculated and assumed to be the true variances and covariances for any

future measurement sets of the same kind. Often, only rough estimates are made. The experimenter may assume covariances are all zero because the measurements are expected to be statistically independent. Standard deviations may be estimated from an analog meter's smallest division or a digital meter's least digit. Of course, estimates are not exact, but for now, all variances and covariances entering into an analysis will be assumed known. Issues associated with uncertainty in $[\sigma_y^2]$ will be put off until Chapter 9.

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 from the true value of the physical quantity the mean is interpreted to represent. With y_t representing this true value and δ_{sys} the systematic error, this relationship can be expressed

$$\mu_y = y_t + \delta_{\text{sys}} \tag{5.2}$$

Sometimes δ_{sys} is constant as y_t varies. In such cases, it is called an offset or zeroing error and μ_y will be always be above or below the true value by the same amount. Sometimes δ_{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, δ_{sys} is a combination of an offset and a scaling error. Or, δ_{sys} might vary in some arbitrary manner.

Combining Eqs. 5.1 and 5.2

$$y = y_t + \delta_y + \delta_{\text{sys}} \tag{5.3}$$

expresses how random and systematic errors contribute to a measurement. *Accuracy* refers to the size of possible systematic errors while *precision* refers to the size of possible random errors.

Systematic errors should typically be neglected in the first round of data analysis in which results and their uncertainties are obtained taking into account random errors only. Then one determines how big systematic errors might be, how they might behave (e.g., offset and/or gain errors), and how they would change the 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 are significant and should be reported separately.

Correlated Data

Measurement values are often statistically independent. One measurement's random error is unlikely to be related to another's. However, there are occasions when correlated measurement errors can be expected. For example, simultaneously measured voltages are prone to correlated random errors because similar electrical interference (e.g., from power lines) might be picked up in both measurements. Measurement timing is a common aspect of correlation because temporal fluctuations are a common manifestation of random error. Measurements made closely in time — shorter than the time scale for the fluctuations — are likely to have correlated random errors.

Correlations can also arise by pre-treating uncorrelated data. For example, the current I in some circuit element might be determined from a measurement of the voltage V across a resistor R wired in series with that element. An application of Ohm's law gives $I = V/R$. In the next chapter you will see how the uncertainty in I is determined. It will depend on the measured values for V and R and their uncertainties. V and R are likely to be statistically independent. However, if the experiment involves making many current determinations using the same resistor, the current values I_i will have correlated errors even if the measured V_i are statistically independent. The correlation arises because a single value of R and its uncertainty are used for calculating all I_i and their uncertainties. The random error in that R-value would then affect all I_i systematically. A common mistake is to treat the I-values as if they were statistically independent.

Correctly dealing with correlated input data is discussed in the next chapter and in Chapter 7. In general, the simplest solution is to work directly with uncorrelated (usually raw) data whenever possible. Thus, the measurements above should be analyzed by substituting V_i/R for I_i in the theoretical predictions so that only the independent variables, V_i and R, appear directly. The quantity R would likely combine algebraically with other model parameters and only at the latest possible stage of the analysis (after any regression analysis, for example) should its value and uncertainty be used.

Chapter 6

Propagation of Error

Direct transformations of random variables will be treated in this chapter. A direct transformation gives M output variables a_k, $k = 1...M$ as defined from N input variables y_i, $i = 1...N$ according to M given functions

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

If the full set of N input variables are acquired repeatedly, they would vary according to their joint probability distribution $p_y(\{y\})$. And as each new set is acquired, the transformation Eqs. 6.1 lead to a new set of M output variables, which then would vary according to their joint distribution $p_a(\{a\})$. Determining the relationship between the input and output distributions is a common analysis task.

Propagation of error provides the variances and covariances that can be expected for the output variables given the values, variances and covariances for the input variables. The treatment will require that the input variations cause only proportional variations in the output variables, i.e., that the output variables follow a first-order Taylor expansion in the input variables. Measurement errors are often small enough to satisfy this requirement and thus propagation of error is one of the more commonly used data analysis procedures. However, it deals only with the input and output covariance matrices and so a couple of special cases will be examined first that deal directly with the joint probability distributions themselves.

Complete Solutions

A complete solution to the problem would determine $p_a(\{a\})$ given $p_y(\{y\})$ and Eqs. 6.1. This calculation can be difficult when several variables are involved. Even with one input and one output variable, extra care is needed if a given value of a can occur for more than one value of y.

The simplest case involves two variables that are single-valued, invertible functions of one another: $a = f(y)$, $y = g(a)$, and $g(f(y)) = y$ for all y. The y and a variables can be considered input and output, respectively, but the transforms must work in either direction. In this special case, the probability distributions $p_a(a)$ and $p_y(y)$ will satisfy

$$p_a(a) = \frac{p_y(y)}{\left|\frac{da}{dy}\right|} \tag{6.2}$$

where the derivative da/dy would be determined directly from $f(y)$ or implicitly from $g(a)$. The inverse function $y = g(a)$ is used to eliminate y from the final expression for $p_a(a)$.

Figure 6.1 shows a simple example where $a = \sqrt{y}$ and the input distribution $p_y(y)$ is a Gaussian of mean μ_y and variance σ_y^2. Assuming the probability for negative y-values is negligible, $a = \sqrt{y}$ and $y = a^2$ are single-valued and $da/dy = 1/2\sqrt{y} = 1/2a$. Equation 6.2 then gives

$$\begin{aligned} p_a(a) &= \frac{p_y(a^2)}{\left|\frac{da}{dy}\right|} \\ &= \frac{2a}{\sqrt{2\pi\sigma_y^2}} e^{-(a^2-\mu_y)^2/2\sigma_y^2} \end{aligned}$$

Equation 6.2 can be extended to two or more variables, but the complexity multiplies quickly. The simplest multivariable case involves two pairs of variables having a unique, one-to-one relationship between each pair — where $a_1 = f_1(y_1, y_2)$, $a_2 = f_2(y_1, y_2)$ and the inverse functions $y_1 = g_1(a_1, a_2)$, $y_2 = g_2(a_1, a_2)$ exist. In this case,

$$p_a(a_1, a_2) = \frac{p_y(y_1, y_2)}{\begin{vmatrix} \frac{\partial a_1}{\partial y_1} & \frac{\partial a_1}{\partial y_2} \\ \frac{\partial a_2}{\partial y_1} & \frac{\partial a_2}{\partial y_2} \end{vmatrix}} \tag{6.3}$$

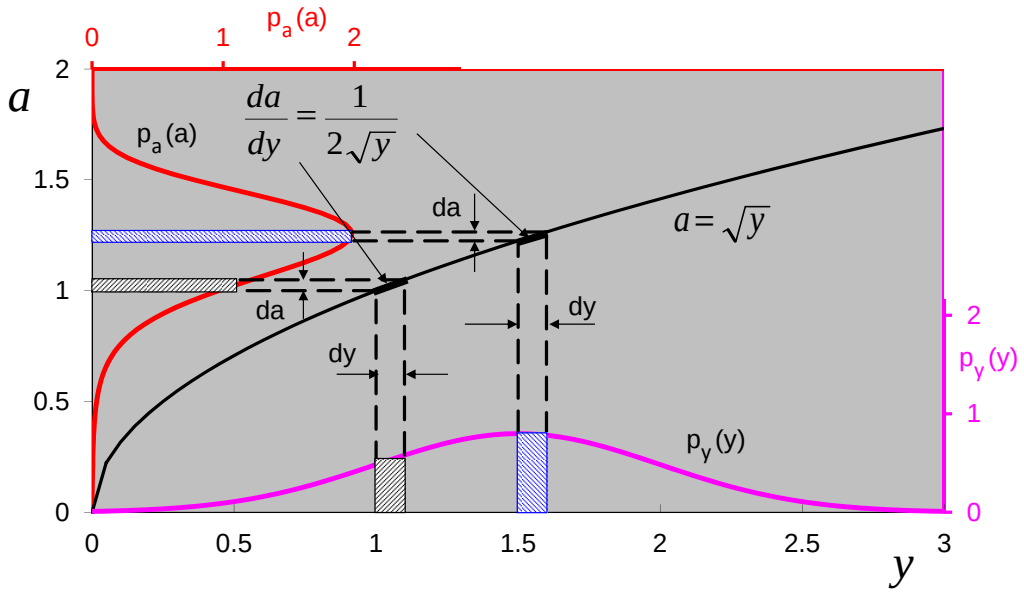


Figure 6.1: Single variable probability density function transformation. $p_y(y)$ is the Gaussian distribution plotted along the horizontal axis. The transformation is $a = \sqrt{y}$ (black curve) and results in $p_a(a)$ — the skewed distribution plotted along the vertical axis. The probabilities in the black shaded bins in each distribution (areas $p_y(y)\,dy$ and $p_a(a)\,da$) must be equal because samples in that y-bin would have square roots that would put them into the corresponding a-bin. The same argument holds for the blue shaded bins. Thus, the ratio $p_a(a)/p_y(y)$ must everywhere be equal to $|dy/da|$ determined by the functional relationship between a and y.

The matrix of derivatives inside the determinant of Eq. 6.3 is the Jacobian for the transformation from the set $\{y\}$ to the set $\{a\}$. Again, inverse functions may be needed to express the result as a function of a_1 and a_2 only.

A useful example for this 2×2 case is the Box-Müller transformation which creates Gaussian random variables from uniform ones. For this transformation y_1 and y_2 are independent and uniformly distributed on the interval $[0, 1]$. Thus, a properly normalized joint pdf for y_1 and y_2 is given by

$$\begin{aligned} p(y_1, y_2) &= 1 \qquad 0 > y_1, y_2 \geq 1 \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{6.4}$$

If y_1 and y_2 are randomly chosen from this joint pdf and then a_1 and a_2 are calculated according to

$$\begin{aligned} a_1 &= \sqrt{-2\ln y_1}\sin 2\pi y_2 \\ a_2 &= \sqrt{-2\ln y_1}\cos 2\pi y_2 \end{aligned} \tag{6.5}$$

then a_1 and a_2 will be independent, Gaussian-distributed random variables — each with a mean of zero and a variance of one. This fact follows after demonstrating that the Jacobian determinant is $2\pi/y_1$ and that Eqs. 6.5 give $y_1 = \exp[-(a_1^2 + a_2^2)/2]$. Equation 6.3 then gives:

$$p(a_1, a_2) = \frac{1}{2\pi}\exp[-(a_1^2 + a_2^2)/2] \tag{6.6}$$

which is a properly normalized joint pdf describing two, independent Gaussian random variables, a_1 and a_2, of zero mean and unit variance.

An integral transformation arises when adding two continuous random variables, say, x and y. Specific values for the sum $z = x + y$ can be made with different combinations of x and y. The general form for the pdf for the sum, $p_z(z)$, can be expressed as a convolution of the pdfs $p_x(x)$ and $p_y(y)$ for x and y.

$$p_z(z) = \int_{-\infty}^{\infty} p_x(x)p_y(z-x)dx \tag{6.7}$$

Note that $p_z(z)$ is the product of the x-probability density at any x with the y-probability density at $y = z - x$ (so that $x + y = z$) integrated over all possible x.

The convolution behavior is illustrated by the frequency distributions shown in Fig. 8.1. The top-left graph shows a histogram for 10,000 samples from a uniform distribution in the range from 0 to 1. The solid line is the expected frequency distribution for this probability distribution. The two graphs at the bottom of this figure show the distributions obtained by adding either two (left) or three (right) such uniform random variables. The solid curves show the expected frequency distributions predicted by one or two applications of Eq. 6.7. Adding two uniform random variables results in a triangular or piecewise linear distribution. Adding a third results in a piecewise quadratic distribution.

Propagation of Error

Propagation of error will here refer to a restricted case of transformations where the ranges for the y_i are 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.

If there were no random error in any of the y_i and they were all equal to their true means μ_i, Eq. 6.1 should then give the true means of the calculated a_k, which will be denoted α_k

$$\alpha_k = f_k(\mu_1, \mu_2, ..., \mu_N) \tag{6.8}$$

Assuming each y_i is always in the range $\mu_i \pm 3\sigma_i$, propagation of error formulas will be valid if, over such ranges, the a_k are accurately represented by a first-order Taylor expansion of each f_k about the values $\mu_1, \mu_2, ..., \mu_N$.

$$\begin{aligned} a_k &= f_k(y_1, y_2, ..., y_N) \\ &= f_k(\mu_1, \mu_2, ..., \mu_N) + \\ &\quad \frac{\partial f_k}{\partial y_1}(y_1 - \mu_1) + \frac{\partial f_k}{\partial y_2}(y_2 - \mu_2) + ... + \frac{\partial f_k}{\partial y_N}(y_N - \mu_N) \\ &= \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i) \end{aligned} \tag{6.9}$$

where Eq. 6.8 has been used in the final step. In linear algebra form, Eq. 6.9 becomes the kth element of the vector equation:

$$\boldsymbol{a} = \boldsymbol{\alpha} + [J_y^a](\boldsymbol{y} - \boldsymbol{\mu}) \tag{6.10}$$

where the $M \times N$ Jacobian has elements given by

$$[J_y^a]_{ki} = \frac{\partial f_k}{\partial y_i} \tag{6.11}$$

Equation 6.10 is often used in the form

$$\Delta\boldsymbol{a} = [J_y^a]\Delta\boldsymbol{y} \tag{6.12}$$

where $\Delta\boldsymbol{a} = \boldsymbol{a} - \boldsymbol{\alpha}$ and $\Delta\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{\mu}$ are the deviations from the means.

To see how the first-order Taylor expansion simplifies the calculations, consider the case where there is only one calculated variable, a, derived from

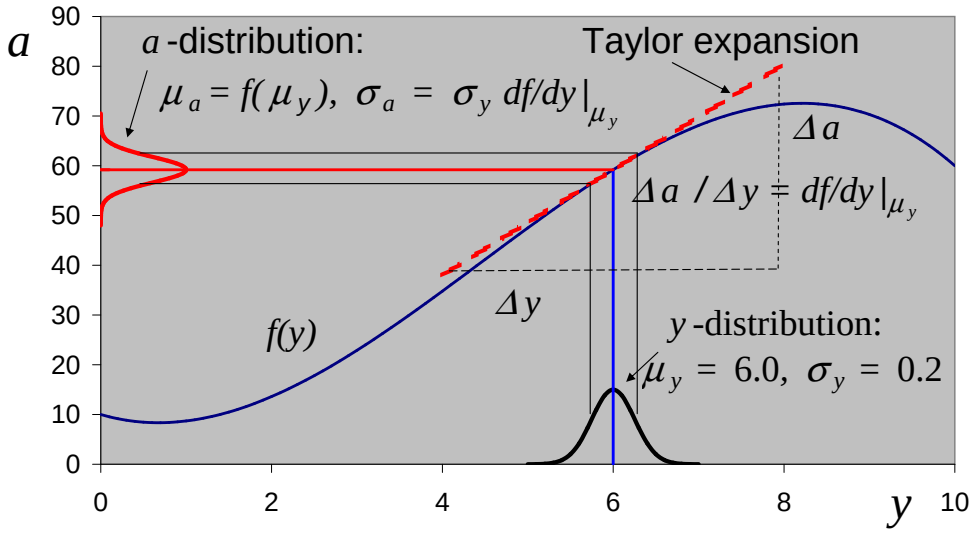


Figure 6.2: Single variable propagation of error. Only the behavior of $f(y)$ over the region $\mu_y \pm 3\sigma_y$ affects the distribution in a.

one random variable, y, according to a given function $a = f(y)$. Figure 6.2 shows the situation where the standard deviation σ_y is small enough that for y-values in the range of their distribution, $a = f(y)$ is well approximated by a straight line — the first-order Taylor expansion of $f(y)$ about μ_y.

$$a = f(\mu_y) + \frac{df}{dy}(y - \mu_y) \tag{6.13}$$

where the derivative is evaluated at μ_y. With a linear relation between a and y, the distribution in y will lead to an identically-shaped distribution in a (or one reversed in shape if the slope is negative). With either sign for the slope, $\mu_a = f(\mu_y)$ and $\sigma_a = \sigma_y|df/dy|$ would hold.

A second-order term in the Taylor expansion — proportional to $(y - \mu_y)^2$ — would warp the linear mapping between a and y. In Fig. 6.2, for example, $a = f(y)$ is always below the tangent line and thus the mean of the a-distribution will be slightly less than $f(\mu_y)$. Such higher order corrections to the mean will be addressed at the end of this chapter, but they will be assumed small enough to neglect more generally.

When more than one random variable contributes to a calculated variable, the one-to-one relationship between the shape of the input and output

distributions is lost. The distribution for the calculated variable becomes something akin to a convolution of the distributions for the contributing variables. The central limit theorem discussed in Chapter 8 states that with enough contributing variables, the calculated quantities will be Gaussian-distributed no matter what distributions govern the contributing variables. But whether the distributions for the a_k turn out to be Gaussian or not and no matter what distributions govern the y_i, if the first-order Taylor expansion is accurate over the likely range of the y_i, propagation of error will accurately predict the most important parameters of $p(\{a\})$ — the means, variances, and covariances as follows.

The mean of a_k is its expectation value and evaluated from Eq. 6.9 becomes

$$\begin{aligned} \mu_{ak} &= \langle a_k \rangle \\ &= \left\langle \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i) \right\rangle \\ &= \alpha_k + \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i} \langle (y_i - \mu_i) \rangle \\ &= \alpha_k \end{aligned} \tag{6.14}$$

where the expectation values $\langle y_i - \mu_i \rangle = 0$ (Eq. 2.17) have 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 an unbiased estimate of the true α_k.

Recall that elements of the covariance matrix for the a_k are defined by:

$$[\sigma_a^2]_{kl} = \langle (a_k - \alpha_k)(a_l - \alpha_l) \rangle \tag{6.15}$$

and that the entire covariance matrix (Eq. 4.22) can be expressed

$$[\sigma_a^2] = \left\langle (\boldsymbol{a} - \boldsymbol{\alpha})(\boldsymbol{a} - \boldsymbol{\alpha})^T \right\rangle \tag{6.16}$$

The kl element of $[\sigma_a^2]$ is the covariance between a_k and a_l and the kk element (kth diagonal element) is the variance of a_k. Substituting Eq. 6.10 and its transpose for $\boldsymbol{a}$ and $\boldsymbol{a}^T$ in Eq. 6.16 then gives:

$$[\sigma_a^2] = \left\langle [J_y^a](\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T [J_y^a]^T \right\rangle \tag{6.17}$$

for which a general element is

$$[\sigma_a^2]_{kl} = \left\langle \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial f_k}{\partial y_i}(y_i - \mu_i)(y_j - \mu_j)\frac{\partial f_l}{\partial y_j} \right\rangle \tag{6.18}$$

Rearranging the terms in the sum, factoring constants (the derivatives) out from the expectation values and then using Eq. 4.21 for the variances and covariances of the y_i, Eq. 6.18 becomes:

$$\begin{aligned} [\sigma_a^2]_{kl} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial f_k}{\partial y_i}\frac{\partial f_l}{\partial y_j} \langle (y_i - \mu_i)(y_j - \mu_j)\rangle \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial f_k}{\partial y_i}\frac{\partial f_l}{\partial y_j} [\sigma_y^2]_{ij} \end{aligned} \tag{6.19}$$

Proceeding from Eq. 6.17, the same logic means the expectation angle brackets can be moved through the Jacobians giving

$$\begin{aligned} [\sigma_a^2] &= [J_y^a] \left\langle (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T \right\rangle [J_y^a]^T \\ &= [J_y^a][\sigma_y^2][J_y^a]^T \end{aligned} \tag{6.20}$$

where Eq. 4.22 was used in the final step. Equations 6.19 and 6.20 give the covariance matrix $[\sigma_a^2]$ associated with the a_k in terms of the covariance matrix $[\sigma_y^2]$ associated with the y_i and the Jacobian describing the relationships between the a_k and the y_i. The partial derivatives in $[J_y^a]$ are simply constants that should be evaluated at the expansion point, $\mu_1, \mu_2, ..., \mu_N$. However, as the true means are typically unknown, the derivatives will have to be evaluated at the measured point $y_1, y_2, ..., y_N$ instead. This difference should not significantly affect the calculations as all $f_k(y_i)$ are assumed to be 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.

Equations 6.19 and 6.20 are the same general formula for propagation of error. Various formulas derived from them are often provided to treat less general cases. One such formula is simply a rewrite for a diagonal element of the covariance matrix. $[\sigma_a^2]_{kk}$, the variance of a_k, denoted σ_{ak}^2, is especially important because its square root is the standard deviation, i.e., the random uncertainty in a_k.

$$\sigma_{ak}^2 = \sum_{i=1}^{N} \left(\frac{\partial f_k}{\partial y_i}\right)^2 \sigma_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\partial f_k}{\partial y_i}\frac{\partial f_k}{\partial y_j}\sigma_{ij} \tag{6.21}$$

The factor of 2 in the double sum arises because Eq. 6.19 would produce two equivalent cross terms while the sum above includes each cross term only once. The first sum includes all terms involving the variances of the y_i, and the second sum — over all pairs i, j where $j > i$ — includes all terms involving the covariances between the y_i. 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 σ_{ak}^2 when all y_i are independent, i.e., when their covariances σ_{ij}, $i \neq j$ are all zero. In this case, Eq. 6.21 simplifies to

$$\sigma_{ak}^2 = \sum_{i=1}^{N} \left(\frac{\partial f_k}{\partial y_i} \right)^2 \sigma_i^2 \tag{6.22}$$

This is the most common propagation of error formula, but it only applies to uncorrelated y_i.

