Statistical Analysis of Data

Robert DeSerio

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

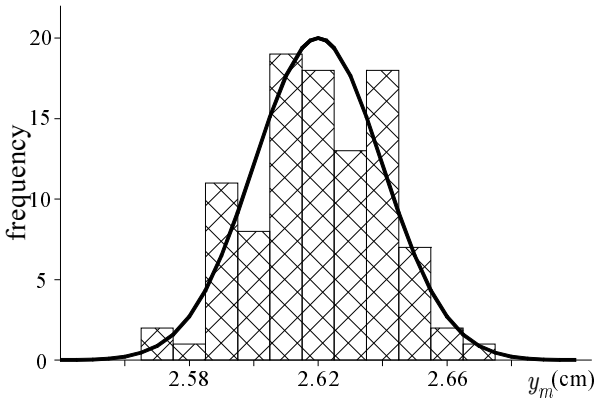


Figure 1: A sample frequency distribution for $N = 100$ measurements of the length of a rod. The parent Gaussian distribution is shown by the smooth curve.

References

P. R. Bevington, and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, (McGraw-Hill, New York 1992).
H. D. Young, *Statistical Treatment of Experimental Data*, (McGraw-Hill, New York, 1962).

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 measurements will not all be the same and a *frequency distribution* (see Fig. 1) is one way to display the resulting data. The height of each vertical bar indicates the number of times a particular measurement value occurs and the value is indicated by the position of the bar along the horizontal axis.

The measurements are referred to as a *sample set,* the number of measurements N is called the *sample size*, and the frequency distribution is called a *sample distribution.* Were new sample sets taken, random errors would cause each new sample distribution to vary. However, as the sample size increases to infinity, the sample distribution stabilizes. As $N \to \infty$, the sample distribution converges to the *parent distribution*—a distribution containing complete statistical information about the particular measurement.

Many of the statistical analysis procedures described here will strictly apply only to data whose parent distributions would have the bell shape of the Gaussian distribution. This restriction need not apply rigorously. Results based on the assumption of a Gaussian parent distribution will often be acceptable even for data with non-Gaussian distributions.[1]

[1]However, beware of predictions in the tails of a distribution, i.e., the (small) probabilities of getting results far from the mean. The probabilities in the tails are typically where real data deviate most sig-

Thus, a single measurement should be regarded as one sample from a parent distribution. Equivalently, a measurement can be modeled as the sum of a signal component and a noise component. The signal component would be the center value or mean of the measurement's parent distribution, and the noise component would be the random error—a quantity equally likely to be positive or negative—that scatters individual measurement values about the mean.

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

Theoretical models provide relationships for physical variables. For example, the temperature of a fixed quantity of gas might be measured along with the pressure and volume in a study of various equations of state such as the van der Waals model. Coming up with and testing theoretical models are typical experimental objectives.

Many statistical analysis procedures can now be broadly summarized as providing some measure of the probability of a data set given the following two hypotheses:

Experimental model: that measurements have random errors about a mean and that any systematic deviations between that mean and the true value of the physical quantity assigned to each measurement are sufficiently small.

Theoretical model: that some particular theoretical relationships or predictions about the physical quantities are correct.

If the statistical analysis then indicates that the data set is too unlikely, at least one of the hypotheses must be rejected. But even if the analysis indicates that the data are reasonably likely, one must still be cautious. The data can appear consistent with the hypotheses even when one or both are false. In particular, systematic errors are often difficult to disentangle from the theoretical model. Sorting out the behavior of measuring instruments from the behavior of the system under investigation and designing experimental procedures to individually and quantifiably verify all aspects of both hypotheses are basic goals of the experimental process.

Random Variables

A *random variable* is a numerical quantity having a value which varies as one repeats the procedure used to obtain it. Furthermore, the different values or outcomes should occur with fixed probabilities (whether or not those probabilities are known). Throwing a pair of dice generates a random variable y_i with the possible outcomes being the integers from two through twelve. A throw of the dice is a *discrete random variable*, as the outcomes are countable. Each possible outcome for a discrete random variable occurs with a particular probability $P(y_i)$. For example the probability of throwing "boxcars" (two sixes) in a

nificantly from a Gaussian and the true probabilities of getting results in the tails can be significantly different from the predictions based on a Gaussian data distribution.

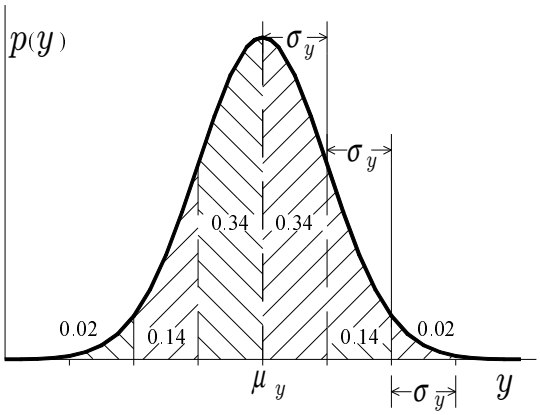


Figure 2: The Gaussian pdf

dice game is $P(12) = 1/36$, while the most likely throw, seven, occurs with a probability of $P(7) = 1/6$.

Outcome probabilities for a *continuous random variable* are specified by a *probability density function* (pdf) $p(y)$ for all values y in the range of possible outcomes. Probabilities for various outcomes are then obtained from the pdf according to either of the following two equivalent statements. Differential statement: The probability $dP(y)$ of an outcome between y and $y + dy$ is given by

$$dP(y) = p(y)dy \tag{1}$$

Integral statement: The probability $P(y_1 < y < y_2)$ of an outcome between y_1 and y_2 is given by

$$P(y_1 < y < y_2) = \int_{y_1}^{y_2} p(y)\, dy \tag{2}$$

In the introduction, we referred to a frequency distribution for a sample set of infinite size as the parent distribution. It is also common to refer to the pdf for a random variable as the variable's *parent.* Knowing the pdf is equivalent to having an infinite sample set.

Exercise 1 *(a) How would you obtain an approximate value of $p(y)$ (for any y) from a large sample set, y_i, $i = 1..N$? Your answer should contain N, Δy, and $\Delta N(y)$, where $\Delta N(y)$ is the number of sample measurements in the interval between $y - \Delta y/2$ and $y + \Delta y/2$. $\Delta N(y)$ can be considered given—obtained from the sample by counting the actual number of y_i's in the interval. Hint: Write an expression for the value of $\Delta N(y)$ which would be expected for a given N, $p(y)$, and Δy. (b) What are the considerations in choosing the size of Δy, i.e., what would happen if it were made too small or too big?*

The Gaussian

The Gaussian probability density function has the mathematical form

$$p_G(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right] \tag{3}$$

and is parameterized by two quantities: the mean μ_y and the standard deviation σ_y. The quantity σ_y^2 is called the variance. Fig. 2 shows the Gaussian pdf and gives various integral probabilities. A more complete listing of probabilities can be found in Table 2. You should be familiar with the one-sigma and two-sigma probabilities; there is a 68% probability the measurement will be within one standard deviation of the mean and a 95% probability it will be within two.

When a measurement is written down with a certain number of digits, the probability of the measurement is assumed to be the integral of the Gaussian over a range $\pm 1/2$ of the least significant digit. For example, if a voltage reading is written as $y = 3.72$ V, then the integral probability ΔP of this reading would be

$$\Delta P(3.72) = \int_{3.715}^{3.725} p_G(y')\, dy'$$

with an implicit dependence on the values of μ_y and σ_y appearing in $p_G(y')$.

Most of the statistical analysis procedures we will develop are based on the ability to use differential probabilities to describe real measurement values. For this to be valid, the size Δy of the least significant digit (0.01 V in the example above) must be small compared to the standard deviation σ_y. Then, $p_G(y)$ will not vary significantly over the range of integration and we can write

$$\begin{aligned} \Delta P(y) &= \int_{y-\Delta y/2}^{y+\Delta y/2} p_G(y')\, dy' \\ &= p_G(y)\, \Delta y \end{aligned} \tag{4}$$

i.e., the integral probability is effectively differential.

If Δy is large compared to σ_y, the integral form would be needed to describe the measurement probability and the ability to use standard analysis techniques would be compromised. In the example above, were σ_y of the order 0.0001 V, a reading of 3.72 V, were it obtained once, would be obtained on virtually all subsequent measurements. Consequently, multiple measurements and averaging (a common statistical procedure) could not be used to improve the determination of the mean.

Uncertainties, errors, deviations

Random errors

A measurement y can be expressed as the sum of the mean of its pdf μ_y and a *random error* δ_y

$$y = \mu_y + \delta_y \tag{5}$$

The quantity $\delta_y = y - \mu_y$ is also referred to as the *deviation* (of the measurement from the mean).

Normally, the mean is not known and thus the random error cannot be determined. Instead, the experimenter provides an *uncertainty* quantifying how big a deviation can reasonably be expected. These uncertainties will often be the basis for any statistical assessment of the agreement between theory and experiment.

In this lab, an uncertainty will be taken as the standard deviation σ_y of the parent Gaussian for the measurements. For example, a rod length reported as 2.64 ± 0.02 cm would indicate a measurement value $y = 2.64$ cm and a standard deviation $\sigma_y = 0.02$ cm. Thus, an uncertainty should not include possible systematic errors; it should only take into account expected random errors. Furthermore, its size should not be chosen extra large to improve the chances of getting results "in range." It should be chosen at the "one-sigma" confidence level—in agreement with a standard deviation—so that the range $y \pm \sigma_y$ should include the mean μ_y 68% of the time or, equivalently, with a 68% probability.

Systematic Errors

While random errors 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 (non-randomly) from the true physical quantity the mean is interpreted to represent. With y_t representing this true value and δ_{sys} the systematic error, this might be expressed

$$\mu_y = y_t + \delta_{\text{sys}} \tag{6}$$

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

Accuracy and precision

Combining Eqs. 5 and 6

$$y = y_t + \delta_y + \delta_{\rm sys} \tag{7}$$

demonstrates that both random and systematic errors contribute to the measurement. Both can be made smaller but neither can ever be entirely eliminated. Accuracy refers to the size of possible systematic errors $\delta_{\rm sys}$ while precision refers to the size of possible random errors δ_y.

While statistical analysis procedures can help determine the size and effects of random errors, they have little to say about systematic errors. Determining the possible size and effects of systematic errors is an important, but often overlooked, part of the experimentalist's work. If systematic errors might be a significant problem, additional measurements (calibrations) should be made to check for and correct them. However, as there is no general way to treat uncorrected systematic errors, they will be presumed insignificant in the sections to follow.[2] Consequently, **the mean of a measurement distribution is always assumed to be the true value.** In effect, this is the experimental model.

[2] A calibration can be treated as an additional hypothesis—either experimental or theoretical—relating the mean of an uncorrected instrument reading (which then becomes its own independent physical quantity) to the true value of the real physical quantity of interest.

Sample and parent averages

Let y_i, $i = 1..N$ represent sample values for a random variable y having probabilities of occurrence governed by a pdf $p(y)$. The *sample average* of any function $f(y)$ (denoted $\overline{f(y)}$) is defined as the value of $f(y)$ averaged over all values of y_i

$$\overline{f(y)} = \frac{1}{N}\sum_{i=1}^{N} f(y_i) \tag{8}$$

For example, for the function $f(y) = y$, application of Eq. 8 represents simple averaging of the y-values

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{9}$$

Note that $\bar{y}$, or the sample average of any function, is a random variable; taking a new sample set would produce a different value. However, in the limit of infinite sample size, the average defined by Eq. 8 converges to a well defined constant depending only on the pdf $p(y)$ and the function $f(y)$. This constant is called the *expectation value* of $f(y)$ and will be denoted by putting angle brackets around the function

$$\langle f(y)\rangle = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(y_i) \tag{10}$$

Since having an infinite sample is equivalent to having the parent frequency distribution, an expectation value is also called a *parent average*. The next exercise demonstrates how to obtain a parent average from $p(y)$.