Special conditions can lead to uncorrelated output variables. In general, however, any two output variables will be correlated (have nonzero off-diagonal $[\sigma_a^2]_{kl}$) whether the input variables are correlated or not. For the special case where all y_i are independent, Eq. 6.19 simplifies to

$$[\sigma_a^2]_{kl} \;=\; \sum_{i=1}^{N} \frac{\partial f_k}{\partial y_i} \frac{\partial f_l}{\partial y_i} \sigma_i^2 \tag{6.23}$$

which is not likely to be zero without fortuitous cancellations.

Exercise 3 *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 $\mu_I = 76$ mA and a standard deviation $\sigma_I = 3$ mA. Take V-values from a Gaussian with a mean $\mu_V = 12.2$ V and a standard deviation $\sigma_V = 0.2$ 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 create a scatter plot for the 1000 R, P sample pairs. Calculate the predicted means (Eq. 6.8) and variances (Eq. 6.22) for the R and P distributions and calculate their predicted covariance (Eq. 6.23). Evaluate the sample means (Eq. 2.14) for the 1000 R and P values, their sample variances (Eq. 2.23), and the sample covariance between R and P (Eq. 4.18).

Quantitatively compare the predictions with the sample values. This comparison requires the probability distributions for the sample means, sample variances, and sample covariances. Some of these distributions will be discussed later. Their standard deviations will be given here as an aid to the comparison. The standard deviation of the mean of N 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. A fractional standard deviation is the standard deviation of a quantity divided by the mean of that quantity. The fractional standard deviation of the two sample variances are predicted to be* $\sqrt{2/(N-1)}$. *For* $N = 1000$, *this is about 4.5%, which is also roughly the fractional standard deviation of the sample covariance in this case. So check if your sample variances and covariance agree with predictions at the 9% or two-sigma level.*

Take the 1000-point samples repeatedly while keeping an eye out for how often $\bar{R}$ *and* $\bar{P}$ *are above and below their predicted means.* $\bar{P}$ *should behave as expected — equally likely to be above or below the predicted mean. However,* $\bar{R}$ *is more likely to be above the predicted mean than below it. The reason for this behavior is the nonlinear dependence of R on I as discussed next.*

Correction to the Mean

To check how nonlinearities will affect the mean, Eq. 6.14 is re-derived — this time starting from a second-order Taylor expansion.

For a function $a = f(y_1, y_2)$ of two random variables y_1 and y_2, the second-order Taylor series expansion about the means of y_1 and y_2 becomes

$$\begin{aligned} a &= f(y_1, y_2) \\ &= f(\mu_1, \mu_2) + \frac{\partial f}{\partial y_1}(y_1 - \mu_1) + \frac{\partial f}{\partial y_2}(y_2 - \mu_2) \\ &\quad + \frac{1}{2!}\left(\frac{\partial^2 f}{\partial y_1^2}(y_1 - \mu_1)^2 + \frac{\partial^2 f}{\partial y_2^2}(y_2 - \mu_2)^2 + 2\frac{\partial^2 f}{\partial y_1 \partial y_2}(y_1 - \mu_1)(y_2 - \mu_2)\right) \end{aligned} \tag{6.24}$$

Taking the expectation value of both sides of this equation noting that $\langle a \rangle = \mu_a$, $\langle y_i - \mu_i \rangle = 0$, $\langle (y_i - \mu_i)^2 \rangle = \sigma_i^2$, and $\langle (y_1 - \mu_1)(y_2 - \mu_2) \rangle = \sigma_{12}$, gives

$$\mu_a = f(\mu_1, \mu_2) + \frac{1}{2}\left(\frac{\partial^2 f}{\partial y_1^2}\sigma_1^2 + \frac{\partial^2 f}{\partial y_2^2}\sigma_2^2 + 2\frac{\partial^2 f}{\partial y_1 \partial y_2}\sigma_{12}\right) \tag{6.25}$$

For the power values of Exercise 3, the three terms in parentheses are all zero — the first two because the second derivatives are zero, the third because the covariance between I and V is zero. For the resistance values, however, the term in parentheses is nonzero (due to the $(1/2)(\partial^2 R/\partial I^2)\sigma_I^2$ term only) and adds 0.25 Ω to μ_R — a relatively insignificant shift compared to the standard deviation of the distribution for R in the exercise: $\sigma_R \approx 7\ \Omega$. However, it is significant when compared to the standard deviation for a thousand-point average $\bar{R}$ where $\sigma_{\bar{R}} \approx 0.2\ \Omega$ is also small.

For the more general case in which $a = f(y_1, y_2, ...y_N)$ is a function of N variables, the second-order Taylor expansion becomes

$$a = f(\mu_1, \mu_2, ...\mu_N) + [J_y^a]\Delta\boldsymbol{y} + \frac{1}{2}\left(\Delta\boldsymbol{y}^T[H_{yy}^a]\Delta\boldsymbol{y}\right) \tag{6.26}$$

where $[J_y^a]$ is now a $1 \times N$ matrix (row vector) with the ith element given by $\partial f/\partial y_i$ and $[H_{yy}^a]$, the *Hessian matrix*, is the $N \times N$ symmetric matrix of second derivatives

$$[H_{yy}^a]_{ij} = \frac{\partial^2 f}{\partial y_i \partial y_j} \tag{6.27}$$

Taking expectation values of both sides of Eq. 6.26 then gives a result that can be expressed:

$$\mu_a = f(\mu_1, \mu_2, ...\mu_N) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}[H_{yy}^a]_{ij}[\sigma_y^2]_{ij} \tag{6.28}$$

Chapter 7

Regression Analysis

Regression analysis refers to procedures 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. A common example is a fit to a set of measured (x_i, y_i) data points predicted to obey a straight-line relationship: $y = mx + b$. Regression analysis then provides an estimate of the slope m and the intercept b based on the data.

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, Poisson, or binomial probability distribution. The independent variables (the x_i in the straight line fit) can also be random variables, but this possibility will only be considered after treating the simpler case where the independent variables are known exactly.

The dependent variables y_i in a regression analysis are typically assumed to be statistically independent with no correlations in their error distributions. If correlations exist, they should be taken into account. One treatment is addressed later in this chapter, but until then, all y_i will be considered statistically independent.

The model is that the mean of the distribution for each y_i depends on the independent variables associated with that data point through a fitting function with M unknown theory parameters α_k, $k = 1...M$

$$\mu_i = F_i(\alpha_1, \alpha_2, ..., \alpha_M) \tag{7.1}$$

where the subscript i in $F_i(\{\alpha\})$ denotes the independent variables. Equation 7.1 is intentionally written without any explicit independent variables.

In a regression analysis they simply distinguish the point by point dependencies of the μ_i on the α_k. As far as regression is concerned, the fitting function is simply N equations for the μ_i in terms of the M values for the α_k.

Equation 7.1 is written as defining the true means μ_i in terms of the true theory parameters α_k. Except in simulations, both are usually unknown. The version

$$y_i^{\rm fit} = F_i(a_1, a_2, ...a_M) \tag{7.2}$$

gives the corresponding quantities determined by the fit. The $y_i^{\rm fit}$ are the estimates of μ_i obtained via the same fitting function as Eq. 7.1 but with each α_k replaced by a corresponding estimate a_k as determined by the fit. Even though the a_k depend on the y_i only indirectly via the fitting process, they are nonetheless associated with one particular data set and are random variables; the fitted a_k would change if the y_i were resampled.

Principle of Maximum Likelihood

The estimates a_k are obtained according to the *principle of maximum likelihood* in that they are chosen to maximize the probability of the data set from which they are derived. As a result, should the experiment and theory be deemed incompatible, they will be incompatible regardless of the parameter values. Any other values will only make the data less likely. Any parameter determined by this principle is called a *maximum likelihood estimate* or MLE.

With the y_i statistically independent, the product rule applies. Variables governed by Gaussian, Poisson, or binomial distributions have joint probabilities given by Eqs. 4.12, 4.13 or 4.14, respectively. These give the actual probability for the data set — larger for data sets that are more likely and smaller for sets that are less likely. This product probability becomes dependent on the a_k when the μ_i are replaced with the estimates $y_i^{\rm fit}$ as expressed through Eq. 7.2. The a_k that produce the $y_i^{\rm fit}$ that produce the largest possible joint probability become the MLE's for the α_k. The $y_i^{\rm fit}$ are commonly called the *best fit* (to the input y_i) and thus the a_k are also called the *best-fit parameters.*

For a continuous function $f(a_1, a_2, ...)$, conditions for a local maximum are that $\partial f/\partial a_k = 0$ for all k. When f represents a joint probability, satisfying these conditions usually finds a global maximum. To find the maximum, a useful trick is to first take the natural logarithm of $f(a_1, a_2, ...)$ and maximize

that. This works because $\partial(\ln f)/\partial a_k = (1/f)\,\partial f/\partial a_k$ and, because f will be nonzero and finite, where one derivative is zero, so is the other.

The natural logarithm of the joint probability is called the *log likelihood* $\mathcal{L}$ and thus defined by

$$\mathcal{L} = \ln P \tag{7.3}$$

Using $\mathcal{L}$ simplifies the math because it transforms the products into sums which are easier to differentiate. The actual value of $\mathcal{L}$ is unimportant. Its only use is in maximizing the probability with respect to the fitting parameters by imposing the condition

$$\frac{\partial \mathcal{L}}{\partial a_k} = 0 \tag{7.4}$$

for each a_k. In general, a_k — and the y_i^{fit} they produce — will refer to MLE values. Where derivatives are taken, they are normally evaluated at the MLE values.

As mentioned, the dependence of $\mathcal{L}$ on the a_k arises when the y_i^{fit} are used as estimates of the μ_i in the joint probability. Because any term that is independent of y_i^{fit} will automatically have a zero derivative with respect to all a_k, only terms that depend on μ_i need to be kept when evaluating $\mathcal{L}$.

For a Gaussian data set where $P(\{y\})$ is given by Eq. 4.12, Eq. 7.3 gives (after dropping terms that are independent of μ_i and substituting y_i^{fit} for μ_i)

$$\mathcal{L} \;=\; -\frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - y_i^{\text{fit}}\right)^2}{\sigma_i^2} \tag{7.5}$$

For a Poisson data set (Eq. 4.13) $\mathcal{L}$ becomes

$$\mathcal{L} \;=\; \sum_{i=1}^{N} y_i \ln y_i^{\text{fit}} - y_i^{\text{fit}} \tag{7.6}$$

And for a binomial data set (Eq. 4.14) $\mathcal{L}$ becomes

$$\mathcal{L} \;=\; \sum_{i=1}^{N} y_i \ln y_i^{\text{fit}} + (\mathcal{N}_i - y_i)\ln(\mathcal{N}_i - y_i^{\text{fit}}) \tag{7.7}$$

Because $\mathcal{L}$ depends only indirectly on the a_k through the y_i^{fit}, the derivatives in Eq. 7.4 are evaluated according to the chain rule

$$\frac{\partial \mathcal{L}}{\partial a_k} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial y_i^{\text{fit}}} \frac{\partial y_i^{\text{fit}}}{\partial a_k} \tag{7.8}$$

where the partial derivatives $\partial y_i^{\rm fit}/\partial a_k$ are determined by the particular form for the fitting function, Eq. 7.2.

For Gaussian-distributed y_i, where $\mathcal{L}$ is given by Eq. 7.5, Eq. 7.4 becomes

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_k}\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - y_i^{\rm fit})^2}{\sigma_i^2}\right] \\ &= \sum_{i=1}^{N}\frac{y_i - y_i^{\rm fit}}{\sigma_i^2}\,\frac{\partial y_i^{\rm fit}}{\partial a_k} \end{aligned} \tag{7.9}$$

For Poisson-distributed y_i, where $\mathcal{L}$ is given by Eq. 7.6, Eq. 7.4 becomes

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_k}\left[\sum_{i=1}^{N}\left(y_i \ln y_i^{\rm fit} - y_i^{\rm fit}\right)\right] \\ &= \sum_{i=1}^{N}\left(\frac{y_i}{y_i^{\rm fit}} - 1\right)\,\frac{\partial y_i^{\rm fit}}{\partial a_k} \\ &= \sum_{i=1}^{N}\frac{y_i - y_i^{\rm fit}}{y_i^{\rm fit}}\,\frac{\partial y_i^{\rm fit}}{\partial a_k} \end{aligned} \tag{7.10}$$

Equation 7.10 is remarkably similar to Eq. 7.9. The numerator in each term of both equations, $\left(y - y_i^{\rm fit}\right)\partial y_i^{\rm fit}/\partial a_k$, is the same and the denominator in Eq. 7.10, $y_i^{\rm fit}$, is, in fact, the variance σ_i^2 of a Poisson distribution (Eq. 3.8) having a mean of $y_i^{\rm fit}$ — and after all, the mean is exactly what $y_i^{\rm fit}$ is estimating. Thus, if

$$\sigma_i^2 = y_i^{\rm fit} \tag{7.11}$$

is used when fitting Poisson-distributed y_i, Eq. 7.10 is exactly the same as Eq. 7.9.

For binomial-distributed y_i, where $\mathcal{L}$ is given by Eq. 7.7, Eq. 7.4 becomes

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_k}\left[\sum_{i=1}^{N} y_i \ln y_i^{\rm fit} + (\mathcal{N}_i - y_i)\ln(\mathcal{N}_i - y_i^{\rm fit})\right] \\ &= \sum_{i=1}^{N}\left(\frac{y_i}{y_i^{\rm fit}} - \frac{\mathcal{N}_i - y_i}{\mathcal{N}_i - y_i^{\rm fit}}\right)\,\frac{\partial y_i^{\rm fit}}{\partial a_k} \\ &= \sum_{i=1}^{N}\frac{y_i - y_i^{\rm fit}}{y_i^{\rm fit}(1 - y_i^{\rm fit}/\mathcal{N}_i)}\,\frac{\partial y_i^{\rm fit}}{\partial a_k} \end{aligned} \tag{7.12}$$

Once again, the numerator is the same as in Eq. 7.9 and the denominator is the variance of a binomial distribution (Eq. 3.6) having a mean of y_i^{fit} with $\mathcal{N}_i$ trials. Thus, if

$$\sigma_i^2 = y_i^{\text{fit}} \left(1 - \frac{y_i^{\text{fit}}}{\mathcal{N}_i}\right) \tag{7.13}$$

is used when fitting binomial-distributed y_i, Eq. 7.12 is also the same as Eq. 7.9.

Thus, with the understanding that the σ_i^2 are appropriately chosen for the y_i at hand, Eqs. 7.9, 7.10 and 7.12 can all be rewritten in the form

$$\sum_{i=1}^{N} \frac{y_i - y_i^{\text{fit}}}{\sigma_i^2} \frac{\partial y_i^{\text{fit}}}{\partial a_k} = 0 \tag{7.14}$$

This is the kth element of a set of M simultaneous equations which must be satisfied by the M unknown a_k in order for them to be the MLE's for the α_k.

A regression analysis finds the a_k that maximize $\mathcal{L}$ or, equivalently, that satisfy equation set 7.14 and it determines their covariance matrix $[\sigma_a^2]$. Before discussing how to do this, it is worthwhile to first examine the relationship between the maximum likelihood principle and the least-squares principle.

Least-Squares Principle

Aside from an overall multiplicative factor of $-1/2$, the Gaussian log-likelihood of Eq. 7.5 is a sum of positive (squared) terms — one for each y_i. This sum is the *chi-square* (χ^2) random variable and is given by

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - y_i^{\text{fit}}\right)^2}{\sigma_i^2} \tag{7.15}$$

Thus for Gaussian-distributed y_i, $\mathcal{L} = -\chi^2/2$ and maximizing $\mathcal{L}$ is the same as minimizing χ^2. Finding the minimum χ^2 proceeds as for maximizing the log likelihood — by setting to zero its derivatives with respect to each a_k — and leads to the same equation set, Eq. 7.14, to solve for the a_k. Since the χ^2 is a "sum of squares," minimizing it is said to be a *least-squares* procedure.

Minimizing the χ^2 is intuitively satisfying no matter what distribution governs the y_i. A smaller χ^2 value means a better fit with deviations, $y_i - y_i^{\text{fit}}$, of smaller magnitude. And there is a sensible dependence on the standard

deviations — with equal contributions to the χ^2 sum for equal deviations in units of σ_i. That is, deviations are compared relative to their uncertainty.

As will be shown next, minimizing a χ^2 can lead to maximum likelihood fitting parameters even for non-Gaussian random variables.

Iteratively Reweighted Least Squares

While maximizing the log likelihood is a straightforward way to find the best-fit a_k, minimizing the χ^2 is a much more common procedure. Moreover, such a least squares analysis can also be applied to any y_i for which the maximum likelihood condition, Eq. 7.4, can be cast in the form of Eq. 7.14 with an appropriate value of σ_i^2. For example, least squares can be applied to Poisson or binomial y_i using "best-fit variances" $\sigma_i^2 = y_i^{\text{fit}}$ or $\sigma_i^2 = y_i^{\text{fit}}(1 - y_i^{\text{fit}}/\mathcal{N}_i)$, respectively.

The logic for using least squares to find maximum likelihood solutions is based on the fact that, as demonstrated for Gaussian y_i, minimizing the χ^2 *with the* σ_i^2 *held fixed* is the same as solving Eq. 7.14 — the equation set for maximum likelihood solutions for Gaussian, Poisson, and binomial y_i. Consequently, for Poisson and binomial y_i, if the σ_i^2 are held fixed at their best-fit values when finding the χ^2 minimum, the a_k at that minimum will be maximum likelihood estimates. Of course, there is a minor "chicken-and-egg" problem because the best-fit σ_i^2 are needed to minimize the correct χ^2, but the a_k at that minimum are needed to determine those σ_i^2. This problem is easily solved by iteration. In this case it's called iteratively reweighted least squares — IRLS for short.

IRLS begins with an initial χ^2 minimization using a constant or other estimate for the σ_i^2. The a_k and y_i^{fit} at this minimum are then used to determine a new set of σ_i^2, which are then held fixed in the next χ^2 minimization. Additional χ^2 minimizations are performed (always with fixed σ_i^2) until they are self-consistent — with the σ_i^2 as evaluated at a previous χ^2 minimum leading to those same (now best-fit) a_k.

The iterations tend to converge quickly because the fitting parameters typically depend only weakly on the σ_i^2. The dependence is weak enough that for Poisson-distributed y_i, it is often assumed that the input y_i should be good enough estimates of the y_i^{fit} for the purposes of calculating the σ_i^2. The fit then uses $\sigma_i^2 = y_i$ without iteration. This is not unreasonable if the y_i are all 100 or more so that the errors in using y_i instead of y_i^{fit} for σ_i^2 are unlikely to be more than 10%. However, if many of the y_i are relatively low,

using $\sigma_i^2 = y_i$ can give significantly different fit parameters. With today's computers and programming tools there is hardly a reason not to use the correct σ_i^2 via iteration.

Sample Mean and Variance

A small detour is now in order — a return to the subject of distribution sampling in light of the principle of maximum likelihood. Recall the definitions for the sample mean (Eq. 2.14) and the sample variance (Eq. 2.23) for samples from a common distribution. Can they be demonstrated to satisfy the principle of maximum likelihood?

The input data in this case now consist of a sample set, y_i, $i = 1...N$, all from the exact same probability distribution. The model is that the true mean is the same for all y_i and a maximum likelihood estimate of its value is sought. This is the simple case of a fit to the constant function — $y_i^{\text{fit}} = \bar{y}$ for all i. The estimate is given the symbol $\bar{y}$ for reasons that will be obvious shortly. As a one-parameter fit, $M = 1$ with $y_i^{\text{fit}} = F_i(a_1) = a_1 = \bar{y}$ and $\partial y_i^{\text{fit}}/\partial a_1 = 1$ for all i. Because all y_i are from the same distribution, the σ_i^2 will also be the same for all y_i; $\sigma_i^2 = \sigma_y^2$ for all i.

After bringing the constant σ_y^2 out of the summations on both sides of Eq. 7.14, this quantity cancels. Setting the derivatives to one, and setting $y_i^{\text{fit}} = \bar{y}$ for all i, this equation becomes simply:

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} \bar{y} \tag{7.16}$$

The right side is simply $N\bar{y}$ and solving for $\bar{y}$ then reproduces the standard definition of the sample mean, Eq. 2.14.