Exercise 2 *Show that Eq. 10 leads to the equivalent expression:*

$$\langle f(y)\rangle = \int_{-\infty}^{\infty} f(y)\, p(y)\, dy \tag{11}$$

Hints: Consider the number of measurements $dN(y)$ that can be expected in an interval between y and $y + dy$. Each measurement in

this interval contributes $f(y)$ and together they contribute $dN(y)f(y)$ to the sum in Eq. 10. Finally, consider the entire range of possible y-values.

Two frequently used properties of expectation values are easily derived:

1. The expectation value of a sum of terms is the sum of the expectation value of each term.

2. Constants can be factored out of the expectation value brackets.

However, the expectation value of a product is not necessarily the product of the expectation values, e.g., $\langle y^2 \rangle \neq \langle y \rangle^2$.

Exercise 3 *Show by direct integration (no integral tables) that the Gaussian pdf (Eq. 3), is properly normalized, i.e., that*

$$\int_{-\infty}^{\infty} p(y)dy = 1 \tag{12}$$

Also directly integrate Eq. 11 to show the following

$$\langle y \rangle = \mu_y \tag{13}$$

$$\left\langle (y-\mu_y)^2 \right\rangle = \sigma_y^2 \tag{14}$$

Obviously, you must use $f(y) = y$ for the former and $f(y) = (y-\mu_y)^2$ for the latter.

Equation 14 is the origin of the name standard deviation in that the left side of that equation gives a standard measure of the deviations $y-\mu_y$ that can be expected from a sample. Equally likely to be positive as negative, deviations have a zero mean ($\langle y-\mu_y \rangle = 0$, Eq. 13) and thus some other measure will be needed to describe their average size. The mean absolute deviation—the mean of the absolute value of the deviation—would be one such measure. However, the standard measure is the root-mean-square or rms value. The rms value is the square root of the mean of the square of the values and Eq. 14 says that σ_y is this standard deviation for the parent distribution.

Principle of maximum likelihood

Often, an experimental or theoretical model includes constants that are predicted to influence the data. For example, they might be calibration constants or the slope and intercept of a predicted linear relationship between two physical quantities. The constants are assumed to be unknown and *estimators* of their true values are to be determined based on the experimental data.

One accepted technique is to choose estimators in such a way as to maximize the probability of the data set. This guiding idea is called the *principal of maximum likelihood* and any parameter value it produces is called a *maximum likelihood estimator.* We will call such an estimator a *best estimate* and we will always give it a symbol separate from the true value of the quantity it estimates.

One consequence of the technique is that if the data still turn out to be too unlikely, the hypotheses can confidently be rejected regardless of the parameter values; any other values will only make the data less likely. Another important property of best estimates is that they will often be *unbiased estimators.* The expectation value of an unbiased estimator is the true value of the quantity being estimated. This is not a necessary consequence of the Principle and is an important property to check for in any results obtained using it. One would like to be certain that if the experiment is repeated many times, the true values would be obtained "on average."

Weighted Average

The first example we will treat using the Principal is a data set $y_i, i = 1..N$ in which each of the values is supposed to be a measure of the same physical quantity. That is, each y_i is assumed to be a sample from a Gaussian pdf with the same mean μ_y, but this mean or true value is unknown and is to be chosen based on the data set. Further assume that the measurements are obtained from a variety of instruments and/or methods so that each measurement y_i has a different standard deviation σ_i associated with its parent Gaussian. The y_i (the measurement values) and the σ_i (the standard deviations) are given. We wish to find only the best estimate of the mean μ_y.

First, we write an expression for the probability $dP(y_1)$ that the first measurement is in a small interval dy_1 about the measured value y_1 in terms of y_1, dy_1, μ_y and σ_1.

$$dP(y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(y_1-\mu_y)^2}{2\sigma_1^2}\right] dy_1 \tag{15}$$

The probability of the whole set will be the product of N such terms, one for each y_i.

$$dP(\{y_i\}) = \frac{dy_1 dy_2 ... dy_N}{\left(\sqrt{2\pi}\right)^N \sigma_1\sigma_2..\sigma_N} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu_y)^2}{\sigma_i^2}\right] \tag{16}$$

The best estimate of μ_y is that value which maximizes this probability. The probability maximizes when the sum in the argument of the exponential is minimized. This sum

$$\chi^2 = \sum_{i=1}^{N}\frac{(y_i-\mu_y)^2}{\sigma_i^2} \tag{17}$$

called the *chi-square*, has several uses. Here, we consider χ^2 to be a function of μ_y and find the value of μ_y where it becomes a minimum. This value (denoted $\bar{y}$ for reasons that will become obvious shortly) occurs where the derivative of χ^2 with respect to μ_y is zero.

$$\left.\frac{d\chi^2}{d\mu_y}\right|_{\mu_y=\bar{y}} = 0 = -2\sum_{i=1}^{N}\frac{y_i-\bar{y}}{\sigma_i^2} \tag{18}$$

Solving for $\bar{y}$ gives

$$\bar{y} = \frac{\sum_{i=1}^{N}\frac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}} \tag{19}$$

Eq. 19 is called a weighted average:

$$\bar{y} = \frac{w_1y_1 + w_2y_2 + ... + w_Ny_N}{w_1 + w_2 + ... + w_N} \tag{20}$$

where each weight is the inverse of the variance:

$$w_i = \frac{1}{\sigma_i^2} \tag{21}$$

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

This value of $\bar{y}$ when used for μ_y minimizes the chi-square of Eq. 17 and maximizes the probability of the data set $\{y_i\}$. Thus, according to the Principle, it is the best estimate of μ_y. But $\bar{y}$ is not expected to be exactly μ_y. It is a random variable, following its own pdf.