The sample mean has now been proven to be the MLE for the distribution mean for variables governed by a Gaussian, Poisson or binomial distribution. The sample mean $\bar{y}$ has previously been shown (see Eq. 2.15) to be an unbiased estimate of the true mean μ_y. Thus, for these three distributions, the MLE is unbiased. The principle of maximum likelihood does not always produce unbiased estimates. A biased estimate will have a distribution mean above or below the true mean and will not give the true mean "on average." Bias is considered a significant flaw and, consequently, corrections for it are sought and applied.

For a sample set of y_i from a common Gaussian distribution, σ_y^2 becomes a second parameter suitable for determination according to the principle of maximum likelihood. Is the sample variance s_y^2 of Eq. 2.23 an MLE for σ_y^2? For this case, the joint probability would now have to include all terms dependent on μ_y and σ_y. Taking the natural logarithm of Eq. 4.12 replacing all μ_i with their estimate $\bar{y}$ and replacing all σ_i^2 with their estimate s_y^2 (and dropping all terms independent of these two variables) gives

$$\mathcal{L} = -N \ln s_y - \frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{s_y^2} \tag{7.17}$$

Now, the derivatives of $\mathcal{L}$ with respect to both $\bar{y}$ and s_y must be set equal to zero in order for them to be the MLE's for μ_y and σ_y. Nothing changes for the $\bar{y}$ equation and the sample mean (Eq. 2.14) stays the MLE for μ_y. Setting the derivative with respect to s_y equal to zero then gives

$$\begin{aligned} 0 &= \frac{\partial}{\partial s_y} \left[-N \ln s_y - \frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{s_y^2} \right] \\ &= -\frac{N}{s_y} + \frac{1}{s_y^3} \sum_{i=1}^{N} (y_i - \bar{y})^2 \end{aligned} \tag{7.18}$$

with the solution

$$s_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{7.19}$$

However, this s_y^2 is seldom used because it is biased — having an expectation value smaller than the true variance. As will be demonstrated in Exercise 5, the more common estimate given by Eq. 2.23, with $N - 1$ in place of N in the denominator, is preferred because it is unbiased.

Note that the mean of a single sample (a sample of size $N = 1$) is well defined. It is that sample value and thus also the MLE of the true mean of its parent distribution. No estimate of the true variance can be obtained with only a single sample. Neither Eq. 7.19, which gives an unphysical estimate of $s_y^2 = 0$, nor Eq. 2.23, which gives an indeterminate value of $0/0$, can be used. It takes at least two samples to get a sample variance, for which Eq. 2.23 gives the unbiased estimate $s_y^2 = (y_1 - y_2)^2/2$.

Of course, samples of size one or two are the smallest possible. Larger samples give sample means and sample variances which are more precise — more closely clustered around the true mean and the true variance. The

variance of $\bar{y}$ is given in the next exercise. The variance of s_y^2 is discussed in Chapter 9.

Exercise 4 *The* variance of the mean $\sigma_{\bar{y}}^2$ *is most easily determined from Eq. 2.19 — 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*

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{N} \tag{7.20}$$

Hint: Re-express $\bar{y}^2$ *as*

$$\begin{aligned} \bar{y}^2 &= \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) \left(\frac{1}{N} \sum_{j=1}^{N} y_j \right) \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \end{aligned} \tag{7.21}$$

before taking the expectation value. Each $\bar{y}$ *in* $\bar{y}^2$ *gets its own summation index to clearly enumerate terms with* $i = j$ *and* $i \neq j$ *for use with Eq. 4.9.*

Eq. 7.20 indicates that the standard deviation of the mean of N samples is $\sqrt{N}$ times smaller than the standard deviation of a single sample, e.g., the average of 100 samples is 10 times more precise an estimate of the true mean than is a single sample. The *sample variance of the mean* is then defined by

$$s_{\bar{y}}^2 = \frac{s_y^2}{N} \tag{7.22}$$

and is an unbiased estimate of the true $\sigma_{\bar{y}}^2$.

Exercise 5 *Show that Eq. 2.23 is unbiased and satisfies Eq. 2.22. Hint 1: Explain why each of the* N *terms in Eq. 2.23 has the same expectation value and use this fact to eliminate 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. 2.14 and 4.9 and/or results from Exercise 4 as needed for the individual terms.*

Weighted Mean

The sample mean $\bar{y}$ is an unweighted average; each y_i has an equal effect on its value. Suppose that a sample of y_i, $i = 1...N$ are again all predicted to have the same distribution mean μ_y. For example, they might be results for the same quantity obtained from different data sets or by different research groups. But now, the y_i don't all have the same uncertainty — each comes with its own standard deviation σ_i. The y_i and the σ_i are given and the MLE for μ_y is to be determined. Taking that MLE as m_y, the regression problem is again a simple fit to a constant — $M = 1$, with $y_i^{\text{fit}} = F_i(a_1) = a_1 = m_y$ and $\partial y_i^{\text{fit}}/\partial a_1 = 1$ for all i. The maximum likelihood solution must then satisfy Eq. 7.14 which becomes the single equation

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} = \sum_{i=1}^{N} \frac{m_y}{\sigma_i^2} \tag{7.23}$$

m_y can be factored from the sum giving the result

$$m_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \Bigg/ \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{7.24}$$

This is a *weighted average* of the y_i

$$m_y = \frac{w_1 y_1 + w_2 y_2 + ... + w_N y_N}{w_1 + w_2 + ... + w_N} \tag{7.25}$$

where the weight w_i for each y_i is given by the inverse of the variance for that y_i

$$w_i = \frac{1}{\sigma_i^2} \tag{7.26}$$

Larger standard deviations indicate less precisely known y-values and, appropriately, smaller weights in the average. Weighting a data point's contribution according to the inverse of its distribution's variance persists in all regression problems. The larger the σ_i, the smaller the weight of that sample and thus the smaller the effect of that sample on the fitting parameters.

Propagation of error then gives the variance of m_y via Eq. 6.22 with

Eqs. 7.25 and 7.26 as

$$\begin{aligned}
\sigma_{my}^2 &= \sum_{i=1}^{N} \left(\frac{\partial m_y}{\partial y_i} \right)^2 \sigma_i^2 \\
&= \sum_{i=1}^{N} w_i^2 \sigma_i^2 \Bigg/ \left(\sum_{i=1}^{N} w_i \right)^2 \\
&= \sum_{i=1}^{N} w_i \Bigg/ \left(\sum_{i=1}^{N} w_i \right)^2 \\
&= 1 \Bigg/ \sum_{i=1}^{N} w_i
\end{aligned} \tag{7.27}$$

Inverting both sides, this result can be expressed more symmetrically as

$$\frac{1}{\sigma_{my}^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{7.28}$$

Effectively, Eqs. 7.24 and 7.28 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_{my}.

Linear Regression

The rest of this chapter is devoted to providing solutions to the maximum likelihood condition (Eq. 7.14) for fixed σ_i^2. It covers linear and nonlinear regression and then several specialized cases such as data sets with uncertainties in the independent variables, data sets with correlated y_i, and data sets collected after an instrument calibration.

In linear algebra form, Eq. 7.14 is just the kth element of the vector equation

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right) = 0 \tag{7.29}$$

Bringing the $\boldsymbol{y}^{\text{fit}}$ term to the right puts this equation in a form more useful for finding solutions.

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y} = [J_a^y]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y}^{\text{fit}} \tag{7.30}$$

In these equations, $[J_a^y]$ is the $N \times M$ Jacobian of partial derivatives of y_i^{fit} with respect to each a_k as determined by the fitting function

$$[J_a^y]_{ik} = \frac{\partial y_i^{\text{fit}}}{\partial a_k} \tag{7.31}$$

and $\left[\sigma_y^2\right]^{-1}$ — the inverse of the covariance matrix — is called the *weighting matrix* and given by

$$\left[\sigma_y^2\right]^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & \cdots \\ 0 & 1/\sigma_2^2 & 0 & \cdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/\sigma_N^2 \end{pmatrix} \tag{7.32}$$

It is recommended that the reader check Eq. 7.29 and verify that it is, indeed, a vector equation having M elements with the kth element reproducing Eq. 7.14 including the proper summation over the i index. Consequently, solving Eq. 7.30 solves all M equations simultaneously and gives the MLE for all a_k at once.

The linear algebra formulas discussed next are demonstrated in Excel for a quadratic fit using array formulas in *Linear Regression Algebra.xlsm* available on the lab website.

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

$$\begin{aligned} y_i^{\text{fit}} &= a_1 f_1(x_i) + a_2 f_2(x_i) + ... + a_M f_M(x_i) \\ &= \sum_{k=1}^{M} a_k f_k(x_i) \end{aligned} \tag{7.33}$$

where the $f_k(x)$ are given functions of x with no unknown parameters. For example, a data set for a cart rolling on an inclined track might consist of the measured cart position y_i versus the time t_i at each measurement. This data might then be checked against a predicted quadratic based on motion at constant acceleration:

$$y_i^{\text{fit}} = a_1 + a_2 t_i + a_3 t_i^2 \tag{7.34}$$

This model is linear in the three parameters a_1, a_2, and a_3 associated with the basis functions: $f_1(t_i) = 1$, $f_2(t_i) = t_i$, and $f_3(t_i) = t_i^2$.

Note that when the true parameters α_k are used in place of the fitted a_k, Eq. 7.33 gives the true means μ_i of the distributions for the y_i.

$$\begin{aligned} \mu_i &= \alpha_1 f_1(x_i) + \alpha_2 f_2(x_i) + ... + \alpha_M f_M(x_i) \\ &= \sum_{k=1}^{M} \alpha_k f_k(x_i) \end{aligned} \tag{7.35}$$

Equation 7.31 for the Jacobian is then the $N \times M$ matrix given by

$$[J_a^y] = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_M(x_2) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_M(x_N) \end{pmatrix} \tag{7.36}$$

Note that this Jacobian is independent of the entire set of a_k. It is this condition that determines when linear regression is appropriate. The columns of this matrix must be linearly independent to produce a unique set of a_k at the χ^2 minimum.

Equation 7.33 for the column vector of y_i^{fit} values can then be expressed by the vector equation

$$\boldsymbol{y}^{\text{fit}} = [J_a^y]\, \boldsymbol{a} \tag{7.37}$$

and Eq. 7.35 for the true means becomes

$$\boldsymbol{\mu} = [J_a^y]\, \boldsymbol{\alpha} \tag{7.38}$$

Substituting Eq. 7.37 into Eq, 7.30 then gives

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y} = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y]\, \boldsymbol{a} \tag{7.39}$$

In this equation, $[J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y]$ is an $M \times M$ symmetric matrix whose inverse turns out to be the parameter covariance matrix. It will be referred to as the X-matrix ($[X]$ in equations).

$$[X] = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y] \tag{7.40}$$

so that Eq. 7.39 becomes

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y} = [X]\boldsymbol{a} \tag{7.41}$$

This equation is solved for the best-fit parameter vector by determining the inverse of the X-matrix and multiplying it from the left on both sides. This gives

$$\boldsymbol{a} = [X]^{-1} \left[J_a^y\right]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y} \tag{7.42}$$

or

$$\boldsymbol{a} = \left[J_y^a\right]^\dagger \boldsymbol{y} \tag{7.43}$$

where

$$\left[J_y^a\right]^\dagger = [X]^{-1} \left[J_a^y\right]^T \left[\sigma_y^2\right]^{-1} \tag{7.44}$$

is an $M \times N$ matrix called the *weighted Moore-Penrose pseudoinverse* of $[J_a^y]$. $\left[J_y^a\right]^\dagger$ is not a true matrix inverse, which is defined for square matrices only. However, note that:

$$\begin{aligned} \left[J_y^a\right]^\dagger [J_a^y] &= [X]^{-1} \left[J_a^y\right]^T \left[\sigma_y^2\right]^{-1} [J_a^y] \\ &= [X]^{-1}[X] \\ &= [1] \end{aligned} \tag{7.45}$$

where $[1]$ is the $M \times M$ identity matrix. This product of the Jacobian and its pseudoinverse yields the $M \times M$ identity matrix — a key inverse-like property.

The a_k are random variables. For each new input set of y_i, the output set of a_k would vary according to Eq. 7.43. What can be expected for the means, variances and covariances for the a_k if the input data sets were resampled over and over again?

That the distribution for a_k will be unbiased (have a mean of α_k) can be demonstrated upon taking the expectation value of both sides of Eq. 7.43

$$\begin{aligned} \langle \boldsymbol{a} \rangle &= \left[J_y^a\right]^\dagger \langle \boldsymbol{y} \rangle \\ &= \left[J_y^a\right]^\dagger \boldsymbol{\mu} \\ &= \left[J_y^a\right]^\dagger [J_a^y] \, \boldsymbol{\alpha} \\ &= \boldsymbol{\alpha} \end{aligned} \tag{7.46}$$

where $\langle \boldsymbol{y} \rangle = \boldsymbol{\mu}$ (from $\langle y_i \rangle = \mu_i$) was used to get to line 2, Eq. 7.38 was used to get to line 3, and Eq. 7.45 was used to get to line 4.

Keep in mind the kth row of Eq. 7.43 is

$$a_k = \left[J_y^a\right]_{k1}^\dagger y_1 + \left[J_y^a\right]_{k2}^\dagger y_2 + ... \left[J_y^a\right]_{kN}^\dagger y_N \tag{7.47}$$

and defines a direct linear relationship for each output parameter a_k from any set of input y_i. Consequently, the propagation of error formula (Eq. 6.20) can be used to determine the parameter covariance matrix $[\sigma_a^2]$ in terms of the input covariance matrix $[\sigma_y^2]$. Recall, the $M \times N$ Jacobian appearing in Eq. 6.20 has elements defined by $\partial a_k / \partial y_i$, which Eq. 7.47 shows are just the elements of $\left[J_y^a\right]^\dagger$. Thus the parameter covariance matrix is simply

$$[\sigma_a^2] = \left[J_y^a\right]^\dagger [\sigma_y^2] \left[J_y^a\right]^{\dagger T} \tag{7.48}$$

$\left[J_y^a\right]^\dagger$ can be eliminated from this equation in favor of $[J_a^y]$ using Eqs. 7.44 and 7.40. The Jacobian $[J_a^y]$ — the derivatives of the fitting function with respect to the fitting parameters — is the simpler of the two and would be needed anyway to determine $\left[J_y^a\right]^\dagger$. Note that the rule for taking the transpose gives $\left[J_y^a\right]^{\dagger T} = \left[\sigma_y^2\right]^{-1} [J_a^y] [X]^{-1}$ because $\left[\sigma_y^2\right]^{-1}$ and $[X]^{-1}$ are symmetric about their matrix diagonals and thus they are their own transpose. With these notes, proceeding from Eq. 7.48 gives

$$\begin{aligned}
[\sigma_a^2] &= [X]^{-1} [J_a^y]^T \left[\sigma_y^2\right]^{-1} [\sigma_y^2] \left[\sigma_y^2\right]^{-1} [J_a^y] [X]^{-1} \\
&= [X]^{-1} [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y] [X]^{-1} \\
&= [X]^{-1} [X] [X]^{-1} \\
&= [X]^{-1}
\end{aligned} \tag{7.49}$$

Taking the inverse of both sides and using Eq. 7.40 then gives

$$\left[\sigma_a^2\right]^{-1} = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y] \tag{7.50}$$

Equation 6.20, $[\sigma_a^2] = [J_y^a][\sigma_y^2][J_y^a]^T$, and Eq. 7.50, above, are complementary relationships for the two most common statistical procedures. Equation 6.20 applies to propagation of error and gives the output covariance matrix in terms of the input covariance matrix and the $M \times N$ Jacobian, $[J_y^a]_{ki} = \partial a_k / \partial y_i$ based on the direct relationships: $a_k = f_k(\{y_i\})$. Equation 7.50 applies to regression analysis and gives the output weighting matrix in terms of the input weighting matrix and the $N \times M$ Jacobian $[J_a^y]_{ik} = \partial y_i^{\text{fit}} / \partial a_k$ based on the fitting model: $y_i^{\text{fit}} = F_i(\{a_k\})$.

As Eq. 7.47 demonstrates, every a_k is a purely linear function of the y_i and thus the first-order Taylor expansion for each a_k about any set of y_i is exact. Recall from Chapter 6, this implies Eq. 6.20 (here, Eq. 7.48 and

Eq. 7.50) is exact. Linear fitting functions lead to linear direct relationships, and for both, the validity of the calculated parameter covariance matrix does not rely on keeping errors small. The calculated $[\sigma_a^2]$ is exact for any given $[\sigma_y^2]$ and for any distribution governing the y_i.

As was demonstrated in Chapter 6, when the a_k are nonlinear functions of the y_i, Eq. 6.20 remains valid, but relies on keeping the σ_i sufficiently small. While Eq. 7.50 was obtained for linear fitting functions, as will be demonstrated later in this chapter, it also remains valid for nonlinear fitting functions if the σ_i are kept small enough.

Equally-Weighted Linear Regression

Occasionally, all y_i are obtained using the same technique and have the same uncertainty. Or, lacking better estimates, the uncertainties might simply be assumed equal. 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 regression equations simplify because the covariance matrix and the weighting matrix are proportional to the identity matrix; $[\sigma_y^2] = \sigma_y^2[1]$ and $\left[\sigma_y^2\right]^{-1} = (1/\sigma_y^2)[1]$, where $[1]$ is the $N \times N$ identity matrix. With this substitution, Eq. 7.40 becomes

$$[X] = \frac{1}{\sigma_y^2}[X_u] \tag{7.51}$$

where

$$[X_u] = [J_a^y]^T\,[J_a^y] \tag{7.52}$$

is the $[X]$ matrix without the intervening weighting matrix and is thus independent of σ_y. The inverse of Eq. 7.51 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. 7.49, $[X]^{-1}$ is the parameter covariance matrix. That is,

$$[\sigma_a^2] \;=\; \sigma_y^2[X_u]^{-1} \tag{7.53}$$

thereby demonstrating that every element of the parameter covariance matrix is proportional to σ_y^2 and thus the standard deviation of every parameter (square root of the corresponding diagonal element) is proportional to σ_y.

Equation 7.42 for the parameter values becomes

$$\boldsymbol{a} = [X_u]^{-1}\,[J_a^y]^T\,\boldsymbol{y} \tag{7.54}$$

showing that σ_y^2 has canceled and thus the parameter values themselves are independent of its value. The independence can also be inferred from the χ^2 of Eq. 7.15, which becomes

$$\chi^2 = \frac{1}{\sigma_y^2} \sum_{i=1}^{N} \left(y_i - y_i^{\text{fit}} \right)^2 \tag{7.55}$$

Recall that the best-fit parameters would be those that minimize this χ^2. Because σ_y^2 factors from the sum, no matter its value, the same sum of squared deviations must be minimized and thus the same a_k will be obtained.

Nonlinear Regression

Linear regression requires that the Jacobian $[J_a^y]$ be 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 sample of a gamma-ray-emitting radioisotope placed in front of a Geiger counter. If the half-life of the isotope is a few minutes or so, the number of detected gamma rays y_i might be measured over consecutive ten-second intervals as a function of the interval starting time t_i relative to the start of the experiment. The y_i would decrease as the sample decays and a model predicting exponential decay would be represented

$$y_i^{\text{fit}} = a_1 e^{-t_i/a_2} + a_3 \tag{7.56}$$

where a_1 is proportional to the initial sample activity, a_2 is the mean lifetime, and a_3 is a constant background level arising from other sources. For this nonlinear fitting function, the derivative of y_i^{fit} with respect to a_1 depends on a_2 and the derivative with respect to a_2 depends on both a_1 and a_2.

The solution for the best-fit parameters must still satisfy Eq. 7.30, but because $[J_a^y]$ depends on the a_k, an iterative solution must be sought. The user provides initial guesses for the fitting parameters that will be used as a starting point. From the initial guesses, a nonlinear fitting program will locate other nearby parameter sets — evaluating the χ^2 each set produces. Each time the program finds that χ^2 has decreased, it uses those improved parameter values as a new starting point for the next iteration. Iterations continue until the solution to Eq. 7.30 is self-consistent — satisfied with $[J_a^y]$ (and perhaps $[\sigma_y^2]$) evaluated at the best fit.

Various algorithms can be used to search parameter space for the best-fit a_k that minimize the χ^2. Three commonly used are the *Gauss-Newton*, the *gradient-descent,* and a hybrid of the two — the *Levenberg-Marquardt.* The Gauss-Newton algorithm is the most efficient when the starting parameters are already sufficiently close to the best fit. Its big drawback is that it tends to fail if the starting parameters are not close enough. The gradient-descent algorithm is better at improving the fit parameters when their starting values are further from the best-fit values. Its big drawback is that it tends to take many iterations to find the best fit. The Levenberg-Marquardt algorithm elegantly addresses the shortcomings of the other two. In effect, it uses gradient-descent when Gauss-Newton fails, but switches to Gauss-Newton as the parameters approach the best fit.