We can find the expectation value of $\bar{y}$ quite simply by taking the expectation value of Eq. 20.

$$\langle\bar{y}\rangle = \left\langle\frac{w_1y_1 + w_2y_2 + ... + w_Ny_N}{w_1 + w_2 + ... + w_N}\right\rangle$$

Noting that the weights w_i are simply constants, and using the rules for the expectation value of a sum this becomes

$$\langle\bar{y}\rangle = \frac{w_1\langle y_1\rangle + w_2\langle y_2\rangle + ... + w_N\langle y_N\rangle}{w_1 + w_2 + ... + w_N}$$

The expectation values $\langle y_1 \rangle$, $\langle y_2 \rangle$, etc., are all μ_y since each y_i is a sample from a Gaussian parent with a mean μ_y (Eq. 13) giving

$$\begin{aligned} \langle \bar{y} \rangle &= \frac{w_1\mu_y + w_2\mu_y + ... + w_N\mu_y}{w_1 + w_2 + ... + w_N} \\ &= \mu_y \end{aligned} \tag{22}$$

Thus, the sample average $\bar{y}$ defined by Eq. 19 would "on average" give the correct mean μ_y. However, as already mentioned, $\bar{y}$ will not equal μ_y for any particular sample. We will not try to prove that $\bar{y}$ varies according to a Gaussian pdf—it does, and we have just shown that the mean of this pdf is μ_y.

We should also want to know the standard deviation of the pdf for $\bar{y}$. To find this value we use the definition of the variance as the expectation value of the squared magnitude of the variable from its mean (Eq. 14)

$$\sigma_{\bar{y}}^2 = \left\langle (\bar{y} - \mu_y)^2 \right\rangle$$

Substituting Eq. 20 for $\bar{y}$ gives

$$\sigma_{\bar{y}}^2 = \left\langle \left(\frac{w_1y_1 + w_2y_2 + ... + w_Ny_N}{w_1 + w_2 + ... + w_N} - \mu_y \right)^2 \right\rangle$$

This equation is easier to evaluate if it is rewritten

$$\sigma_{\bar{y}}^2 = \left\langle \left(\frac{w_1(y_1 - \mu_y) + ... + w_N(y_N - \mu_y)}{w_1 + ... + w_N} \right)^2 \right\rangle \tag{23}$$

The sum inside this expectation value is squared and this squaring must be done before the expectation value can be treated as a sum of individual terms. The squaring leads to quadratic terms $(y_i - \mu_y)^2$ and cross terms $(y_i - \mu_y)(y_j - \mu_y)$, $i \neq j$.

Exercise 4 *Show that the expectation value of any cross term is zero.*

$$\langle (y_i - \mu_y)(y_j - \mu_y) \rangle = 0 \tag{24}$$

Hint: the probability for any pair of sample values y_i, y_j to occur is given by Eq. 16

$$dP(y_i, y_j) = \frac{dy_i dy_j}{2\pi\sigma_i\sigma_j} \exp\left[-\frac{(y_i - \mu_y)^2}{2\sigma_i^2} - \frac{(y_j - \mu_y)^2}{2\sigma_j^2} \right]$$

The expectation value in Eq. 24 should then be evaluated from the expression

$$\int_{y_i=-\infty}^{\infty} \int_{y_j-\infty}^{\infty} (y_i - \mu_y)(y_j - \mu_y) dP(y_i, y_j)$$

Dropping all cross terms in Eq. 23 leaves only the quadratic terms

$$\sigma_{\bar{y}}^2 = \left\langle \frac{w_1^2(y_1 - \mu_y)^2 + ... + w_N^2(y_N - \mu_y)^2}{(w_1 + w_2 + ... + w_N)^2} \right\rangle$$

Taking expectation values term by term, using $\langle (y_i - \mu_y)^2 \rangle = \sigma_i^2$ (Eq. 14), and noting that $\sigma_i^2 = 1/w_i$ (Eq. 21), gives

$$\begin{aligned} \sigma_{\bar{y}}^2 &= \frac{w_1 + w_2 + ... + w_N}{(w_1 + w_2 + ... + w_N)^2} \\ &= \frac{1}{w_1 + w_2 + ... + w_N} \end{aligned}$$

or

$$\frac{1}{\sigma_{\bar{y}}^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \tag{25}$$

Equation 25 always gives a $\sigma_{\bar{y}}$ smaller than any of the individual σ_i. Effectively, Eqs. 19 and 25 are a prescription for turning a group of independent samples (with known standard deviations) into a single sample $\bar{y}$ with a reduced standard deviation $\sigma_{\bar{y}}$.

Sample mean and standard deviation

In this section, the Principal will be used to obtain best estimates of both μ_y and σ_y from a sample set y_i, $y = 1...N$ all from the same Gaussian parent distribution.

All of the previous discussion for weighted averaging still applies. With all σ_i equal (we

will use $\sigma_i = \sigma_y$ in this case), the σ's then drop out of Eq. 19 giving

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{26}$$

This is just the sample average, Eq. 9. Previously shown to have an expectation value equal to μ_y, it is now seen to be the best estimate for μ_y.

Equation 25 applied to this case leads to

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{N}} \tag{27}$$

It states that the standard deviation of the mean is $\sqrt{N}$ times smaller than the standard deviation of a single value, i.e., the average of 100 values is 10 times more precise than a single value alone. Equation 27 is an important result, but it does not provide an estimate of σ_y.

To determine a best estimate of σ_y from a sample set of y_i's (all from the same parent) we will first assume that μ_y is known. Then the probability of the whole set is constructed (with all equal $\sigma_i = \sigma_y$) leading to

$$dP(\{y_i\}) = \frac{dy_1 dy_2 ... dy_N}{\left(\sqrt{2\pi}\sigma_y\right)^N} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i-\mu_y)^2}{\sigma_y^2}\right] \tag{28}$$

Exercise 5 *Show that maximizing $dP(\{y_i\})$ with respect to σ_y leads to the following equation for the best estimate (which we will call s_y^2) of the parent variance σ_y^2*

$$s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_y)^2 \tag{29}$$

s_y^2 is called the sample variance. s_y is called the sample standard deviation.

Typically, μ_y is not known. Can its best estimate $\bar{y}$ be used for μ_y in Eq. 29? Not exactly. If one uses $\bar{y}$, the best estimate of the parent variance is

$$s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2 \tag{30}$$

One can get a glimpse into the verity of the $1/(N-1)$ factor in Eq. 30 by considering a sample set consisting of a single measurement y_1. In this case, $N = 1$ and the sample average is well defined; $\bar{y} = y_1$, indicating the best estimate of parent mean is the single measured value y_1. But Eq. 30 gives $s_y^2 = 0/0$, an indeterminate expression indicating—quite reasonably—that an estimate of the parent variance cannot be obtained from a single measurement.

Exercise 6 *Both Eqs. 29 and 30 give the same expectation value $\left\langle s_y^2 \right\rangle = \sigma_y^2$ for any value of N. (a) Show this is true for Eq. 29. This is trivial using Eq. 14. (b) Show this is true for Eq. 30. This is not trivial at all. You must use Eq. 26 and keep track of all terms. Hint 1: Explain why each of the N terms in Eq. 30 has the same expectation value and then use this fact to get rid of the sum over i—replacing it with a factor of N times the expectation value of one term (say $i = 1$). Hint 2: Show that both Eqs. 14 and 24 can be expressed by the equation*

$$\langle y_i y_j \rangle = \mu_y^2 + \sigma_y^2\,\delta_{ij} \tag{31}$$

where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. (This form is useful in summations.) Hint 3: To reexpress $\bar{y}^2$ use

$$\bar{y}^2 = \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)\left(\frac{1}{N}\sum_{j=1}^{N} y_j\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \tag{32}$$

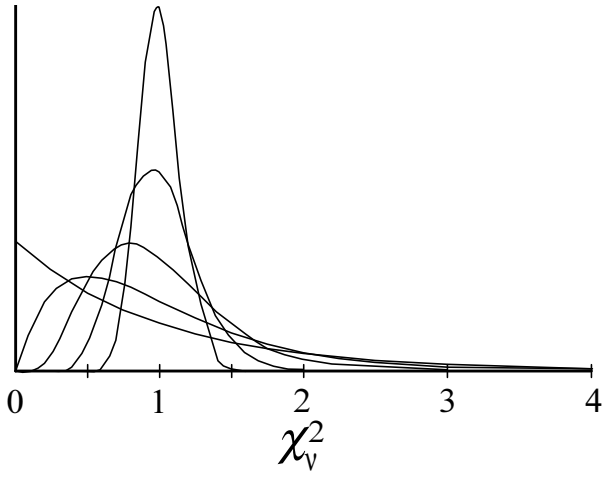


Figure 3: The pdf $p_\nu(\chi^2_\nu)$ vs. the value of χ^2_ν for $\nu = 2, 4, 10, 30, 100$. The curve for $\nu = 100$ has the sharpest and highest distribution. For the other values of ν the curves get broader and lower. For $\nu = 1$ (not shown) the pdf is actually (integrably) infinite at $\chi^2_\nu = 0$.

That is, each $\bar{y}$ must use its own private dummy index. Hint 4: Show how Eq. 32 with Eq. 31 gives

$$\left\langle \bar{y}^2 \right\rangle = \mu_y^2 + \frac{\sigma_y^2}{N} \tag{33}$$

which is basically a restatement of Eq. 27 and useful for this exercise.

We have shown that the sample variance s_y^2 is the best estimate of the true variance σ_y^2 and that the true variance of the mean is given by $\sigma_{\bar{y}}^2 = \sigma_y^2/N$. If σ_y is unknown and one uses the sample standard deviation s_y in its place, one obtains the sample standard deviation of the mean

$$s_{\bar{y}} = \frac{s_y}{\sqrt{N}} \tag{34}$$

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

The reduced chi-square pdf χ^2_ν

The sample variance is a random variable; each new sample set would give a new variance. However, rather than discuss the pdf for s_y^2 directly, it is more useful to introduce a normalized sample variance

$$\chi^2_\nu = \frac{s_y^2}{\sigma_y^2} \tag{35}$$

and switch the discussion to this quantity. This new quantity is a *reduced chi-square* random variable and its pdf will be written $p_\nu(\chi^2_\nu)$. With this normalization, the pdf for χ^2_ν becomes universal—independent of σ_y. If σ_y is not known, as is often the case, simply keep in mind it is a constant.

The subscript ν, used in labeling the reduced chi-square random variable and its pdf, is called the number of *degrees of freedom* and is equal to the sample size minus the number of parameters obtained from that sample set and used in the calculation of χ^2_ν. If $\bar{y}$ is determined first and then s_y^2 is determined using Eq. 30, a single derived parameter ($\bar{y}$) is used in calculating χ^2_ν and $\nu = N - 1$.

If a sample set is obtained over and over again (with N new y_i for each set), the resulting distribution of values for s_y^2 is obviously determined by the randomness of the y_i values. That is, $p_\nu(\chi^2_\nu)$ is completely determined by the Gaussian pdf governing the y_i. For example, properties of the Gaussian pdf were used in Exercise 6 to demonstrate that $\left\langle s_y^2 \right\rangle = \sigma_y^2$ for all values of N. Thus trivially, this exercise also proves

$$\left\langle \chi^2_\nu \right\rangle = 1 \tag{36}$$

for all values of ν.