Of course, all three algorithms require user input: the N values of the independent variable y_i and their variances σ_i^2, initial guesses for the M fitting parameters a_k, the N values of y_i^{fit} using those a_k, and the $N \times M$ Jacobian $[J_a^y]$. The y_i^{fit}, $[J_a^y]$ and perhaps $[\sigma_y^2]$ must be in the form of a spreadsheet formula or computer code for evaluation at any set of a_k. Chapter 10 shows how to specify this input for regression analysis in Excel.

The Gauss-Newton algorithm begins with the Jacobian of Eq. 7.31 evaluated at the starting parameters. If the starting parameters are near the best-fit values, the χ^2 will be near its true minimum and in that neighborhood will lie on an M-dimensional parabola. The Gauss-Newton algorithm uses the $[J_a^y]$ evaluated at the starting parameters to determine this parabola and then jumps directly to the predicted χ^2 minimum.

For a linear fitting function the parabolic shape is guaranteed — even when the starting parameters are far from the best fit. In effect, this is why linear regression formulas can find the best-fit parameters without iteration. For nonlinear fitting functions, the parabolic shape may only extend to a small neighborhood around the best-fit parameters. If the starting parameters are in that neighborhood, the Gauss-Newton algorithm would jump almost directly to the correct best-fit parameters in one try. However, if the starting parameters are too far from the best fit, the local derivatives may not predict where the true minimum in χ^2 will be. When the algorithm jumps to the predicted best-fit parameters, it may find the χ^2 has decreased, but not to its minimum. In that case, it can simply reiterate from there. However, if at any time it finds that the χ^2 has increased rather than decreased, it would then have no recourse to improve the fit.

Each iteration of the gradient-descent algorithm is guaranteed to improve

the fit (lower the χ^2) and so it can be used when the Gauss-Newton algorithm fails. It is based on the "χ^2 gradient" — a vector of derivatives $\partial\chi^2/\partial a_k$ with components giving the rate of change of χ^2 with changes in each parameter. Recall that χ^2 is at a minimum and these derivatives are all zero at the best fit. Conversely, if any derivatives are found to be nonzero,[1] the χ^2 is not at its minimum and the parameters are not at the best fit. The algorithm then simultaneously changes each parameter by an amount Δa_k proportional to the corresponding derivative but in the opposite direction — toward decreasing χ^2. That is,

$$\Delta a_k \;=\; -\kappa_k \frac{\partial \chi^2}{\partial a_k} \tag{7.57}$$

where the κ_k are (positive) constants of proportionality determining the size of the change. While this equation is guaranteed to move a_k in the direction of decreasing χ^2, κ_k can be made too large. The new a_k (after the change is applied) can overshoot the location of the χ^2 minimum and the χ^2 can increase instead of decrease. But if the proportionality constants κ_k are all made small enough, Eq. 7.57 guarantees the χ^2 will decrease. On the other hand, κ_k values that are too small will give Δa_k that are too small. The new a_k values will move only a small fraction of the way to the χ^2 minimum and many iterations will be required to locate it.

To get appropriately sized κ_k values, first note that the fitting parameters typically have different units of measure and thus the χ^2 derivatives will have different units. To make Eq. 7.57 dimensionally consistent, κ_k must have the units of a_k^2. This can be expressed by rewriting Eq. 7.57

$$\Delta a_k = -\kappa\, u_k^2 \,\frac{\partial \chi^2}{\partial a_k} \tag{7.58}$$

with the proportionality constants $\kappa_k = \kappa\, u_k^2$ now expressed as the product of a positive, unitless factor κ common to all parameters and a positive, parameter-specific, or relative, scale factor u_k^2 having the same units as a_k^2.

A poorly scaled set of u_k^2 values leads to a κ value smaller than necessary and a slow approach to the best fit. Consequently, a gradient-descent

[1]Starting from Eq. 7.15 and treating the a_k therein as a set of unoptimized trial parameters, it is easy to show that $\partial\chi^2/\partial a_k$ is just the kth component of the M-component vector $-2\,[J_a^y]^T\,[\sigma_y^2]^{-1}\left(\boldsymbol{y}-\boldsymbol{y}^{\text{trial}}\right)$, with $[J_a^y]$ and $\boldsymbol{y}^{\text{trial}}$ evaluated with those parameters.

algorithm typically adjusts κ and the u_k^2 as the iterations proceed. Good u_k^2 values can be determined from the local X-matrix as described shortly. In any case, the u_k^2 need only have the correct order of magnitude. If the χ^2 does not decrease after using the current κ and u_k^2, a gradient-descent algorithm will leave the current fit parameters unchanged, decrease κ by some factor, and try again. Sooner or later the Δa_k will all be small enough that χ^2 will decrease — leading to a new, improved set of parameters from which a new iteration can proceed.

The elegance of the Levenberg-Marquardt algorithm is in how it monitors the χ^2 during the iterations and smoothly switches between Gauss-Newton-like and gradient-descent-like algorithms. More details on the gradient-descent and the Levenberg-Marquardt algorithms will be provided after discussing the Gauss-Newton algorithm.

If the σ_i^2 depend on the best fit, iteratively reweighted least squares will be needed. One might recalculate the σ_i^2 after each successful χ^2 decrease, or only after finding the minimum χ^2 with the current set of σ_i^2. Because self-consistency must ultimately be achieved for both σ_i^2 and $[J_a^y]$, the choice is simply a matter of whether and how fast the solution converges.

The Gauss-Newton Algorithm

Regression formulas for the Gauss-Newton algorithm are essentially identical to their linear regression counterparts. And while the range of fit parameters where the algorithm can be applied may be limited, in most cases it will include the all-important region within a few standard deviations of the best fit.

The treatment will require distinguishing between the best-fit parameters a_k, $k = 1...M$ and another set — nearby, but otherwise arbitrary. This nearby "trial" solution will be labeled $a_k^{\rm trial}$, $k = 1...M$, and gives a trial fitting function

$$y_i^{\rm trial} = F_i(\{a^{\rm trial}\}) \tag{7.59}$$

This initial solution must be close enough to the best fit that, for all data points, a first-order Taylor series expansion about $y_i^{\rm trial}$ will accurately reproduce the best-fit $y_i^{\rm fit}$.

$$y_i^{\rm fit} = y_i^{\rm trial} + \sum_{k=1}^{M} \frac{\partial y_i^{\rm trial}}{\partial a_k^{\rm trial}} (a_k - a_k^{\rm trial}) \tag{7.60}$$

Where these expansions are accurate, the χ^2 surface is parabolic.

Differentiating Eq. 7.60 shows that the elements of the Jacobian $[J_a^y]_{ik} = \partial y_i^{\rm fit}/\partial a_k$ are

$$[J_a^y]_{ik} = \frac{\partial y_i^{\rm trial}}{\partial a_k^{\rm trial}} \tag{7.61}$$

That is, the first-order Taylor expansion implies the derivatives at the best fit are the derivatives at the nearby starting point.

For well-behaved fitting functions, the first-order Taylor expansion is guaranteed accurate for values of $a_k^{\rm trial}$ that are sufficiently close to the a_k. If the expansion remains valid for a wider range of $a_k^{\rm trial}$, the Gauss-Newton algorithm will likewise find the best-fit a_k from those more distant trial solutions. Moreover, the parameter covariance matrix relies on the first-order Taylor expansion and so its range of validity will have important implications for parameter uncertainties.

To see how the first-order Taylor expansion will lead to linear regression-like formulas, first define the modified input data Δy_i and the modified best-fit $\Delta y_i^{\rm fit}$ as relative to the trial solution.

$$\Delta y_i = y_i - y_i^{\rm trial} \tag{7.62}$$
$$\Delta y_i^{\rm fit} = y_i^{\rm fit} - y_i^{\rm trial} \tag{7.63}$$

Or, in vector notation

$$\Delta \boldsymbol{y} = \boldsymbol{y} - \boldsymbol{y}^{\rm trial} \tag{7.64}$$
$$\Delta \boldsymbol{y}^{\rm fit} = \boldsymbol{y}^{\rm fit} - \boldsymbol{y}^{\rm trial} \tag{7.65}$$

Subtracting $[J_a^y]^T \left[\sigma_y^2\right]^{-1} \boldsymbol{y}^{\rm trial}$ from both sides of the defining equation for the maximum likelihood solution, Eq. 7.30, then gives

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \Delta \boldsymbol{y} = [J_a^y]^T \left[\sigma_y^2\right]^{-1} \Delta \boldsymbol{y}^{\rm fit} \tag{7.66}$$

Next, define the modified best-fit parameters Δa_k as the difference between the actual best-fit parameters and the trial parameters.

$$\Delta a_k = a_k - a_k^{\rm trial} \tag{7.67}$$

or

$$\Delta \boldsymbol{a} = \boldsymbol{a} - \boldsymbol{a}^{\rm trial} \tag{7.68}$$

With these definitions, the first-order Taylor expansion (Eq. 7.60) can be written

$$\Delta y_i^{\text{fit}} = \sum_{k=1}^{M} [J_a^y]_{ik} \, \Delta a_k \tag{7.69}$$

or

$$\Delta \boldsymbol{y}^{\text{fit}} = [J_a^y] \, \Delta \boldsymbol{a} \tag{7.70}$$

Using Eq. 7.70 in Eq. 7.66 then gives the linear-regression-like result

$$[J_a^y]^T \left[\sigma_y^2\right]^{-1} \Delta \boldsymbol{y} = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y] \, \Delta \boldsymbol{a} \tag{7.71}$$

This equation is now in a form analogous to Eq. 7.39 with the solution for the best-fit Δa_k analogous to Eq. 7.43.

$$\Delta \boldsymbol{a} = \left[J_y^a\right]^\dagger \Delta \boldsymbol{y} \tag{7.72}$$

where $\left[J_y^a\right]^\dagger = [X]^{-1} [J_a^y]^T \left[\sigma_y^2\right]^{-1}$ and $[X] = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y]$ as evaluated at the trial solution.

After Eq. 7.72 is applied to determine the best-fit $\Delta \boldsymbol{a}$, Eq. 7.67 must then be applied to each element to find the best-fit a_k

$$a_k \;\; = \;\; a_k^{\text{trial}} + \Delta a_k \tag{7.73}$$

The resulting values for the a_k should then be used as new trial parameters a_k^{trial} for another iteration of the algorithm. y_i^{trial}, Δy_i, $[J_a^y]$, and if necessary, $\left[\sigma_y^2\right]^{-1}$ should be reevaluated there and the Gauss-Newton algorithm reiterated. Iterations can be stopped when there are no significant changes to the a_k, i.e., when $\Delta \boldsymbol{a} = 0$. Equation 7.71 with Eq. 7.64 shows that $\Delta \boldsymbol{a} = 0$ when $[J_a^y]^T \left[\sigma_y^2\right]^{-1} (\boldsymbol{y} - \boldsymbol{y}^{\text{trial}}) = 0$, which is just the condition for $\boldsymbol{y}^{\text{trial}}$ to be the best-fit solution, namely, Eq. 7.29.

Equation 7.49 then provides $[X]^{-1}$ as the covariance matrix for the Δa_k. Because of the constant offset transformation between Δa_k and a_k expressed by Eq. 7.73, propagation of error implies the a_k have the exact same covariance matrix.

The Gauss-Newton, gradient-descent, and Levenberg-Marquardt algorithms are demonstrated for simulated exponential decay in the two Excel workbooks *Nonlinear Regression.xlsm* and *Nonlinear Regression Poisson.xlsm* for Gaussian- and Poisson-distributed y_i, respectively. All three

algorithms are iterative. They differ in how they modify the X-matrix before taking its inverse and using it to calculate $\Delta \boldsymbol{a}$ (from Eq. 7.72) for the current iteration.

Leaving $[X]$ unmodified corresponds to the Gauss-Newton algorithm.

Zeroing out off-diagonal elements of $[X]$ and multiplying the diagonal elements by a single, programatically adjusted scale factor λ is a gradient-descent algorithm with

$$\kappa u_k^2 = \left(2\lambda \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(\frac{\partial y_i^{\text{fit}}}{\partial a_k} \right)^2 \right)^{-1} \tag{7.74}$$

The unitless scale factor λ common to all a_k starts at one, say, and is adjusted as follows. If the χ^2 fails to decrease, leave the a_k unchanged and increase λ by a factor of ten. This decreases all parameter step sizes for the next iteration and guarantees that with enough failed iterations, the steps will ultimately become small enough that the χ^2 will decrease. If the χ^2 successfully decreases, keep the new a_k and decrease λ by a factor of ten. Decreasing λ increases the parameter step sizes for the next iteration thereby helping to keep step sizes near the optimum values.

The Levenberg-Marquardt algorithm multiplies diagonal elements of $[X]$ by a programatically adjusted factor of $1+\lambda$ and leaves off-diagonal elements unchanged. As a consequence, this algorithm follows a near-gradient-descent algorithm with decreasing step sizes as λ increases above one and it approaches the Gauss-Newton algorithm as λ decreases below one. λ starts at one and is adjusted as for the gradient-descent algorithm. If the χ^2 fails to decrease, leave the a_k unchanged and increase λ by a factor of ten for the next iteration. If the χ^2 successfully decreases, keep the new a_k and decrease λ by a factor of ten.

The parameter covariance matrix $[\sigma_a^2] = [X]^{-1}$ (Eq. 7.49) should always be obtained using an unmodified X-matrix (Eq. 7.40) with the Jacobian $[J_a^y]$ and the input covariance matrix $[\sigma_y^2]$ evaluated at the best fit.

Uncertainties in Independent Variables

Up to now, only the dependent variables 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 weights provides a good statistical model.

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

$$y_i^{\text{fit}} = F(x_i; \{a\}) \tag{7.75}$$

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

$$y_i^{\text{fit}} = F(\mu_{xi}; \{a\}) + \frac{\partial F(x_i; \{a\})}{\partial x_i} (x_i - \mu_{xi}) \tag{7.76}$$

for any x_i in this range.

Under these conditions, propagation of error implies that random variations in x_i with a variance of σ_{xi}^2 would cause random variations in y_i^{fit} with a variance

$$\sigma_{y_i^{\text{fit}}}^2 = \left(\frac{\partial F(x_i; \{a\})}{\partial x_i} \right)^2 \sigma_{xi}^2 \tag{7.77}$$

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

$$\sigma_i^2 = \sigma_{yi}^2 + \left(\frac{\partial F(x_i; \{a\})}{\partial x_i} \right)^2 \sigma_{xi}^2 \tag{7.78}$$

where now the σ_{yi} are the standard deviations of the y_i (σ_i previously).

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. 7.78. These values will give the proper weighting matrix for the fit with the correct dependence on σ_{xi} and σ_{yi}. Most importantly, the adjusted σ_i^2 will give the correct covariance matrix for the fitting parameters and when used in the χ^2 of Eq. 7.15 will maintain its proper expectation value — a critical aspect of the chi-square test discussed in Chapter 9.

The best-fit parameters would need to be known in order to determine the σ_i^2 of Eq. 7.78; the σ_i^2 depend on the derivatives of the fitting function

with respect to x_i and these derivatives will depend on the a_k. In keeping with a key result of iteratively reweighted least squares, the σ_i^2 should be recalculated from Eq. 7.78 at the start of each iteration using the current set of a_k and kept fixed while a fit to find the next set of a_k is performed. The iterations should be repeated until a self-consistent solution is obtained.

Regression with Correlated y_i

Performing a fit to a set of y_i having a nondiagonal covariance matrix $[\sigma_y^2]$ is relatively simple if the joint probability distribution for the y_i is reasonably described by the correlated Gaussian of Eq. 4.24. In this case, the regression formulas already presented remain valid without modification. One need only substitute the nondiagonal 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. 4.24 (multiplied by -2 so it is to be minimized to maximize the probability) depends on the μ_i only via a χ^2 of the form

$$\chi^2 = (\boldsymbol{y} - \boldsymbol{\mu})^T \left[\sigma_y^2\right]^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{7.79}$$

To be a maximum likelihood solution, Eq. 7.79 must produce the minimum χ^2 when y_i^{fit} is used for μ_i.

$$\chi^2 = \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right)^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right) \tag{7.80}$$

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:

$$\begin{aligned} 0 &= \frac{\partial}{\partial a_k} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right)^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right) \\ &= -\left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right)^T \left[\sigma_y^2\right]^{-1} \frac{\partial \boldsymbol{y}^{\text{fit}}}{\partial a_k} - \frac{\partial \boldsymbol{y}^{\text{fit}\,T}}{\partial a_k} \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right) \end{aligned} \tag{7.81}$$

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

$$\frac{\partial \boldsymbol{y}^{\text{fit}\,T}}{\partial a_k} \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\text{fit}}\right) = 0 \tag{7.82}$$

This scalar equation must be true for each of the M fitting parameters a_k and with the definition of the Jacobian (Eq. 7.31), all M equations can be rewritten in the vector form of Eq. 7.29. Because Eq. 7.29 was the starting point for the regression results already presented, and because its solution does not rely on $[\sigma_y^2]$ being diagonal, the equations for the best-fit parameters and their covariance matrix do not change when $[\sigma_y^2]$ is nondiagonal.

Calibrations and Instrument Constants

Systematic errors are common when using scientific instruments. Reducing them typically involves a calibration. As an example, consider a grating spectrometer used to determine visible wavelengths from measured diffraction angles where spectral features such as emission lines from a discharge tube are observed. Optics, diffraction theory, and spectrometer construction details predict the wavelengths (the dependent variables y_i) based on the diffraction angles (the independent variables x_i) and a set of apparatus parameters (the instrument constants b_j). For a spectrometer, the grating line spacing and an incidence angle might be instrument constants.

Consider a second example. Many instruments involve electronic circuits which are susceptible to offset and gain errors. It would then be appropriate to assume that instrument readings will suffer such effects and that corrected y_i should be obtained from raw instrument readings x_i according to $y_i = b_1 + b_2 x_i$, where b_1 is the offset error and the deviation of b_2 from unity is the gain error. Based on a factory calibration, a voltmeter offset and gain might be specified as: $b_1 = 0 \pm 1$ mV and $b_2 = 1 \pm 0.001$. The best estimates $b_1 = 0$ and $b_2 = 1$ result in $y_i = x_i$, i.e., there are no corrections to the raw readings. However, the uncertainties in b_1 and b_2 acknowledge that every y_i may be systematically offset by a few mV and systematically off in scale by a few parts per thousand.

The experimental model is that the instrument (or calibration) constants do not change during the acquisition of a measurement set, but their true values are subject to some uncertainty. Making sure the instrument constants do not change significantly is an important experimental consideration. For example, temperature affects all kinds of instrumentation and is often associated with gain and offset errors. That's why measurements should always be made after the electronics have warmed up and why swings in ambient temperature should be avoided.

A calibration function for a data set will be expressed

$$y_i = G(x_i; \{b\}) \tag{7.83}$$

With a given set of b_j, $j = 1...L$, the calibration function transforms each raw instrument reading x_i into an "instrument-measured" y_i of some physical significance.

In some cases, calibrations involve careful investigation of the instrument resulting in a statistically independent determination of each b_j and its uncertainty. In other cases, the b_j and their covariance matrix are determined by using the instrument to measure *standards*, i.e., samples or sources for which the y_i are already known. With the corresponding x_i measured for a set of known y_i, the (x_i, y_i) data points are then fit to the calibration equation treating it as a fitting function for the b_j.

$$y_i^{\text{fit}} = G(x_i; \{b\}) \tag{7.84}$$

A regression analysis then determines the best-fit b_j and their covariance matrix $[\sigma_b^2]$. This analysis involves uncertainty in the x_i and, as discussed previously, estimates of σ_{xi} will be needed and Eq. 7.78 would be used for determining the σ_i^2 for the fit. Furthermore, the reference y_i values for this analysis are often highly accurate and their associated σ_{yi} may be small enough to neglect in Eq. 7.78.

In essence, a calibration is simply a determination of the best estimates b_j and their standard deviations (if independent) or, more generally, their $L \times L$ covariance matrix $[\sigma_b^2]$. The calibration constants should be considered a single sample from some joint probability distribution having means given by the true values $\{\beta\}$ and having a covariance matrix $[\sigma_b^2]$.

With the b_j and $[\sigma_b^2]$ in hand, the instrument is then calibrated and ready for use in the main investigation — one where the y-values are not known in advance. For example, wavelengths measured with a spectrometer are used in a wide range of studies associated with spectral sources and excitation conditions. To this end, a new set of x_i are measured and used with the calibration constants in Eq. 7.83 to determine a new set of y_i-values. These y_i will now be considered "measured" y-values to be used as the dependent variables in a some new regression analysis associated with that study. How should the uncertainty in the instrument constants, as represented by $[\sigma_b^2]$, be treated and how will it affect the uncertainty in the fitting parameters of the new, or main, analysis?

The main regression analysis begins with the assumption that the b_j are exact. The input y_i must first be obtained from $y_i = G(x_i; \{b\})$ where the x_i are measured, i.e., random variables. Independent variations in the x_i of variance σ_{xi}^2 propagate to independent variations in the y_i of variance

$$\sigma_i^2 = \left(\frac{\partial y_i}{\partial x_i}\right)^2 \sigma_{xi}^2 \tag{7.85}$$

In matrix notation, the covariance matrix becomes

$$[\sigma_y^2] = [J_x^y][\sigma_x^2][J_x^y]^T \tag{7.86}$$

where the $N \times N$ Jacobian $[J_x^y]$ has only diagonal elements

$$[J_x^y]_{ii} = \frac{\partial G(x_i; \{b\})}{\partial x_i} \tag{7.87}$$

Off-diagonal elements are all zero because of the one-to-one relationship between the x_i and y_i in Eq. 7.83. The diagonality of $[J_x^y]$ implies $[\sigma_y^2]$ of Eq. 7.86 will likewise be diagonal (and the y_i can be treated as statically independent) if, as is usually the case, the x_i are statistically independent.

The main regression analysis is then performed with the $[\sigma_y^2]$ of Eq. 7.86, but is otherwise an ordinary regression analysis. It determines the M best-fit parameters a_k to the main fitting function $y_i^{\text{fit}} = F_i(\{a\})$ and it determines their covariance matrix $[\sigma_a^2]$. All previous regression results apply. The effects of $[\sigma_b^2]$ are determined only after this solution is obtained.

In the linear regime, a first-order Taylor expansion gives the small changes to the y_i that can be expected from small changes to the b_j. And of course, small changes to the y_i lead to small changes in the a_k. Once these first-order Taylor expansions are specified, propagation of error can be used to determine the contribution to $[\sigma_a^2]$ arising from $[\sigma_b^2]$. This new contribution will be in addition to the $[\sigma_a^2]$ determined from the main analysis and is readily predicted as follows.

Changes in the y_i due to small changes in the b_j are assumed to be well described by a first-order Taylor expansion about their means. In linear algebra form it is simply

$$\Delta \boldsymbol{y} = [J_b^y]\Delta \boldsymbol{b} \tag{7.88}$$

where

$$[J_b^y]_{ij} = \frac{\partial G(x_i; \{b\})}{\partial b_j} \tag{7.89}$$

Equation 7.72 ($\Delta\boldsymbol{a} = \left[J_y^a\right]^\dagger \Delta\boldsymbol{y}$) gives the changes in the main fitting parameters a_k that would occur with changes to the y_i. Combining this with Eq. 7.88 then gives the first-order Taylor expansion for the a_k as a function of the b_j.

$$\Delta\boldsymbol{a} = \left[J_y^a\right]^\dagger [J_b^y] \Delta\boldsymbol{b} \tag{7.90}$$

Propagation of error Eq. 6.16 then gives the covariance matrix for the a_k in terms of that for the b_j

$$\begin{aligned}
[\sigma_a^{2(b)}] &= \left\langle \Delta\boldsymbol{a}\Delta\boldsymbol{a}^T \right\rangle \\
&= \left\langle \left[J_y^a\right]^\dagger [J_b^y] \Delta\boldsymbol{b}\Delta\boldsymbol{b}^T [J_b^y]^T \left[J_y^a\right]^{\dagger T} \right\rangle \\
&= \left[J_y^a\right]^\dagger [J_b^y] \left\langle \Delta\boldsymbol{b}\Delta\boldsymbol{b}^T \right\rangle [J_b^y]^T \left[J_y^a\right]^{\dagger T} \\
&= \left[J_y^a\right]^\dagger [J_b^y] [\sigma_b^2] [J_b^y]^T \left[J_y^a\right]^{\dagger T}
\end{aligned} \tag{7.91}$$

where the (b) in $[\sigma_a^{2(b)}]$ indicates that Eq. 7.91 is the contribution to $[\sigma_a^2]$ arising from $[\sigma_b^2]$ and must be added to that due to $[\sigma_x^2]$. The contribution due to $[\sigma_x^2]$ is obtained from Eq. 7.48 (with Eq. 7.86 for $[\sigma_y^2]$) giving

$$\begin{aligned}
[\sigma_a^{2(x)}] &= \left[J_y^a\right]^\dagger [\sigma_y^2] \left[J_y^a\right]^{\dagger T} \\
&= \left[J_y^a\right]^\dagger [J_x^y] [\sigma_x^2] [J_x^y]^T \left[J_y^a\right]^{\dagger T}
\end{aligned} \tag{7.92}$$

The total covariance matrix is the sum of Eqs. 7.91 and 7.92.

$$\begin{aligned}
[\sigma_a^2] &= [\sigma_a^{2(b)}] + [\sigma_a^{2(x)}] \\
&= \left[J_y^a\right]^\dagger \left([J_b^y] [\sigma_b^2] [J_b^y]^T + [J_x^y] [\sigma_x^2] [J_x^y]^T \right) \left[J_y^a\right]^{\dagger T}
\end{aligned} \tag{7.93}$$

Note that the term in parentheses is the final covariance matrix for the y_i.

$$[\sigma_y^2] = [J_b^y] [\sigma_b^2] [J_b^y]^T + [J_x^y] [\sigma_x^2]] [J_x^y]^T \tag{7.94}$$

which is simply the propagation of error formula applied to $y_i = G(x_i; \{b\})$ assuming the main fit x_i and the calibration b_j are statistically independent. Because changes in the instrument constants propagate through all y_i, the contribution to $[\sigma_y^2]$ from $[\sigma_b^2]$ will lead to correlated y_i even if the x_i are uncorrelated.

While Eq. 7.94 is the final covariance matrix for the y_i, recall that only the $[\sigma_y^2]$ due to $[\sigma_x^2]$ (Eq. 7.86) should be used in the main regression analysis — including the calculation of the χ^2. This is because there is only one set

of b_j determining all y_i.[2] Consequently, deviations in the b_j from their true values do not add random scatter to the y_i. They propagate to systematic deviations in the y_i and then to systematic deviations in the a_k. Moreover, because the a_k and the χ^2 depend on the $[\sigma_y^2]$ used for the main fit, if the wrong $[\sigma_y^2]$ is used, the a_k and χ^2 would be wrong as well.

Transforming the Dependent Variables

This chapter finishes with a simple but common case where the dependent variables for the main regression analysis are not the measured y_i directly, but are, instead, calculated from the y_i. For example, one might fit to dependent variables $z_i = 1/y_i$ or $z_i = \ln y_i$ or some arbitrary function:

$$z_i = H(y_i) \tag{7.95}$$

where $H(y_i)$ contains no random variables except y_i.

If $H(y_i)$ is invertible, one could simply adjust the fitting function to predict y_i rather than z_i. But if the deviations in the y_i are small enough, this situation can be easily handled using the transformed z_i as the dependent variable and getting $[\sigma_z^2]$ for the fit using propagation of error. In matrix form:

$$[\sigma_z^2] = [J_y^z][\sigma_y^2][J_y^z]^T \tag{7.96}$$

where $[J_y^z]$ is diagonal with elements

$$[J_y^z]_{ii} = \frac{\partial H(y_i)}{\partial y_i} \tag{7.97}$$

The parameter covariance matrix (Eq. 7.48) becomes

$$[\sigma_a^2] = [J_z^a]^\dagger[\sigma_z^2][J_z^a]^{\dagger T} \tag{7.98}$$

where $[J_z^a]^\dagger$ is analogous to Eq. 7.44, but with the z_i as the dependent variables. Substituting Eq. 7.96 into the equation above then gives

$$[\sigma_a^2] \;=\; [J_z^a]^\dagger[J_y^z][\sigma_y^2][J_y^z]^T[J_z^a]^{\dagger T} \tag{7.99}$$

[2] One could imagine an odd scenario where a new calibration is performed and a new set of b_j are determined before each y_i is determined. In this case, Eq. 7.94 would be the correct $[\sigma_y^2]$ to use in the main fit and in the calculation of the χ^2.

If the y_i are obtained from a calibration function: $y_i = G(x_i; \{b\})$, $[\sigma_y^2]$ is then given by Eq. 7.94 and using this in Eq. 7.96 gives the covariance matrix for the z_i.

$$[\sigma_z^2] = [J_y^z] \left([J_x^y][\sigma_x^2][J_x^y]^T + [J_b^y][\sigma_b^2][J_b^y]^T \right) [J_y^z]^T \tag{7.100}$$

As discussed in the prior section, only the first term should be used in the regression analysis and for calculating the χ^2.

Using Eq. 7.100 in Eq. 7.98 then gives the parameter covariance matrix

$$[\sigma_a^2] \;=\; [J_z^a]^\dagger [J_y^z] \left([J_x^y][\sigma_x^2][J_x^y]^T + [J_b^y][\sigma_b^2][J_b^y]^T \right) [J_y^z]^T [J_z^a]^{\dagger T} \tag{7.101}$$

with the first and second terms providing the random and systematic error, respectively.

Chapter 8

Central Limit Theorem

The central limit theorem (CLT) states that the sum or mean of a sample set from an arbitrary distribution follows a Gaussian distribution more and more closely as the sample size increases. In this chapter two Monte Carlo simulations are presented. One shows the CLT for samples from a uniform distribution. The other shows how the CLT might apply to procedures such as regression analysis which produce multiple output parameters.

How closely the output parameters will follow a Gaussian distribution depends not only on the number of contributing variables, but also the shapes of their distributions and how each variable contributes to the results. Keep in mind that the distribution for the results may have very nearly Gaussian probabilities within one or two standard deviations of the mean, but have significant differences further from the mean. It typically takes more variables to get agreement with a Gaussian in the tails of the distribution.

Going hand-in-hand with the central limit theorem are procedures such as propagation of error and regression analysis which provide the covariance matrix for the output variables given the covariance matrix for the input variables. While both procedures require the validity of a first-order Taylor expansion, this condition is either guaranteed when the relationships between the input and output variables are linear, or it can be assured for nonlinear relationships by keeping input errors small enough.

For the simulations discussed here, the output variables will be linear functions of the input variables. Because the first-order Taylor expansion will be exact, the calculated covariance matrix will be exact for any size measurement errors having any distribution and whether the distribution for the output has converged to a Gaussian or not.

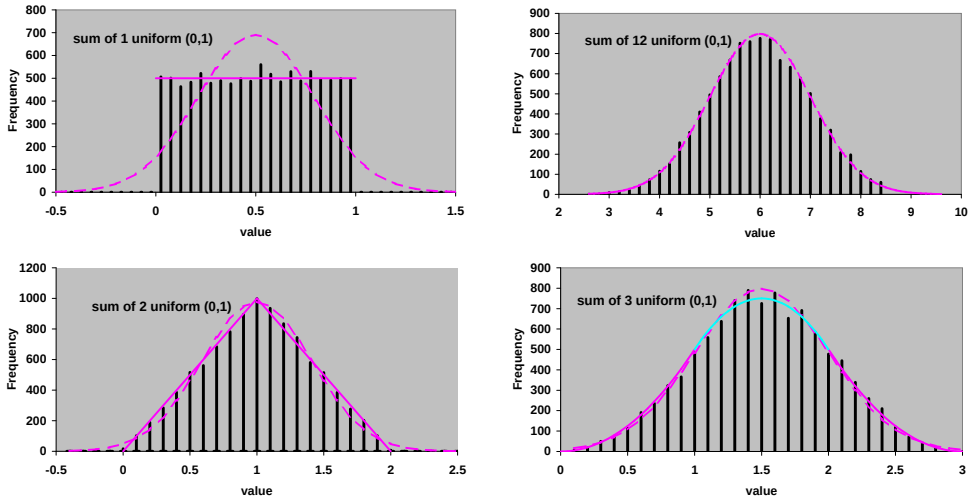


Figure 8.1: Counterclockwise from top left: Sample frequency distributions for 10,000 samples of the sum of 1, 2, 3 and 12 uniformly-distributed random numbers. The solid lines in the first three histograms are based on the true pdfs. The dotted lines in all four histograms are based on Gaussian pdfs with that histogram's true mean and variance.

Single Variable Central Limit Theorem

A simple demonstration of the central limit theorem involves a single output variable Y determined as the sum of N independent random variables y_i.

$$Y = \sum_{i=1}^{N} y_i \tag{8.1}$$

where each y_i comes from the same distribution with a mean μ_y and variance σ_y^2 but is otherwise arbitrary. Propagation of error then gives the mean of the sum (Eq. 6.8) $\mu_Y = N\mu_y$ and it gives the variance of the sum (Eq. 6.22) $\sigma_Y^2 = N\sigma_y^2$.

For a single output variable, there are no covariances to consider and according to the central limit theorem, the probability distribution in the limit of large N is a Gaussian distribution having the form of Eq. 3.1 with a mean and variance as given above.

Simulations of this kind are demonstrated in Fig. 8.1 where four sample frequency distributions are shown. Counterclockwise from top left, 10,000

samples for each histogram were created as the sum of $N = 1$, 2, 3, or 12 random numbers uniformly distributed on the interval from $[0, 1]$. This uniform distribution has a mean of 1/2 and a variance of 1/12 and so propagation of error implies the resulting Y-distributions have means of 0.5, 1.0, 1.5, and 6, and variances of 1/12, 2/12, 3/12, and 1, respectively.

The top-left graph of Fig. 8.1 includes a solid line giving the expected frequency distribution for samples from a uniform parent distribution ($N = 1$). The bottom-left ($N = 2$) and bottom-right ($N = 3$) graphs include solid lines giving the expected frequency distributions for samples from parent distributions obtained by one or two convolutions according to Eq. 6.7. All four cases include a dashed curve giving the expected frequency distribution for samples from a Gaussian distribution with a mean and variance equal to that of the true distribution.

Note how the approach to a Gaussian shape is becoming evident after summing as few as three uniform random variables. But keep in mind that the true distribution for this three-variable sum extends only from 0 to 3 and has zero probability density outside this ±3-sigma range, whereas a true Gaussian random variable would have around 0.3% probability of being there. Note that the $N = 12$ frequency distribution (top-right in Fig. 8.1) agrees very well the Gaussian, dashed-line prediction and demonstrates that the 12-sample sum has become nearly Gaussian-distributed as predicted by the central limit theorem.

Multivariable Central Limit Theorem

The simulations for multiple output variables will come from a common regression analysis task — fitting a slope m and intercept b to N data points (x_i, y_i), $i = 1...N$, all of which are predicted to lie along a straight line: $y = mx + b$.

The simulated data are generated according to $y_i = 3x_i + 15 +$ error. That is, for some fixed set of x_i, the y_i are created as the sum of a true mean $\mu_i = 3x_i + 15$ and a zero-mean random variable — the measurement error — created using a random number generator.

To demonstrate the central limit theorem, the random error for the y_i are all drawn from a uniform distribution in the interval from -6 to 6, which has a mean of zero and a variance $\sigma_y^2 = 12$. The equally-weighted regression formula, Eq. 7.54, is used to get the least-squares estimates for m and b for

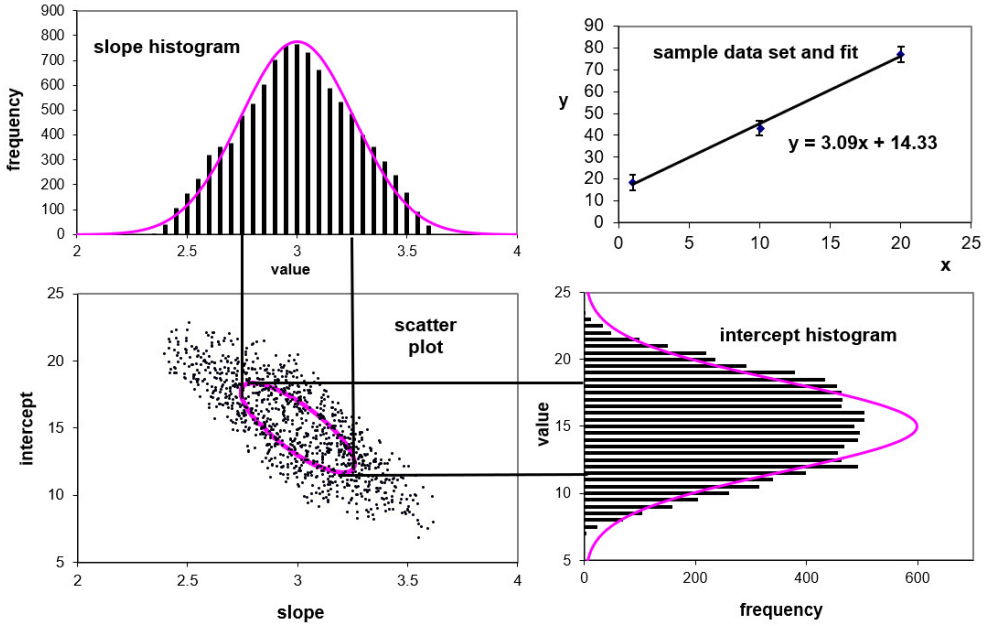


Figure 8.2: 3-point data sets. Top right: one sample data set with its fitted line. Top left and bottom right: frequency distributions for the slopes and intercepts. Bottom left: scatter plot for the slopes and intercepts.

each data set. With $\sigma_y^2 = 12$, Eq. 7.53 will give the true parameter covariance matrix in all cases.

Figures 8.2 and 8.3 summarize the results from two simulations. For each simulation, 10,000 data sets were created and fit to a straight line—producing 10,000 pairs of (m, b) values, one for each set. Because of the added random error, the fitted slopes and intercepts deviate from the true values of 3 and 15, respectively. Because each data set has different random errors, each set produces a different pair; m and b are a random variable pair.

Each figure shows one sample data set (top right) as well as a scatter plot (1500 points, bottom left) and frequency distributions for both m (top left) and b (bottom right) for all 10,000 data sets. For Fig. 8.2 every data set has three data points (for x-values of 1, 10, and 20). Data sets for Fig. 8.3 have 20 data points (for integer x-values from 1 to 20).

Qualitatively, the negative correlation indicated in both scatter plots can be understood by considering the top right graphs in the figures showing a

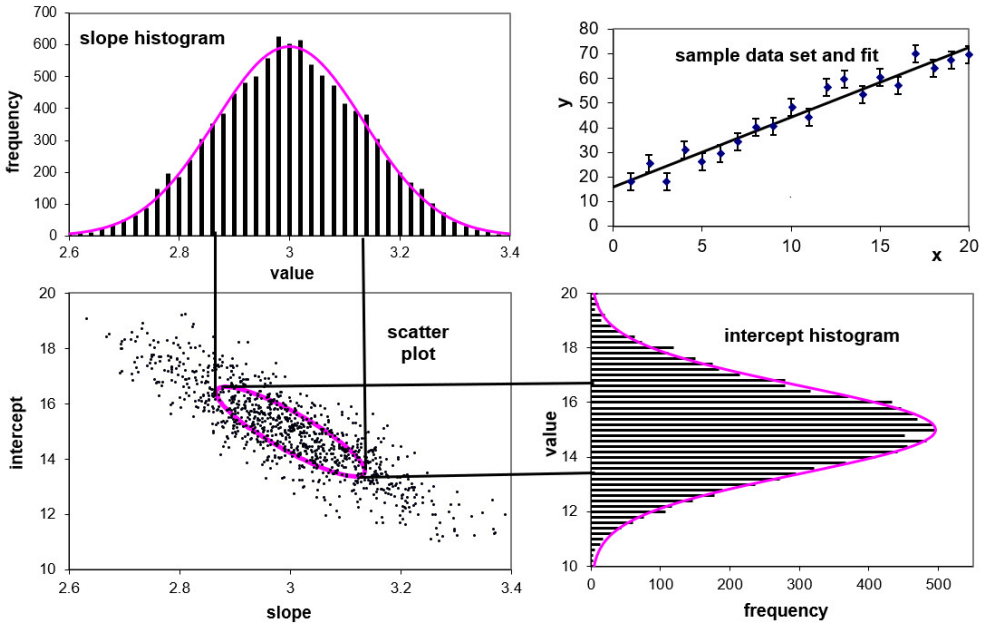


Figure 8.3: 20-point data sets. Top right: one sample data set with its fitted line. Top left and bottom right: frequency distributions for the slopes and intercepts. Bottom left: scatter plot for the slopes and intercepts. Solid lines provide Gaussian predictions.

typical data set and fitted line. Simultaneously high values for both the slope and intercept would produce a fitted line lying entirely above the true line with the deviation between them growing as x increases. Random errors are unlikely to reproduce this behavior by chance. A similar argument would apply when the slope and intercept are simultaneously low. However, with a high slope and a low intercept, or vice versa, the fitted line would cross the true line and the deviations between them would never get as large. Data sets showing this behavior are not as unlikely and lead to negative correlations as evidenced in the scatter plots.

In the histograms and scatter plots for Figs. 8.2 and 8.3, the solid lines are associated with predictions based on the assumption that m and b vary according to the correlated Gaussian distribution of Eq. 4.24 with the predicted covariance matrix of Eq. 7.53.

The solid vertical and horizontal guidelines are positioned $\pm\sigma_m$ and $\pm\sigma_b$ about the predicted means of 3 and 15, respectively, and would be expected

to include 68% of the points if m and b followed Gaussian distributions. For the 20-point data sets, this fraction was about 0.5% below the Gaussian prediction for both the slopes and intercepts. The 3-point data sets had a significantly lower fraction — around 63% for the intercepts and 65% for the slopes. For the 20-point data sets, the fraction of m- and b-values within 2-sigma of their means agreed well with the 95% prediction based on a Gaussian-distributed variable. For the 3-point data sets, the two-sigma fractions were a percent or two higher than the Gaussian prediction.