The reduced chi-square pdf's are a class of pdf's distinguished by the integer ν. Several are shown in Fig. 3. Their dependence on ν should come as no surprise. Just as the pdf for $\bar{y}$ becomes more sharply peaked around μ_y as the sample size increases, so too the pdf for s_y^2 become more sharply peaked around σ_y^2 as ν increases. Consequently, $p_\nu(\chi^2_\nu)$ becomes more sharply peaked around the value

1 as ν increases. For large values of ν, $p_\nu(\chi^2_\nu)$ becomes approximately Gaussian-shaped with a mean of 1 and a standard deviation near $\sqrt{2/\nu}$. However, for low values of ν, $p_\nu(\chi^2_\nu)$ is asymmetric and fairly broad.

Exercise 7 *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 and evaluate the sample mean $\bar{y}$. For the uncertainty, you use the sample standard deviation of the mean $s_{\bar{y}}$. What is the minimum number of samples N needed if one is to be around 60% confident that $s_{\bar{y}}$ is within 10% of $\sigma_{\bar{y}}$. Hint: s_y will also have to be within 10% of σ_y. Thus, you want to find the value of N such that the probability $P(0.9\sigma_y < s_y < 1.1\sigma_y) = 0.6$. Convert this to a condition on χ^2_ν and then use Table 3.*

Student-T probabilities

Consider a sample mean $\bar{y}$ and a sample standard deviation of the mean $s_{\bar{y}}$ obtained from a sample set $y_i, i = 1..N$. These quantities are best estimates of the true mean μ_y and true standard deviation of the mean $\sigma_{\bar{y}}$ of the Gaussian pdf from which $\bar{y}$ is a sample. Now suppose one wishes to find a 68% or 95% confidence intervals for the unknown mean μ_y.

If a random variable y follows a Gaussian pdf, a y value in the range $\mu_y \pm \sigma_y$ occurs 68% of the time and a y value in the range $\mu_y \pm 2\sigma_y$ occurs 95% of the time. This logic is invertible and thus 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 value μ_y will likewise follow from the Gaussian pdf; 68% for $z = 1$, 95% for $z = 2$, etc. Such confidence intervals and associated probabilities are seldom reported because they are well known and completely specified once y and σ_y are given.

To find confidence intervals using a sample mean and a sample standard deviation of the mean, one can again express the interval in the form

$$\bar{y} \pm z s_{\bar{y}}$$

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

A difference arises because $s_{\bar{y}}$ might, by chance, come out larger or smaller than $\sigma_{\bar{y}}$; recall s^2_y/σ^2_y values are distributed according to the reduced chi-square pdf. When the probabilities for all possible values of $s_{\bar{y}}$ are properly taken into account, the confidence level for any z is always smaller than would be predicted based on the Gaussian pdf.

In effect, the uncertainty in how well $s_{\bar{y}}$ estimates $\sigma_{\bar{y}}$ decreases the confidence level for a given z. Because this uncertainty depends on the sample size N, the Student-T confidence intervals also depend on N (via the number of degrees of freedom $\nu = N - 1$). The larger the ν, the better the estimate and the closer the Student-T intervals will be to the corresponding Gaussian intervals.

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

a 95% confidence interval would be, as for a Gaussian, $\bar{y} \pm 2\, s_{\bar{y}}$.

Exercise 8 *Three sample values from a Gaussian pdf are 1.20, 1.24, and 1.19. (a) Find the (sample) mean, standard deviation, and standard deviation of the mean. Give the 68% and 95% confidence intervals for the true mean based on this data alone. (b) Now assume those three sample values are known to come from a pdf with a standard deviation $\sigma_y = 0.02$. With this assumption, what are the 68% and 95% confidence intervals? Hint: determine the standard deviation of the mean of three numbers with this σ_y. (c) Determine the reduced chi-square for this data set (assuming $\sigma_y = 0.02$) and the probability it would be this big or bigger.*

Correlation

Generally, when there are several measured variables in an experiment, they will be related. Indeed, the typical objective of experimentation is to discover a dependence between variables or verify a particular mathematical relationship. Nonetheless, more often than not, the measurements are correctly termed independent thereby indicating a statistical independence between the measured quantities. Statistical independence between two quantities implies that their random errors are independent. The quantities themselves may or may not be physically related. For example, the length L and period T of a simple pendulum are not physically independent; they are related by $T = 2\pi\sqrt{L/g}$, where g is the local acceleration due to gravity. However, the random errors in measurements of the pendulum's length and period will be unrelated and thus L and T would be statistically independent.

While statistical independence is generally the rule for measured quantities, there are the occasional exceptions. For example, simultaneously measured voltages are prone to pickup of common random errors. On the other hand, statistical independence is the exception rather that rule for two quantities derived from a single set of measurements, even if the measurements themselves are statistically independent. For example, the slope and intercept obtained from a linear regression analysis are random variables, and are not, in general, statistically independent. In this section, the quantitative measures of statistical dependence, correlation and covariance are defined. Later sections will demonstrate some effects of correlation.

Consider a sampling experiment in which there are two variables, x and y. The variables are always in one-to-one correspondence so that a single sample consists of an ordered x, y pair. The pairs are sampled repeatedly to make the complete sample set of ordered pairs x_i, y_i, $i = 1..N$. The sample set is taken under unchanging experimental conditions and thus only random variations are expected.

The analysis begins by considering the data as two separate sample sets: x_i, $i = 1..N$ and y_i, $i = 1..N$. We assume each set is consistent with a Gaussian parent. The sample means $\bar{x}$ and $\bar{y}$ and the sample variances s_x^2 and s_y^2 are computed. These are, respectively, best estimates for the means μ_x and μ_y and variances σ_x^2 and σ_y^2 for each variable's parent.

Consider first the limiting case of total positive correlation. In this case, if an x_i value is above (or below) $\bar{x}$ by the amount s_x, the corresponding y_i value will be above (or below) $\bar{y}$ by the amount s_y. More generally, for any x_i, y_i in the data set

$$\frac{x_i - \bar{x}}{s_x} = \frac{y_i - \bar{y}}{s_y}$$

Note that, looked at separately, each variable can still appear to be randomly varying ac-

cording to a Gaussian parent, but the variations are not independent, they are totally correlated. There is really only one random variation which completely determines both the x_i and y_i variations.

With less than total correlation, one variable's value only partially predicts the other. There remains an independent random component in addition to the predictive, or correlated, component. With negative correlation, total or partial, the predictive components become oppositely signed; when one variable in a pair is above its mean, the predictive component of the other is below its mean. With no correlation, there is no predictive component; knowing the value of one member of the pair provides no information about the other. This is the case of statistical independence.

Quantitatively, correlation is described by the *correlation coefficient* or equivalently by the *covariance*. The *sample covariance* s_{xy} is defined

$$s_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{37}$$

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

$$\sigma_{xy} = \lim_{N\to\infty}\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{38}$$

or equivalently as the expectation value

$$\sigma_{xy} = \langle (x - \mu_x)(y - \mu_y)\rangle \tag{39}$$

Consider an x_i, y_i pair chosen randomly from an infinite sample. Whether or not the variables are correlated, the x-value (or y-value) is equally likely to be above or below its mean, and thus the value of $x_i - \bar{x}$ (or $y_i - \bar{y}$) is equally likely to be positive or negative. But only if the variables are uncorrelated will the product $(x_i - \bar{x})(y_i - \bar{y})$ have equally probable positive and negative values and an expectation value of zero. With a positive correlation this product has a positive expectation value and with a negative correlation it has a negative expectation value. Consequently, the covariance $\sigma_{xy} = \langle (x_i - \bar{x})(y_i - \bar{y})\rangle$ will be zero, positive, or negative, respectively, depending on whether there is a no correlation, positive, or negative correlation.

The covariance σ_{xy} is limited by the size of σ_x and σ_y and can vary between

$$-\sigma_x\sigma_y \le \sigma_{xy} \le \sigma_x\sigma_y \tag{40}$$

Thus, σ_{xy} is also often written

$$\sigma_{xy} = \rho\sigma_x\sigma_y \tag{41}$$

where ρ, called the *correlation coefficient,* is between -1 and 1.

The inequality expressed by Eq. 40 is also true for the sample standard deviations and sample covariance with the substitution of s's for σ's. The sample correlation coefficient r is then defined $s_{xy} = r s_x s_y$ and also varies between -1 and 1.

It is common to define the (symmetric) *covariance matrix* denoted $[\boldsymbol{\sigma}]$ to describe all the variances and covariances possible between two or more variables. For 3 variables x, y, z it would be

$$[\boldsymbol{\sigma}] = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \tag{42}$$

with the extension to more variables obvious. Note that $\sigma_{xx} = \sigma_x^2$ is the variance of x with similar relations for σ_{yy} and σ_{zz}

Propagation of Errors

One is often interested in a quantity derived from one or more measurements $y_1, y_2, ..., y_M$. For example, y_1 might represent a measurement value for the voltage V across a circuit

element, while y_2 might represent a measurement value for the current I through the element. The derived quantity might be the element's resistance $R = V/I$ or the power dissipated $P = IV$.

The theoretical model doesn't really provide the relationship between the measurement *values* and the derived quantity. It is better to think of the model as providing the relationship between the measurement's parent *means* (true values of the physical quantities) and the true value of the derived quantity. Denoting this latter quantity α, the theoretical model provides the relation

$$\alpha = f(\mu_1, \mu_2, ..., \mu_M) \tag{43}$$

The measurement uncertainties (standard deviations) $\sigma_1, \sigma_2, ..., \sigma_M$ are assumed known and thus each y_i is a single sample from a Gaussian pdf of (unknown) mean value μ_i and known standard deviation σ_i.

The first objective is to find a best estimate of α (which will be denoted a) based on a single set of values $y_1, y_2, ..., y_M$. Since the data set probability is maximized if all means μ_i are taken equal to the measured values y_i, the Principle of Maximum Likelihood suggests that a is obtained by using y_i for μ_i in Eq. 43

$$a = f(y_1, y_2, ..., y_M) \tag{44}$$

As a is a random variable, consider the distribution of a values obtained by repeatedly evaluating $a = f(y_1, y_2, ..., y_M)$, each time from a newly sampled set of y_i's. As additional sample sets are taken, the y_i's would vary according to their individual Gaussian parents and thus the a's would vary as well—according to some new pdf.

Since each y_i is from a Gaussian parent, better than 99% of the time the value for any y_i will be in the range $\mu_i \pm 3\sigma_i$. We will assume all the σ_i are small enough that over these ranges, the function f is accurately represented by a first-order Taylor expansion about the values $\mu_1, \mu_2, ..., \mu_M$.

$$\begin{aligned} a &= f(y_1, y_2, ..., y_M) \\ &= f(\mu_1, \mu_2, ..., \mu_M) + \\ &\quad \frac{\partial f}{\partial y_1}(y_1 - \mu_1) + \frac{\partial f}{\partial y_2}(y_2 - \mu_2) \\ &\quad +... + \frac{\partial f}{\partial y_M}(y_M - \mu_M) \\ &= \alpha + \sum_{i=1}^{M} \frac{\partial f}{\partial y_i}(y_i - \mu_i) \end{aligned} \tag{45}$$

where Eq. 43 has been used in the final step. The partial derivatives are simply constants that should be evaluated at the expansion point $\mu_1, \mu_2, ..., \mu_M$. However, as the means are typically unknown, they will normally be evaluated at the measured point $y_1, y_2, ...y_M$. This should not be a problem because f is assumed to be linear over a range of several σ_i about each μ_i.