Demonstrating the sample and predicted m, b correlations, the ellipse in each scatter plot is an iso-probability contour assuming m and b are Gaussian-distributed with the predicted covariance matrix. All values for m and b on the ellipse would make the argument of the exponential in Eq. 4.24 equal to $-1/2$. That is, those m, b values would occur with a probability density that is down by a factor of $\exp(-1/2)$ from the peak density at $m = 3$ and $b = 15$. A two dimensional integration of a Gaussian joint pdf gives the probability enclosed by the ellipse and evaluates to $1 - \exp(-1/2) = 0.39$. Thus, if the results followed Gaussian predictions, 39% should fall inside the ellipse. The 20-point data sets had about 38% inside the ellipse, while the 3-point data sets had a significantly lower fraction — around 31%.

In conclusion, both the 3-point data sets and the 20-point data sets show non-Gaussian distributions for m and b. Compared to Gaussian expectations, the histograms for both the slope and intercept show fewer values in both the peak and tails of the distributions and more in the shoulders — likely a result of the same property of the uniform distribution used for the input noise. The difference is quite pronounced for the 3-point data sets. It is much more subtle for the 20-point data sets — indicating that with more data points, m and b follow a more Gaussian-like distribution as predicted by the central limit theorem.

Basically, the central limit theorem guarantees that if enough data go into determining a set of results and their covariance matrix, then the distribution for those results will be approximately Gaussian with that covariance matrix.

Chapter 9

Evaluating a Fit

Graphical Evaluation

Evaluating the agreement between a fitting function and a data set typically begins with a graph. The steps will be described for the common case of a fit to a single dependent variable y_i as a given function of a single independent variable x_i. The main graph should show the fitting function as a smooth curve without markers for a set of x-values that give a good representation of the best-fit curve throughout the fitting region. Plotting it only for the x_i of the data points may be insufficient if they are too widely spaced. It is also sometimes desirable to extrapolate the fit above or below the range of the data set x_i.

The input data points (x_i, y_i) should not be connected by lines. They should be marked with an appropriately sized symbol and 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. Alternatively, the x-uncertainties can be folded into a σ_i calculated according to Eq. 7.78, which would then be used for vertical error bars only.

Figure 9.1 shows a case where the error bars are too small to show clearly on the main graph. The fix is shown below the main graph — a plot of the *residuals* $y_i - y_i^{\text{fit}}$ vs. x_i 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 residuals $(y_i - y_i^{\text{fit}})/\sigma_i$ (without error bars) could be used.

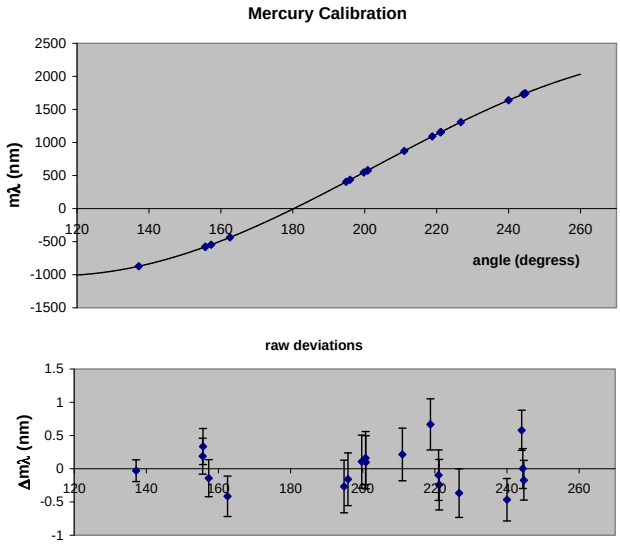


Figure 9.1: Top: main graph for a fit to a calibration function for a visible spectrometer. Bottom: corresponding residual plot.

The purpose of these graphs is to make it easy to see each data point's residual relative to its standard deviation and to assess the entire set of residuals for their expected randomness.

Specifically look for the following fitting problems and possible causes.

- Residuals seem nonrandom 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. Residuals should be random. In particular, positive and negative residuals are equally likely and should be present in roughly equal numbers. Trends in residuals may indicate a systematic error, a problem with the fitting program, or an incorrect fitting model.

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

- There are outliers. Outliers are 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. This behavior is not all that unlikely if there are only a few data points, but with larger data sets it indicates the σ_i have been overestimated — the measurements are more precise than expected. 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.

The χ^2 Distribution

The χ^2 appearing in Eq. 7.15 will be called the "best-fit" chi-square to distinguish it from another version — the "true" chi-square that uses the true theory parameters α_k and thus the true means μ_i.

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \tag{9.1}$$

The chi-squares of Eq. 7.15 and 9.1 are different random variables with different distributions.

Equations 9.1 and 7.15 can also be used, respectively, to define true and best-fit versions of the chi-square random variable for Poisson- and binomial-distributed y_i. For Poisson y_i, simply use $\sigma_i^2 = \mu_i$ for the true χ^2 and $\sigma_i^2 = y_i^{\text{fit}}$ for the best-fit χ^2. For binomial y_i, use $\sigma_i^2 = \mu_i(1 - \mu_i/\mathcal{N}_i)$ and $\sigma_i^2 = y_i^{\text{fit}}(1 - y_i^{\text{fit}}/\mathcal{N}_i)$, respectively, for the true and best-fit χ^2. Except in simulations the true parameter values are not normally known and thus the true means and the true chi-square cannot be determined. Hence, *the* χ^2, without a qualifier, should be understood to be the best-fit value.

The χ^2 is commonly used as a "goodness of fit" statistic. If the σ_i and the fitting function are correct, likely values for the χ^2 variable can be quite predictable. An unlikely, i.e., "bad," χ^2 value is an indication that theory and measurements are incompatible. To determine whether or not a particular χ^2 value is reasonable, the χ^2 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 is a major factor in determining the distribution. Each data point adds one to the degrees of freedom and each fitting parameter subtracts one.

Evaluating the expectation value or mean of the χ^2 distribution begins with the definition, Eq. 2.18, for the variance of each y_i — rewritten in the form

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

Summing over all N data points gives

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

The sum in this expression is the true χ^2 of Eq. 9.1. Thus, the mean of the distribution for the true χ^2 is equal to the number of data points N.

How does the mean of the χ^2 distribution change when χ^2 is evaluated with Eq. 7.15 using the y_i^{fit} in place of the μ_i? It turns out that the best-fit χ^2 satisfies

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

The mean of the best-fit χ^2 is equal to the number of degrees of freedom, i.e., smaller than that of the true χ^2 by the number of fitting parameters M.

The proof of Eq. 9.4 is given in the *Regression Analysis Addendum.* In fact, in Exercise 5 you demonstrated Eq. 9.4 for the special case of a fit to the constant function: ($y_i^{\text{fit}} = \bar{y}$, $M = 1$) with equally weighted y-values ($\sigma_i = \sigma_y$). The proof is based on Eq. 2.18 and thus valid for any distribution for the y_i, but it does require that the y_i^{fit} follow a first-order Taylor expansion in the a_k over the range of likely fitting parameters.

For fixed σ_i it is easy to see why some reduction in the χ^2 should always be expected when y_i^{fit} replaces μ_i. After all, the fitted a_k and their corresponding y_i^{fit} are specifically chosen to produce the lowest possible χ^2 for one particular data set — lower even than would be obtained (for that data set) with the true α_k and their corresponding μ_i. Thus, for any data set, the χ^2 using y_i^{fit} can only be equal to or less than the χ^2 using the μ_i. The average decrease is M, but the actual decrease can be smaller or larger for any particular data set.

According to Eq. 9.4, the mean of the chi-square distribution is $N - M$. How far above (or below) the 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.

The variances of the best-fit and true χ^2 distributions depend on the probability distribution for the y_i. For the true χ^2, the variance is readily predicted starting from Eq. 2.19.

$$\begin{aligned}
\sigma^2_{\chi^2} &= \left\langle \left(\chi^2\right)^2 \right\rangle - \left(\left\langle \chi^2 \right\rangle\right)^2 \\
&= \left\langle \left(\sum_{i=1}^{N} \frac{\left(y_i - \mu_i\right)^2}{\sigma_i^2} \right) \left(\sum_{j=1}^{N} \frac{\left(y_j - \mu_j\right)^2}{\sigma_j^2} \right) \right\rangle - N^2 \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \left(\frac{\left(y_i - \mu_i\right)^2}{\sigma_i^2} \right) \left(\frac{\left(y_j - \mu_j\right)^2}{\sigma_j^2} \right) \right\rangle - N^2 \qquad (9.5)
\end{aligned}$$

where Eq. 9.3 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. 4.6 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 (each of which is unity by Eq. 9.2). 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. 9.5 gives

$$\sigma^2_{\chi^2} = \sum_{i=1}^{N} \left\langle \frac{\left(y_i - \mu_i\right)^4}{\sigma_i^4} \right\rangle - N \qquad (9.6)$$

The quantity

$$\beta_i = \left\langle \frac{\left(y_i - \mu_i\right)^4}{\sigma_i^4} \right\rangle \qquad (9.7)$$

is called the *normalized kurtosis* and is a measure of the peakedness of the distribution for y_i. The more probability density further from the mean (in units of the standard deviation), the higher the kurtosis. The expectation values determining β_i were evaluated for several probability distributions using Eq. 2.9 or 2.10. The results follow.

For y_i governed by a Gaussian distribution, the result is $\beta_i = 3$ and using this in Eq. 9.6 gives

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

For y_i governed by a Poisson distribution, $\beta_i = 3 + 1/\mu_i$ giving

$$\sigma^2_{\chi^2} = 2N + \sum_{i=1}^{N} \frac{1}{\mu_i} \qquad (9.9)$$

For y_i governed by a binomial distribution, $\beta_i = 3 + \mathcal{N}_i/\mu_i(\mathcal{N}_i - \mu_i) - 6/\mathcal{N}_i$ giving

$$\sigma^2_{\chi^2} = 2N + \sum_{i=1}^{N} \left[\frac{\mathcal{N}_i}{\mu_i(\mathcal{N}_i - \mu_i)} - \frac{6}{\mathcal{N}_i} \right] \tag{9.10}$$

For y_i governed by a uniform distribution, $\beta_i = 1.8$ giving

$$\sigma^2_{\chi^2} = 0.8N \tag{9.11}$$

Equations 9.8-9.11 give four different values for the variance of the true χ^2 distribution for y_i that follow four different distributions — Gaussian, binomial, Poisson, and uniform. Equation 9.3 gives the common value for the mean in all four cases. While the mean and variance do not provide a complete description of these chi-square distributions, they are a good start. Tabulations and other information about the χ^2 distribution found in textbooks and statistics software typically apply only to Gaussian-distributed y_i.

As to the distribution for the best-fit χ^2, only for Gaussian-distributed y_i is it well known. It is the textbook chi-square distribution with $N - M$ degrees of freedom; it has a mean of $N - M$ and a variance of $2(N - M)$. Simple Monte Carlo simulations[1] with Gaussian-distributed y_i confirmed the predicted means and variances for both the true and best-fit χ^2-distributions.

The same simulations, but with Poisson, binomial, and uniformly-distributed y_i confirmed the predicted means for the true and best-fit χ^2 as well as Eqs. 9.9-9.11 for the true χ^2 variances. As to the best-fit χ^2 variance, the simulations demonstrated that it moved toward better agreement with the textbook χ^2 distribution. For Poisson y_i (with $\mu_y = 10$), for example, where the true χ^2 variance is larger than it is for Gaussian y_i, the best-fit χ^2 variance decreased from the true χ^2 variance by more than it did for Gaussian y_i. In contrast, for uniformly-distributed y_i, where the true χ^2 variance is considerably less than it is for Gaussian y_i, the best-fit χ^2 variance actually exceeded the true χ^2 variance.

Textbook chi-square distributions will be assumed appropriate in the following discussions, but if evaluating χ^2 probabilities is an important aspect

[1] Sample sets of y_i of size $N = 20$ were randomly generated from a common distribution with a given mean μ_y and variance σ^2_y. The sample mean $\bar{y}$, the true χ^2 and the best-fit χ^2 were calculated for each of 20,000 such sets to get precise estimates of the means and variances of these χ^2 variables.

of the analysis, keep this assumption in mind. For example, the noteworthy chi-square test, discussed next, would depend on the distribution for the y_i.

The χ^2 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. While reporting the best-fit χ^2 should be standard practice when describing experimental results, the test itself is no substitute for a graphical evaluation of a fit.

The following discussion applies to Gaussian-distributed y_i and thus the textbook χ^2 distributions.

Consider a fit to a data set with $N - M = 50$ degrees of freedom. The mean of the χ^2 distribution would be 50 and its standard deviation would be $\sigma_{\chi^2} = \sqrt{2(N-M)} = 10$. While the χ^2 distribution is not Gaussian, χ^2 values outside the two-sigma range from 30 to 70 should be cause for concern.

So suppose the actual χ^2 value from a fit is significantly above $N - M$ (indicating that the data are missing the fit by more than expected) and the analysis must decide if the χ^2 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.

The χ^2 may sometimes come out too small — well under the expected value of $N - M$ thereby suggesting that the data and fit may agree too well. To check if the χ^2 is too low, the chi-square distribution is 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. One caveat seems reasonable. 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 probability density is nonzero at $\chi^2 = 0$ and decreases monotonically as χ^2 increases. Thus, for these two cases, smaller χ^2 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.

Uncertainty in the Uncertainty

It is not uncommon to be unsure of the true σ_i associated with a set of measurements. If the σ_i are unknown, the χ^2 cannot be calculated and the chi-square test is unusable. More importantly, the σ_i determine the weighting matrix $\left[\sigma_y^2\right]^{-1}$ (Eq. 7.32) and with Eqs. 7.49 and 7.40, they also determine the parameter covariance matrix $[\sigma_a^2]$. If the σ_i are unknown or unreliable, the parameter uncertainties would be unknown or unreliable as well.

What can be done in this case? One accepted practice is to adjust the measurement uncertainties so that the χ^2 is equal to its expectation value of $N - M$. 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.

The technique is straightforward for an equally-weighted data set. Find that one value for σ_y that gives $\chi^2 = N - M$. Equation 7.55 shows this would happen if the following *sample variance of the fit* were used for σ_y^2:

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

Eq. 7.53 demonstrates that the parameter covariance matrix $[\sigma_a^2]$ — every element — would then be proportional to this sample variance.

Equation 9.12 is a generalized version of Eq. 2.23 for determining a sample variance. Equation 9.12 is the general case where the y_i^{fit} are based on the M best-fit parameters of some fitting function. Equation 2.23 can be considered a special case corresponding to a fit to the constant function, where all y_i^{fit} are the same and given by the single ($M = 1$) best-fit parameter $\bar{y}$.

Exercise 5 and Eq. 9.4 demonstrate that Eqs. 2.23 and 9.12 give a sample variance s_y^2 that is an unbiased estimate of the true variance σ_y^2. For an equally-weighted fit then, forcing $\chi^2 = N - M$ is the simple act of using that unbiased estimate for σ_y^2 in the formulas for the parameter covariance matrix.

If the σ_i are known to vary from point to point, forcing $\chi^2 = N - M$ would proceed a bit differently. The relative sizes of the σ_i must be known in advance so that forcing $\chi^2 = N - M$ determines a single overall scale factor for all of them. Relatively-sized σ_i might occur when measuring wide-ranging quantities with instruments having multiple scales or ranges. In such cases the measurement uncertainty might scale with the instrument range used for the measurement.

Initial estimates for the input σ_i would be set in accordance with the known ratios. The fit is performed and the χ^2 is evaluated. Then a single scale factor multiplying all σ_i would be determined to achieve a chi-square value of $N - M$. Scaling all the σ_i by the factor $\kappa = \sqrt{\chi^2/(N-M)}$ does this. It scales the present χ^2 by a factor of $1/\kappa^2$ (to $N - M$). The input covariance matrix $[\sigma_y^2]$ would scale by κ^2 and Eq. 7.49 with Eq. 7.40 implies that the parameter covariance matrix $[\sigma_a^2]$ would scale by κ^2 as well. On the other hand, Eq. 7.43 with the equations for $\left[J_y^a\right]^\dagger$ and $[X]$ show that the fit parameters are unaffected by the scale factor. There is no effect on the parameters because the scaling 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 χ^2 value for the fit. When the σ_i are adjusted to make $\chi^2 = N - M$, that randomness is transferred to the measurement uncertainties and to the parameter uncertainties. The covariance matrices $[\sigma_y^2]$ and $[\sigma_a^2]$ that result from forcing $\chi^2 = N - M$ become random variables. They become *sample covariance matrices* and should be written $[s_y^2]$ and $[s_a^2]$.

For example, parameter standard deviations obtained from the diagonal elements of $[s_a^2]$ become sample standard deviations. As discussed shortly, confidence intervals constructed with sample standard deviations differ somewhat from those constructed with true standard deviations.

Forcing $\chi^2 = N - M$ uses the random fit residuals to determine the s_i used as estimates for the true σ_i. Consequently, the technique is better suited for large data sets where the large number of residuals ensures a reasonable precision for those estimates.

To appreciate the sample size issue, consider an experiment with $N = M + 200$ data points all with equal uncertainty. The expectation value of χ^2 is then 200 with a standard deviation of 20. Suppose that an initial estimate of σ_y makes the the best-fit χ^2 come out 100. The chi-square distribution with 200 degrees of freedom shows that a value outside the range 200 ± 60 is less than 1% likely, so a value of 100 means something is amiss. Scaling the σ_y down by a factor of $\sqrt{2}$ will raise the χ^2 to 200 — the predicted average χ^2 for this case. Your confidence in the original σ_y is not all that high and scaling it that amount is deemed reasonable. The residuals are examined and seem random. This is a good case for forcing $\chi^2 = N - M$. The degrees of freedom are so high that using s_y for σ_y may give a better estimate of the true parameter covariance matrix than one obtained using an uncertain

estimate of σ_y based on other considerations.

Next consider an equally-weighted data set where $N - M = 5$ so that the expectation value of χ^2 is five. Suppose that using an initial estimate of σ_y, the best-fit chi-square comes out ten. The residuals are random and show no systematic deviations from the fit. This time, σ_y would have to be scaled up by a factor of $\sqrt{2}$ to bring the χ^2 down to five. Again, your confidence in σ_y is not high and you are not uncomfortable scaling them this much. But should you? With five degrees of freedom, a χ^2 value of 10 or more is not all that rare and is expected to occur with about an 8% probability. With fewer degrees of freedom, it is not as clear whether to abandon the original σ_y and force $\chi^2 = N - M$ instead.

When should the σ_i be scaled to give $\chi^2 = N - M$? The technique is appropriate whenever the σ_i are unknown or uncertain. Of course, the chi-square test is then unusable as a compatibility test between measurements and theory. After all, the measurement uncertainties have been adjusted specifically to achieve the "ideal" value of $\chi^2 = N-M$. A rough compatibility test would then rely on rough experimental estimates of the true σ_i. Scaling the σ_i by factor of two to achieve $\chi^2 = N - M$ might be deemed acceptable, but scaling them by a factor of ten might not. In general, one should be wary of accepting a scale factor much bigger than 3 or much smaller than 1/3. Even a coarse assessment of experimental variables should be able to distinguish measurement uncertainties at that level. Moreover, one must still examine residuals for systematic deviations. Masking mistakes, systematic errors or incorrect models by forcing $\chi^2 = N-M$ is obviously a poor practice.

The Reduced χ^2 Distribution

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

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

Dividing any random variable by a constant results in a new random variable with a mean divided down from that of the original by that constant and with a variance divided down from that of the original 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 of $2/(N - M)$.

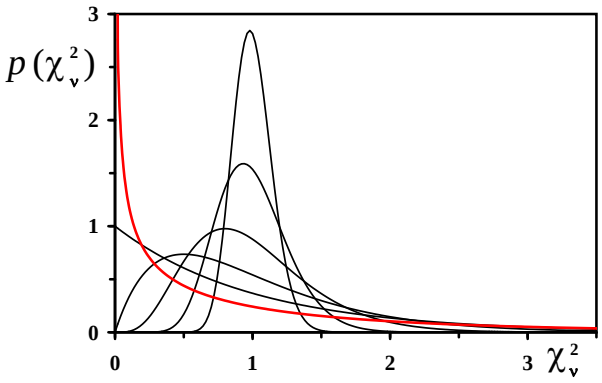


Figure 9.2: The reduced chi-square pdfs $p(\chi^2_\nu)$ for degrees of freedom (dof) $1, 2, 4, 10, 30, 100$. The tall distribution peaking at $\chi^2_\nu = 1$ is for dof = 100. The curves get broader and lower as the dof decrease. For dof = 1, the distribution (red) is singular at zero.

Reduced chi-square distributions for various degrees of freedom are shown in Fig. 9.2. Table 10.3 can be used to look up reduced chi-square probabilities for up to 200 degrees of freedom. The table can also be used for determining χ^2 probabilities using the scaling described 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.

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 $\sqrt{2(N-M)}$, and the latter with a mean of one and a standard deviation of $\sqrt{2/(N-M)}$. This approximation is used in the next exercise.

Dividing both sides of Eq. 9.12 by σ_y^2 and eliminating the sum using Eq. 7.55 gives

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

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

The fact that the χ^2_ν distributions narrow as the sample size N increases, makes sense with respect to the law of large numbers. The distribution

becomes more sharply peaked around its expectation value of one (where $s_y^2 = \sigma_y^2$) and so indicates that the sample variance s_y^2 becomes a more precise estimate of the true variance σ_y^2 as the sample size increases.

Exercise 6 *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 in $\bar{y}$, 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 number of degrees of freedom 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.*

Confidence Intervals

Consider a set of M fitting parameters a_k and their covariance matrix $[\sigma_a^2]$ obtained after proper data collection and analysis procedures. Recall that the set of a_k should be regarded as one sample governed by some joint probability distribution having the known covariance matrix and some unknown true means α_k, which are the ultimate targets of the experiment. What can be said about them?

The a_k and $[\sigma_a^2]$ can simply be reported, leaving the reader to draw conclusions about how big a deviation between the a_k and α_k would be considered reasonable. More commonly, the results are reported using *confidence intervals*. Based on the best-fit a_k and the $[\sigma_a^2]$, a confidence interval consists of a range (or interval) of possible parameter values and a probability (or confidence level) that the true mean will lie in that interval.

To create confidence intervals requires that the shape of the joint distribution be known. Its covariance matrix alone is insufficient. Often, confidence intervals are created assuming the variables are governed by a Gaussian joint distribution, but they can also take into account other information, such as a known non-Gaussian distribution for the a_k or uncertainty in the $[\sigma_a^2]$.

If a random variable y follows a Gaussian distribution, a y-value in the range $\mu_y \pm \sigma_y$ occurs with a 68% probability and a y-value in the range

$\mu_y \pm 2\sigma_y$ occurs with a 95% probability. 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 be 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.

Note that the discussion above is unmodified if y is one variable of a correlated set having the given σ_y^2 but otherwise arbitrary covariance matrix. For a Gaussian joint distribution, the unconditional probability distribution for y — obtained by integrating over all possible values for the other variables in the set — can be shown to be the single-variable Gaussian distribution with a variance of σ_y^2 and a mean μ_y. It does not depend on any of the other random variable values, variances, or covariances.

Occasionally, one might want to describe confidence intervals associated with multiple variables. For example, what is the confidence level that both of two means are in some stated range. If the variables are statistically independent, the probabilities for each variable are independent and the probability for both to be in range is simply the product of the probabilities for each to be in range. When the variables are correlated, the calculations are more involved and only one will be considered.

A constant-probability contour for two correlated Gaussian variables, say y_1 and y_2, governed by the joint distribution of Eq. 4.24, is an ellipse where the argument of the exponential in the distribution is some given constant. The ellipse can be described as the solution to

$$\chi^2 = \left(\boldsymbol{y}^T - \boldsymbol{\mu}^T\right) \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right) \tag{9.15}$$

where χ^2 is a constant. Eq. 4.24 shows that $e^{-\chi^2/2}$ would give the probability density along the ellipse as a fraction of its peak value at $\boldsymbol{y} = \boldsymbol{\mu}$ where the ellipse is centered. For example, on the $\chi^2 = 2$ ellipse the probability density is down by a factor of $1/e$.

The probability for (y_1, y_2) values to be inside this ellipse is the integral of the joint distribution over the area of the ellipse and is readily shown to be $1 - \exp(-\chi^2/2)$. For $\chi^2 = 1$ (akin to a one-sigma one-dimensional interval) the probability is about 39% and for $\chi^2 = 4$ (akin to a two-sigma interval) the probability is about 86%.

Reversing the logic, this same ellipse can be centered on the measured (y_1, y_2) rather than on the means (μ_1, μ_2). If the (y_1, y_2) value were inside the ellipse centered at (μ_1, μ_2), then (μ_1, μ_2) would be inside that same ellipse centered on (y_1, y_2). Since the former occurs with a probability of $1 - \exp(-\chi^2/2)$, this is also the probability (confidence level) for the latter.

Two-dimensional confidence ellipses do not change if the two variables involved are part of a larger set of correlated variables. One can show that integrating the joint Gaussian distribution over all possible values for the other variables leaves the two remaining variables described by a two-dimensional joint Gaussian distribution with the same means, variances, and covariance as in the complete distribution.

Student-T Probabilities

When the χ^2 is forced to $N - M$, the parameter covariance matrix is a random variable — a sample covariance matrix — and confidence levels described above change somewhat. Only single-variable confidence intervals and Gaussian-distributed variables will be considered for this case.

When using a fitting parameter a_k and its sample standard deviation s_{ak}, one can again express a confidence interval in the form

$$a_k \pm z s_{ak}$$

However, now that the interval is constructed with a sample s_{ak} rather than a true σ_{ak}, intervals for $z = 1$ (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_{ak} might, by chance, come out larger or smaller than σ_{ak}. 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_{ak} are properly taken into account, the confidence level for any z is always smaller than would be predicted based on a known σ_{ak} of the same size as s_{ak}.

In effect, the uncertainty in how well s_{ak} estimates σ_{ak} decreases the confidence level for any given z when compared to an interval constructed with a true σ_{ak} of the same size. Because the uncertainty in s_{ak} 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_{ak} becomes and the closer the Student-T probabilities will be to the corresponding Gaussian probabilities.

Table 10.4 gives some Student-T probabilities. As an example of its use, consider five sample y_i-values obtained from the same Gaussian distribution, which are then used to calculate $\bar{y}$ and $s_{\bar{y}}$. There are four degrees of freedom for a $\bar{y}$ calculated from five samples. Looking in the row for four degrees of freedom, the 95% probability interval for the true mean is seen to be $\bar{y} \pm 2.78 s_{\bar{y}}$. 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 7 *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 probability distribution 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.*

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

An optimization routine, such as Excel's *Solver* program, is quite suitable for performing linear or nonlinear regression. A tutorial on using *Solver* for regression analysis is given in Chapter 10. However, *Solver* does not provide the fitting parameter variances or covariances. The entire covariance matrix $[\sigma_a^2]$ can be determined analytically from Eq. 7.50, or elements of $[\sigma_a^2]$ can be determined numerically based on the "$\Delta\chi^2 = 1$ rule" as described next. As an added benefit the numerical procedure can also provide a check on the range of validity of the first-order Taylor expansions.

The best-fit χ^2 is the reference value for calculating $\Delta\chi^2$. That is, $\Delta\chi^2$ is defined as the difference between a chi-square calculated using some unoptimized trial solution and the chi-square at the best fit. The best-fit chi-square is defined by Eq. 7.15 (or Eq. 7.80, which is also valid for nondiagonal $[\sigma_y^2]$). The trial parameters, denoted a_k^{trial}, will be in the neighborhood of the best-fit a_k and, when used with the fitting function, will give an unoptimized

fitting function

$$y_i^{\rm trial} = F_i(\{a_k^{\rm trial}\}) \tag{9.16}$$

and an unoptimized χ^2 determined from Eq. 7.15 or Eq. 7.80 using the $y_i^{\rm trial}$ for $y_i^{\rm fit}$ in those formulas. Using the linear algebra form of Eq. 7.80, $\Delta\chi^2$ is thus defined

$$\begin{aligned} \Delta\chi^2 &= \left(\boldsymbol{y} - \boldsymbol{y}^{\rm trial}\right)^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\rm trial}\right) \\ &\quad - \left(\boldsymbol{y} - \boldsymbol{y}^{\rm fit}\right)^T \left[\sigma_y^2\right]^{-1} \left(\boldsymbol{y} - \boldsymbol{y}^{\rm fit}\right) \end{aligned} \tag{9.17}$$

Note that $\left[\sigma_y^2\right]^{-1}$ must be the same, best-fit value in both terms above. For Poisson y_i, for example, use $\sigma_i^2 = y_i^{\rm fit}$ even for the unoptimized χ^2—do not use $\sigma_i^2 = y_i^{\rm trial}$.

The first-order Taylor expansion for each $y_i^{\rm trial}$ as a function of $a^{\rm trial}$ about the best fit becomes

$$y_i^{\rm trial} = y_i^{\rm fit} + \sum_{k=1}^{M} \frac{\partial F_i}{\partial a_k}(a_k^{\rm trial} - a_k) \tag{9.18}$$

Or, in linear algebra form

$$\boldsymbol{y}^{\rm trial} = \boldsymbol{y}^{\rm fit} + \left[J_a^y\right] \Delta\boldsymbol{a} \tag{9.19}$$

where the Jacobian is evaluated at the best fit and $\Delta\boldsymbol{a} = \boldsymbol{a}^{\rm trial} - \boldsymbol{a}$ gives the deviations of the trial fitting function parameters from their best-fit values.

Defining

$$\Delta\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{y}^{\rm fit} \tag{9.20}$$

and substituting Eq. 9.19 into Eq. 9.17 then gives

$$\begin{aligned} \Delta\chi^2 &= \left(\Delta\boldsymbol{y}^T - \Delta\boldsymbol{a}^T \left[J_a^y\right]^T\right) \left[\sigma_y^2\right]^{-1} \left(\Delta\boldsymbol{y} - \left[J_a^y\right] \Delta\boldsymbol{a}\right) \\ &\quad -\Delta\boldsymbol{y}^T \left[\sigma_y^2\right]^{-1} \Delta\boldsymbol{y} \\ &= \Delta\boldsymbol{a}^T \left[J_a^y\right]^T \left[\sigma_y^2\right]^{-1} \left[J_a^y\right] \Delta\boldsymbol{a} \\ &\quad -\Delta\boldsymbol{a}^T \left[J_a^y\right]^T \left[\sigma_y^2\right]^{-1} \Delta\boldsymbol{y} - \Delta\boldsymbol{y}^T \left[\sigma_y^2\right]^{-1} \left[J_a^y\right] \Delta\boldsymbol{a} \\ &= \Delta\boldsymbol{a}^T [X] \Delta\boldsymbol{a} \qquad (9.21) \\ &= \Delta\boldsymbol{a}^T \left[\sigma_a^2\right]^{-1} \Delta\boldsymbol{a} \qquad (9.22) \end{aligned}$$

where the best-fit condition — Eq. 7.29, $[J_a^y]^T \left[\sigma_y^2\right]^{-1} \Delta \boldsymbol{y} = 0$, and its transpose — were used to eliminate the last two terms in the second equation. Eq. 7.40 ($[X] = [J_a^y]^T \left[\sigma_y^2\right]^{-1} [J_a^y]$) was used to get to the third equation and Eq. 7.49 ($[X]^{-1} = [\sigma_a^2]$) was used to get the final result.

Equation 9.21 describes a multidimensional parabola with a minimum ($\Delta\chi^2 = 0$) at the best fit ($\Delta \boldsymbol{a} = 0$). It gives the second-order Taylor expansion for $\Delta\chi^2$ given the validity of Eq. 9.19 — the first-order Taylor expansion for the fitting function. Because it's an expansion about the χ^2 minimum, the first-order terms — linear in the Δa_k — are all zero and Eq. 9.21 gives the second-order terms — quadratic in the Δa_k. Equation 9.21 is exact for linear fitting functions, but for nonlinear fitting functions its validity is limited to Δa_k values that are small enough for higher-order terms to be negligible. Because $[X]$ provides the quadratic coefficients, it is referred to as the *curvature matrix.*

The $\Delta\chi^2 = 1$ rule effectively solves Eq. 9.22 for elements of $[\sigma_a^2]$ based on $\Delta\chi^2$ values obtained using Eq. 9.17. The $\Delta\chi^2 = 1$ rule is derived in the *Regression Algebra* addendum and can be stated as follows.

> If a fitting parameter is offset from its best-fit value by its standard deviation, i.e., from a_k to $a_k \pm \sigma_{ak}$, 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 Eq. 9.22 is valid, so is the following equation — a more general form of the $\Delta\chi^2 = 1$ rule showing the expected quadratic dependence of $\Delta\chi^2$ on the change in a_k.

$$\sigma_{ak}^2 = \frac{(a_k^{\text{trial}} - a_k)^2}{\Delta\chi^2} \tag{9.23}$$

Here, $\Delta\chi^2$ is the increase in χ^2 after changing a_k from its best-fit value to a_k^{trial}. However, it is not the increase immediately after the change. It is the increase only after refitting all other fitting parameters for a minimum χ^2 obtained while keeping the one selected a_k^{trial} fixed. The immediate change should follow Eq. 9.22 and is likely to be larger than predicted by Eq. 9.23. However, depending on the degree of correlation among the a_k, some of the immediate increase upon changing to a_k^{trial} will be canceled after the refit. Re-minimizing by adjusting the other parameters is needed to bring the $\Delta\chi^2$ into agreement with Eq. 9.23.

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

$$[\sigma_a^2]_{km} = \frac{(a_k^{\text{trial}} - a_k)(a'_m - a_m)}{\Delta\chi^2} \tag{9.24}$$

A coarse check on the validity of the first-order Taylor expansion of 9.19 and the resultant second-order Taylor expansion of Eq. 9.22 is important when working with nonlinear fitting functions and is easily performed with the same calculations. One simply checks that the variances and covariances obtained with Eqs. 9.23 and 9.24 give the same result using any a_k^{trial} in the interval: $a_k \pm 3\sigma_{ak}$ — a range that would cover most of its probability distribution. The check should be performed for each parameter of interest, varying a_k^{trial} by small amounts both above and below the best-fit value and sized so that $\Delta\chi^2$ values come out around 0.1 and around 10. If all four checks give the same results, the χ^2 is parabolic over that parameter range. If the results vary significantly, the χ^2 is not parabolic and the first-order Taylor expansion is not a good approximation over the $a_k \pm 3\sigma_{ak}$ range. Higher-order terms can bias the a_k and skew their probability distribution.

Chapter 10

Regression with Excel

Excel is a suitable platform for all the statistical analysis procedures discussed in the preceding chapters. Familiarity with Excel is assumed in the following discussion, e.g., the ability to enter and graph data and evaluate formulas.

Linear algebra expressions are evaluated using *array formulas* in Excel. To use an array formula, select the appropriate rectangular block of cells (one- or two-dimensional), 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. Errors with various behaviors result if array dimensions are incompatible with the particular operation or if the selected block is not of the appropriate size. Excel will not allow you to modify parts of any array area defined by an array formula. For example, Clear Contents only works if the cell range covers the array completely.

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

MMULT(array1, array2): Returns the matrix product of the vectors and/or matrices in the order given in the argument list.

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

MINVERSE(array): Returns the inverse of an invertible matrix.

MUNIT(integer): Returns the unit (or identity) matrix of size *integer*.

MDETERM(array): Returns the determinant of a square matrix.

array1 $*$ *array2:* Returns the result of an element by element multiplication. If *array1* and *array2* are identically shaped, *array1* $*$ *array2* returns the same size array with each element equal to the product of the corresponding elements in each array. If one argument is a column vector and the other is a row vector, the $*$ product is an outer product having the same number of rows as the column vector and the same number of columns as the row vector. If one argument is a matrix and the other is a column or row vector, the vector is replicated in adjacent rows or columns until it has the same number of rows or columns as the matrix. If the two matrices to be multiplied do not have the same size, elements are deleted from the rightmost columns and/or bottom-most rows until they do.

Other binary operations, including $+$, $-$, and $/$, between the two arrays behave similarly.

The array functions above can be used to make the vectors and matrices needed in statistics formulas. For example, a diagonal covariance matrix or weighting matrix is easily constructed using the following **vector to diagonal matrix** array expression.

MUNIT(N) $*$ *vector:* Returns an $N \times N$ diagonal matrix with elements equal to the corresponding elements of the column or row *vector* (of size N).

To make a *diagonal covariance matrix* from a *vector* of standard deviations simply replace *vector* with *vector*^2 so that elements of the vector are squared before being used to make the diagonal matrix. Replace $*$*vector* with /*vector*^2 to make the corresponding *diagonal weighting matrix*. The vector to diagonal matrix construction is also useful for making a diagonal Jacobian matrix such as in Eq. 7.87 or Eq. 7.97. Simply construct a column for the needed derivatives and use it as the *vector* above.

A column vector of standard deviations can be extracted from a covariance matrix using the following **diagonal elements from matrix** array expression.

MMULT(MUNIT(N) $*$ *array, ROW(1:N)^0):* Returns an $N \times 1$ column vector containing the diagonal elements of the $N \times N$ matrix *array*.

Note that *ROW(1:N)* returns a column vector of the integers from 1 to N. Raising it to the zeroth power makes all N elements equal to one. Thus,

ROW(1:N)^0 is a column vector of 1's. *MUNIT(N)* $*$ *array* creates an $N \times N$ array with the diagonal elements of *array* unchanged, but with all other elements zeroed out. Matrix multiplying by the column vector of 1's then creates a column vector containing the diagonal elements of *array*. To create a column of standard deviations from a covariance matrix *array,* simply take the square root of the expression above, i.e., *SQRT(...)*.

An array need not be created in a worksheet cell range before using it in another array formula. Even for small data sets, putting an $N \times N$ diagonal weighting matrix or an $N \times N$ diagonal Jacobian matrix on a worksheet serves little purpose. Such matrices are better represented by the column of variances or derivatives used to define them. How can such matrices be created and referenced if not in and by their cell range? One method is to use the expression for the matrix in the other array formula. For example, wherever a 12×12 weighting matrix $\left[\sigma_y^2\right]^{-1}$ is needed, one could use the array expression *MUNIT(12) / D1:D12^2* assuming D1:D12 is the cell range where the column vector of 12 standard deviations is located.

A better method is to use the Name Manager located in the FORMULAS toolbar. You supply a unique reference name along with a spreadsheet formula as could be evaluated in a cell or cell range and the Name Manager registers the association. The name can then be used in other spreadsheet formulas. For example, N in the expression *MUNIT(N)* must resolve to the actual value needed. It could be given explicitly as with the 12 in the weighting matrix example above. N could also be defined by the *ROWS(array)* function which returns the number of rows in *array*, for example, by *ROWS(D1:D12)* above. And it could be created in the Name Manager using either the explicit value or the expression. However, the argument *1:N* in the expression *ROW(1:N)^0* for the column vector of 1's must be handled differently. The function *ROW(range)* returns a column vector of the worksheet row numbers where the elements of *range* are located and, consequently, *range* must resolve to an explicit cell range.[1] Thus, with a 3×3 covariance matrix in cells C5:E7, a column vector with the three standard deviations could be created using the array formulas:

*SQRT(MMULT(MUNIT(3)*C5:E7,ROW(1:3)^0))* or
*SQRT(MMULT(MUNIT(ROWS(C5:E7))*C5:E7,ROW(C5:E7)^0))*

[1]The cell range *1:12* in the expression *ROW(1:12)* (missing an explicit column value) resolves to the range A1:A12.

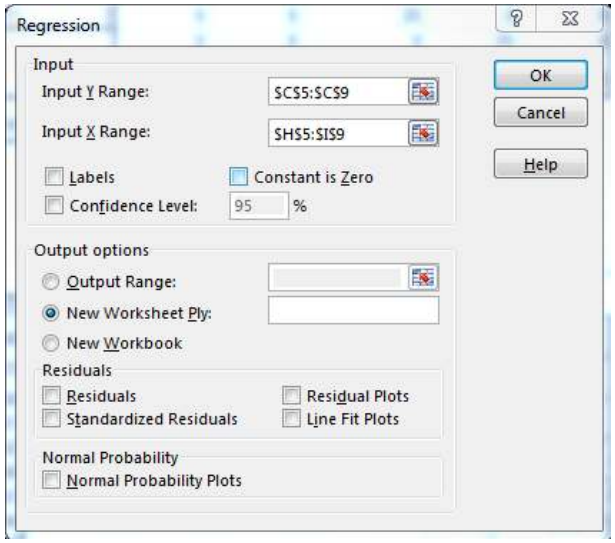


Figure 10.1: Excel's linear regression dialog box.

Array formulas are used for a simple linear regression problem in the *Linear Regression Algebra.xlms* spreadsheet on the lab website. Examine the names in the Name Manager for examples of named ranges and named variables created with these formulas. However, the body of the spreadsheet uses only explicit cell ranges. Try changing them to the named ranges and named variables to learn about the convenience of these constructs.

Excel's Linear Regression Program

Excel's linear regression program is for equally-weighted fits only. Moreover, it does not allow the user to provide σ_y. It uses the s_y of Eq. 9.12 for σ_y in determining the parameter covariance matrix. That is, it forces $\chi^2 = N - M$. It must first be installed from the FILE|Options|Add-Ins page. Select Excel Add-ins, click Go..., check the Analysis ToolPak and OK. The Regression program can then be found inside the Data Analysis program group in the Analysis area of the DATA toolbar. The dialog box appears as in Fig. 10.1.

To use the regression program, first construct columns for x_i and y_i — each as a column of length N. The steps will be described for a quadratic

fit: $y_i^{\text{fit}} = a_1 + a_2 x_i + a_3 x_i^2$ ($M = 3$) having the basis functions: $f_1(x_i) = 1$, $f_2(x_i) = x_i$, and $f_3(x_i) = x_i^2$. Next, create a side-by-side block of M more columns, one for each of the $f_k(x_i)$. For the example, a columns of 1's, then a column of x_i and then a column of x_i^2 would make the parameters in the vectors and matrices appear in the order a_1, a_2 then a_3. In turns out unnecessary to create the column of ones for the a_1 parameter. The Excel regression program can add a constant to any fitting function without one. If linear algebra formulas will be used, however, creating the column of ones for the additive constant would be required.

Select the column containing the y_i-values for the Input Y-Range. For the Input X-Range, select the rectangular block containing all $f_k(x_i)$ values — the Jacobian $[J_a^y]$. For the quadratic fit, select the two columns containing x_i and x_i^2 or select all three columns including the column of ones. Leave the Constant is Zero box unchecked if only the x_i and x_i^2 columns were provided. It would be checked if the fitting function did not include a constant term or if a constant term is included as a column of ones in the Jacobian. Leave the Labels box unchecked. If you would like, check the Confidence Level box and supply a probability for a Student-T interval next to it. Intervals for the 95% confidence level are provided automatically. Select the New Worksheet Ply radio button or the Output Range. For the latter, also specify the upper left corner of an empty spreadsheet area for the results. Then click OK.

The output includes an upper *Regression Statistics* area containing a parameter labeled Standard Error. This is the sample standard deviation s_y from Eq. 9.12. 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 parameter values — labeled *Coefficients* — are their sample standard deviations s_{ak} — labeled *Standard Error*. The last two double columns are for the lower and upper limits of Student-T intervals at confidence levels of 95% and the user specified percentage.

General Regression with Excel

Excel's *Solver* can find the best-fit parameters for both linear or nonlinear fitting functions. The *Solver* can handle weighted y_i and correlated y_i. It must first be installed as an Excel add-in — as described previously for Excel's linear regression program — and will then be found in the Analysis area of the DATA toolbar.

The *Solver* requires the user to construct a cell containing either a chi-square to minimize or a log likelihood to maximize. For pure Gaussian-, binomial-, or Poisson-distributed y_i with negligible uncertainty in the independent variables, one could construct the appropriate log likelihood function $\mathcal{L}$ and maximize that once as described below. For pure Gaussian y_i, a single χ^2 minimization is equivalent and is the more common practice. In fact, χ^2 minimization — with iterative reweighting — has been shown to be equivalent to maximizing likelihood for Poisson- or binomial-distributed data and it is one of the few plausible alternatives if uncertainties in the independent variables must also be taken into account. Recall, iterative reweighting is appropriate whenever the σ_i^2 depend on the best fit. It requires iteration until minimizing the χ^2 — using fixed σ_i^2 that have been evaluated at the best fit — returns the best fit.

Even if a binomial or Poisson log likelihood can be maximized, a χ^2 calculation is still needed for a χ^2 test. In addition, χ^2 calculations can be used to find the parameter uncertainties using the $\Delta\chi^2 = 1$ rule and they can be used to check the validity of the first-order Taylor expansions.

The following instructions describe a single iteration and the iteration method assuming that the σ_i^2 or the input covariance matrix $[\sigma_y^2]$ depends on the best fit.

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

2. Enter the data in columns, starting with x_i and y_i.

3. Use the fitting function to create a column for y_i^{fit} from the x_i and the initial guesses for the a_k.

4. Add additional columns for other input information or derived quantities as needed. For example, if the x_i and/or y_i are Gaussian-distributed random variables, with known uncertainties, add a column for the raw standard deviations σ_{xi} and/or σ_{yi}. If the y_i are from a Poisson or binomial distribution, create a column for σ_i^2 according to Eq. 7.11 or 7.13. If there is uncertainty in the independent variables, create a column for the derivatives of the fitting function $\partial F_i/\partial x_i$ and another for the final σ_i^2 of Eq. 7.78.

5. Create the main graph. Plot the (x_i, y_i) data points with error bars, but without connecting lines. Add a plot of y_i^{fit} vs. x_i as a smooth curve without markers. You can improve the smoothness of the fitted curve by checking the Smoothed line option in the Format Data Series property tab.

   Excel draws the fitted curve connecting the (x_i, y_i^{fit}) points in the order given from first to last. If the x_i are not in either ascending or descending order, the resulting curve will not represent the fit. Either sort the data rows using the Sort function in the DATA tab, or make another column with appropriate and ordered x-values, evaluate the y^{fit}-values there, and use these two columns for the plot. The spacing of the x-values should be chosen small enough to guarantee a smooth curve. If desired, the range of x-values can be chosen to show the fitted curve outside the measured range.

6. Verify that the column of y_i^{fit}-values and its graph depend on the values placed in the cells for the fit parameters. Adjust the fit parameters manually to get the fitted curve close to the data points. *Solver* may fail to find the best-fit parameters if the starting values are not close enough.

7. Columns for the deviations $y_i - y_i^{\text{fit}}$ (residuals) and normalized residuals $(y_i - y_i^{\text{fit}})/\sigma_i$ are also handy — both for evaluating χ^2 and for making plots of these quantities versus x_i for fit evaluation.

8. Construct the χ^2 cell. If a log likelihood maximization is to be performed, construct the $\mathcal{L}$ cell as well.

If needed, the $\mathcal{L}$ cell is easily constructed according to Eq. 7.6 or 7.7. If the y_i are independent, a column of their associated standard deviations σ_i would be needed for a calculation of the χ^2 according to Eq. 7.15. An extra column for the N terms in χ^2 (and in $\mathcal{L}$, if needed) should be constructed for summing. The *SUM(array)* and *SUMSQ(array)* spreadsheet functions are useful for summing elements of an array or their squares.

If the y_i are correlated, the χ^2 of Eq. 7.80 would be needed and the $N \times N$ weighting matrix $\left[\sigma_y^2\right]^{-1}$ would have to be specified. Correlated y_i arise, for example, when they are obtained by preprocessing raw data. Details of the raw data and processing steps then determine the variances and covariances of the y_i.

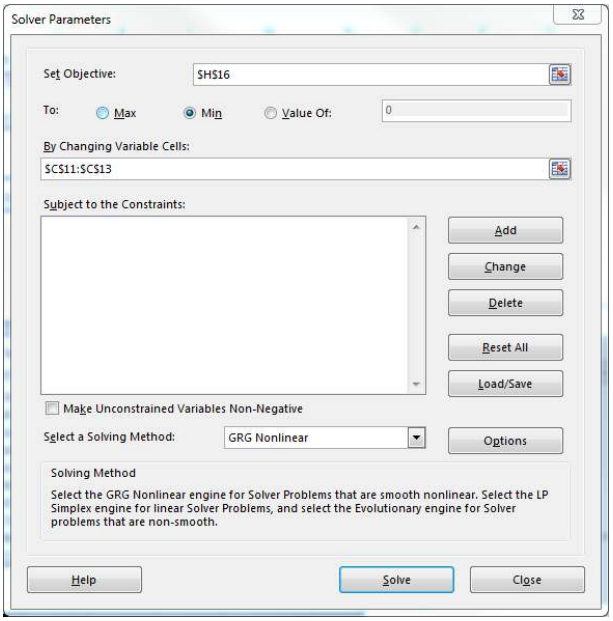


Figure 10.2: Excel's *Solver* dialog box.

If the σ_i^2 or the weighting matrix $\left[\sigma_y^2\right]^{-1}$ depend on the best fit, a constant copy will be needed for an IRLS procedure. Recall that IRLS requires minimizing the χ^2 with the σ_i^2 (or $\left[\sigma_y^2\right]^{-1}$) held fixed. For uncorrelated y_i, the σ_i^2 could be calculated in a spreadsheet column with an adjacent column reserved for the constant copy to be used in Eq. 7.15. If a constant copy of $\left[\sigma_y^2\right]^{-1}$ is needed for a calculation of the χ^2 according to Eq. 7.80 (or for a calculation of the parameter covariance matrix $[\sigma_a^2]$ according to Eq. 7.50), it would best be created in the Name Manager. Simply use the a_k and/or $y_i^{\rm fit}$ to create any needed calculated columns and reserve adjacent columns for the constant copies to be used in creating $\left[\sigma_y^2\right]^{-1}$.

A simple way to make a constant copy is to use the spreadsheet Copy command from the calculated column and then the Paste Special|Values command to the column reserved for the copy. This sets the copied cells to fixed numbers while leaving the cells with the formulas undisturbed. Be sure to use the constant σ_i^2 or $\left[\sigma_y^2\right]^{-1}$ for the calculation of the χ^2 cell.

To run one iteration of the *Solver:*

9. Run the *Solver.* The dialog box is shown in Fig. 10.2.

10. Provide the cell address of the χ^2 or $\mathcal{L}$ in the Set Objective: box. Click the Min radio button next to To: for a χ^2 minimization or click the Max button for an $\mathcal{L}$ maximization.

11. Provide the cell addresses for the fitting parameters in the By Changing Variable Cells: box. Be sure to uncheck the Make Unconstrained Variables Non-Negative box if any of your fitting parameters may be negative. This box could be checked to constrain all fitting parameters to be positive if that is the only appropriate solution. Many other *Solver* options are available, but the default values will typically give good results. For example, use the default solving method, GRG Nonlinear, which is similar to the gradient-descent algorithm.

12. Click on the Solve button. The *Solver's* algorithm then starts with your initial fitting parameters — varying them to find those values which minimize χ^2 or maximize $\mathcal{L}$.

If iterative reweighting is needed, remember to repeat the Copy–Paste Special|Values commands to update any constant columns from their source columns based on the most recent y_i^{fit} and/or a_k. Repeat iterations of the *Solver* until there are no significant changes to the a_k.

Parameter Variances and Covariances

Solver does not provide parameter variances or covariances. The best way to get them is to construct the Jacobian matrix (according to Eq. 7.31) so that the entire covariance matrix $[\sigma_a^2]$ can be obtained from Eq. 7.50 using array formulas. Otherwise, each parameter's variance and its covariances with the other parameters can be individually determined using the following procedure based on the $\Delta\chi^2 = 1$ rule. This procedure is also useful for checking whether the first-order Taylor expansions are valid over each parameter's likely range.

The procedure is a numerical recipe for determining elements of $[\sigma_a^2]$ by checking how χ^2 changes when parameters are varied from their best-fit values. Remember, only the y_i^{fit} should be allowed to change as the parameters are adjusted; the σ_i^2 should not. If the σ_i^2 or $\left[\sigma_y^2\right]^{-1}$ are calculated, use the constant copy evaluated at the best fit.

13. Save the best-fit parameters and the best-fit χ^2 to a new block of cells using the Copy–Paste Special|Values procedure so they will not update in the subsequent steps.

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

15. 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 best-fit χ^2.

16. If χ^2 increased by one, then the amount that a_k was changed is its standard deviation: $\sigma_{ak}^2 = (a_k^{\text{trial}} - a_k)^2$. If the increase in $\Delta\chi^2$ is more (or less) than one, the tested change in a_k is larger (or smaller) than σ_{ak} and the quadratic dependence of Eq. 9.23 can then be used to determine σ_{ak}^2.

17. The covariances between a_k and the other parameters can also be determined by keeping track of the changes in the other parameters after the re-optimization. Equation 9.24 can then be used to solve for the covariances.

18. Check that this procedure gives roughly the same parameter variances and covariances using a_k^{trial} values both above and below a_k and sized such that $\Delta\chi^2$ values are around 0.1 and around 10. If σ_{ak}^2 is reasonably constant for all four cases, one can be reasonably assured that the χ^2 is roughly parabolic throughout the likely parameter range: $a_k \pm 3\sigma_{ak}$.

Multicollinearity and Other Problems

Multicollinearity is a problem in regression analysis that arises when columns of the Jacobian $[J_a^y]$ are not all linearly independent, i.e., when one or more columns can be expressed (exactly or nearly) as a linear superposition of other columns. Exact multicollinearity causes an X-matrix that is not invertible and results in fitting parameters at the χ^2 minimum that are not unique. Near multicollinearity causes an X-matrix that is *ill-conditioned* — numerically difficult to invert accurately and sensitive to small variations

in the Jacobian. It results in highly correlated fitting parameters that are likewise numerically difficult to determine accurately. Diagnosing and treating multicollinearity — topics only briefly touched upon below — depend on whether the regression is linear or nonlinear and other details of the model and the data.

Collinearity between two parameters, say a_j and a_k, may be observable in the parameter covariance matrix by a correlation coefficient $\rho_{jk} = [\sigma_a^2]_{jk}/\sqrt{[\sigma_a^2]_{jj}[\sigma_a^2]_{kk}}$ near 1 or -1. Multicollinearity among more than two parameters can be diagnosed by performing an equally-weighted linear regression on each column of the Jacobian, column k say, against all the other columns. An R-square[2] near one then indicates column k is highly correlated with one or more of the other columns and parameter a_k may be part of a multicollinearity problem.

One might ignore the problem of multicollinearity if all parameters of interest are well determined by the fit. If important parameters turn out to be problematic, a more detailed analysis of how the χ^2 varies with the fit parameters could be helpful. Fixing, constraining, eliminating, or combining fitting parameters and their associated functional dependencies might be appropriate. There are also regression techniques specialized for multicollinearity such as *principal component analysis* and *ridge regression.*

In addition to multicollinearity problems, the *Solver* may fail to find the best fit for a variety of other reasons. If the initial parameter guesses are not close enough to the best-fit values, they may need to be readjusted before the program will proceed to the solution. If a parameter wanders into unphysical domains leading to an undefined function value, constraints may need to be placed on its allowed values.

Solver sometimes has problems with poorly scaled models where y-values, parameters, or the target cell are extremely large or small. *Solver* has a default option to Use Automatic Scaling, which does not always fix the problem. If it is unchecked, check it to see if enabling it fixes the problem. If not, try explicit parameter scaling. For example, if the y-values of a measured exponential decay are of order 10^6 while the mean lifetime is of order 10^{-6} s, and the background is around 10^3, rather than fit to Eq. 7.56, *Solver* performs

[2]R-square or R^2 is a measure of how well the y_i fit the linear model by comparing the best-fit deviations $y_i - y_i^{\text{fit}}$ to the deviations $y_i - \bar{y}$ for a model where all y_i are simply samples from a common distribution with a mean of $\bar{y}$. Its value can be expressed: $R^2 = 1 - \Sigma(y_i - y_i^{\text{fit}})^2/\Sigma(y_i - \bar{y})^2$.

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 regression exercises.

better with

$$y_i^{\text{fit}} = 10^6 a_1 e^{-10^6 t_i/a_2} + 10^3 a_3$$

so that all three fitting parameters are of order unity.

Regression Exercises

In these last exercises you will perform equally-weighted fits for the data of Table 10.1 to a quadratic formula: $y_i^{\text{fit}} = a_1 + a_2 x_i + a_3 x_i^2$. You will solve it three ways: using Excel's linear regression program, evaluating the linear algebra regression formulas, and using the *Solver* program. Augment and save the spreadsheet as you work through the exercises, but the questions must be answered on a separate sheet clearly labeled with the question part and the answer. Feel free to cut and paste from the spreadsheet, but it will not be opened or graded. This is a fit of $N = 6$ points to a formula with $M = 3$ fitting parameters; it has $N - M = 3$ degrees of freedom. Keep this in mind when discussing χ^2, χ^2_ν or Student-T probabilities, which depend on the degrees of freedom.

Exercise 8 *Start by creating the 6-row column vector for* $\boldsymbol{y}$*, i.e., the* y_i *in Table 10.1 and the* 6×3 *Jacobian matrix* $[J_a^y]$*, i.e., three side-by-side columns of 1's for* a_1*,* x_i *for* a_2*, and* x_i^2 *for* a_3*. This matrix will also be needed shortly when the regression formulas are used. Use it now as you solve the problem using Excel's linear regression program. Remember that if you include the column of 1's in the* Input X Range*, the* Constant is Zero *box must be checked. (a) Locate the parameters* a_k *and their sample standard deviations* s_{ak}*. (b) Locate the sample standard deviation* s_y *and show a calculation of its value*

according to Eq. 9.12. Hint: Add a column for y_i^{fit} *and another for* $y_i - y_i^{fit}$. *Then use the Excel* SUMSQ(array) *function to do the sum of squares.*

Exercise 9 *Now use the equally weighted linear regression formulas to solve the problem. First create the* 3×3 *matrix for* $[X_u]^{-1}$ *based on Eq. 7.52 for* $[X_u]$. *Hint: This formula and those to follow will require array formulas and Excel's* MMULT, TRANSPOSE, *and* MINVERSE *array functions. Next, construct the 3-row column vector for the best fit parameters* $\boldsymbol{a}$ *according to Eq. 7.54 and show they agree with Excel's regression results. Finally, construct the parameter covariance matrix* $[\sigma_a^2]$ *from Eq. 7.53 using* $\sigma_y = s_y$ *and show how the parameter sample standard deviations* s_{ak} *provided by Excel's regression program can be obtained from this matrix.*

Exercise 10 *Excel's linear regression program assumes no prior knowledge of* σ_y *(except that it is the same for all* y_i*). The parameter standard deviations returned by the program are calculated using the random variable* s_y *as an estimate of* σ_y. *Consequently, they are sample standard deviations* s_a *and this is why Student-T probabilities are used when Excel constructs confidence intervals. Give the linear regression program's 95% confidence interval (for the quadratic coefficient only) and show how it can be obtained from* a_3, s_{a3} *and the Student-T table.*

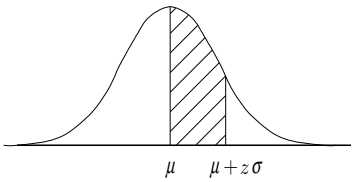
Exercise 11 *If the true* σ_y *is known, its value should be used in Eq. 7.53 for determining* $[\sigma_a^2]$, *which then becomes a true covariance matrix. For this question assume the* y_i *of Table 10.1 are all known to come from distributions with* $\sigma_y = 0.5$. *The data are still equally weighted and thus the best-fit parameter values and the sample standard deviation* s_y *do not change. Give the true parameter standard deviations* σ_{ak} *for this case. Note the important scaling implied by Eq. 7.53, namely, that* σ_{ak}/σ_y *is a constant. It can be applied here for translating from Excel's* s_{ak} *to an* σ_{ak} *appropriate for a given* σ_y. *Give the 95% confidence interval for the quadratic coefficient. Should you use Student-T or Gaussian probabilities for this case?*

Exercise 12 *(a) Do a graphical evaluation of the fit assuming* $\sigma_y = 0.5$. *Add a cell for* σ_y *and reference it to evaluate* χ^2 *according to Eq. 7.15 or Eq. 7.55. Evaluate* χ^2_ν *according to Eq. 9.13 or* $\chi^2_\nu = s_y^2/\sigma_y^2$. *Note that these four formulas for* χ^2 *and* χ^2_ν *are just different ways to calculate and describe the exact same information about the actual and expected deviations. What*

is the probability that a χ^2 or χ^2_ν random variable would come out as big or bigger than it did here? (b) For this data and fit, how small would σ_y have to get before one would have to conclude (at the 99% level, say) that the fit deviations are too big to be reasonable?

Exercise 13 *(a) Use* Solver *to do the fit. The* Solver *will be used in a single χ^2 minimization; no iteration will be needed. Demonstrate that the a_k do not depend on the value used for σ_y — that with any σ_y the optimized a_k are the same as those from the linear regression program. (b) Use the $\Delta\chi^2 = 1$ rule to determine the parameter standard deviations. Start by assuming σ_y is unknown and use the sample standard deviation s_y for σ_y in the calculations. Recall, this means the parameter covariance matrix and standard deviations will be sample values. What value of χ^2 does this produce? Why is this value expected? Use the $\Delta\chi^2 = 1$ rule to determine the sample standard deviation s_{a3} for the a_3 fit parameter only. Show it is the same as that obtained by Excel's linear regression. (c) Now assume it is known that $\sigma_y = 0.5$ and use it with the $\Delta\chi^2 = 1$ rule to determine the standard deviation σ_{a3} in the a_3 parameter. Compare this σ_{a3} value with the s_{a3} value from part (b) and show that they scale as predicted by Eq. 7.53.*

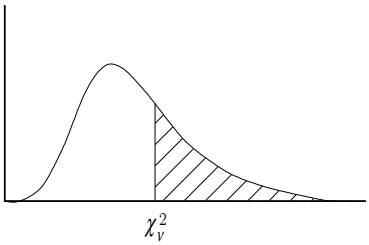
Gaussian Probabilities



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 / ν	0.99	0.95	0.90	0.80	0.70	0.68	0.60	0.50
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.