If keeping only up to the linear terms (Eq. 45) is a satisfactory approximation, then a's pdf will be Gaussian. A small amount of curvature will not usually cause a problem, but f should not have significant nonlinearities over a range of about $\pm 3\sigma_i$ in each variable.

Within the linear approximation, the mean and standard deviation of a's pdf can easily be determined. The mean is defined by Eq. 13 ($\mu_a = \langle a \rangle$) and the expectation value is easily evaluated from Eq. 45

$$\begin{aligned} \mu_a &= \langle a \rangle \\ &= \left\langle \alpha + \sum_{i=1}^{M} \frac{\partial f}{\partial y_i}(y_i - \mu_i) \right\rangle \\ &= \langle \alpha \rangle + \sum_{i=1}^{M} \frac{\partial f}{\partial y_i} \langle (y_i - \mu_i) \rangle \\ &= \alpha \end{aligned} \tag{46}$$

where we have used the expectation values $\langle y_i - \mu_i \rangle = 0$ to eliminate all terms in the sum.

Thus, we have demonstrated the important result that if a is taken as $f(y_1, y_2, ..., y_M)$, it will "on average" be the true value α.

The standard deviation is determined via Eq. 14 ($\sigma_a^2 = \langle (a - \mu_a)^2 \rangle$). Using Eq. 45 for a and canceling μ_a ($= \alpha$), the expectation value becomes

$$\sigma_a^2 = \left\langle \left(\sum_{i=1}^{M} \frac{\partial f}{\partial y_i} (y_i - \mu_i) \right)^2 \right\rangle \qquad (47)$$

Expanding the square of the sum into two sums, one including the quadratic terms and one including the cross terms gives

$$\sigma_a^2 = \left\langle \sum_{i=1}^{M} \left(\frac{\partial f}{\partial y_i} (y_i - \mu_i) \right)^2 + 2 \sum_{i>j=1}^{M} \frac{\partial f}{\partial y_i} \frac{\partial f}{\partial y_j} (y_i - \mu_i)(y_j - \mu_j) \right\rangle \qquad (48)$$

where the second sum is over all pairs i, j where $i > j$. (The factor of 2 arises because Eq. 47 would produce two equivalent cross terms, while the sum includes each cross term only once.)

Now we can use the definitions for the variances of σ_i^2

$$\sigma_i^2 = \langle (y_i - \mu_i)^2 \rangle \qquad (49)$$

and the definition of the covariance between variables

$$\sigma_{ij} = \langle (y_i - \mu_i)(y_j - \mu_j) \rangle \qquad (50)$$

to express σ_a^2 as

$$\sigma_a^2 = \sum_{i=1}^{M} \left(\frac{\partial f}{\partial y_i} \right)^2 \sigma_i^2 + 2 \sum_{i>j=1}^{M} \frac{\partial f}{\partial y_i} \frac{\partial f}{\partial y_j} \sigma_{ij} \qquad (51)$$

In the typical situation, only a single set of measurements $y_1, y_2, ., y_M$ and standard deviations (uncertainties) $\sigma_1, \sigma_2, ., ., \sigma_M$ are available. The quantity $a = f(y_1, y_2, ., y_M)$ is then a single sample from a Gaussian parent having a standard deviation (uncertainty) σ_a—obtained through Eq. 51—and having an unknown mean which will be the true value α of the sought-after quantity. As such, a has much the same interpretation as a single sample or measurement value.

Exercise 9 *Assume there are two derived quantities u_1 and u_2 obtained from measurements y_a and y_b according to the relations*

$$u_1 = f_1(y_a, y_b) \qquad (52)$$
$$u_2 = f_2(y_a, y_b) \qquad (53)$$

Further assume y_a and y_b come from correlated Gaussian pdf's having variances σ_a^2 and σ_b^2 and a covariance σ_{ab}^2 and that the functions f_1 and f_2 are approximately linear over the range of y's likely to occur. Show that the covariance between u_1 and u_2 is given by

$$\sigma_{12} = \frac{\partial f_1}{\partial y_a} \frac{\partial f_2}{\partial y_a} \sigma_a^2 + \frac{\partial f_1}{\partial y_b} \frac{\partial f_2}{\partial y_b} \sigma_b^2 + \left(\frac{\partial f_1}{\partial y_a} \frac{\partial f_2}{\partial y_b} + \frac{\partial f_1}{\partial y_b} \frac{\partial f_2}{\partial y_a} \right) \sigma_{ab} \qquad (54)$$

Note that σ_{12} will generally be non-zero whether or not $\sigma_{ab} = 0$.

Linear Regression

Consider the kind of experiment in which one measures a dependent or responding variable y, as an independent or manipulated variable x is systematically varied.[3] The data set consists of corresponding pairs (x_i, y_i), $i = 1...N$.

[3] There can be more than one independent variable. Although only a single x-variable is treated explicitly, most of the regression equations and ideas would remain the same if the variable x were considered to be a set of independent variables.

Often, theoretical considerations predict a relationship between x and y that can be expressed by a function (called the fitting function)

$$y = F(x) \tag{55}$$

where $F(x)$ includes M unknown constants α_k, $k = 1...M$ called fitting parameters. For example, $F(x) = \alpha_1 + \alpha_2 x$ would specify a straight line fitting function with two unknown fitting parameters—an intercept α_1 and a slope α_2. Regression analysis is used to obtain best estimates of the fitting parameters from the data set and assess whether the data are consistent with the fitting function.

Linear regression is used when $F(x)$ can be expressed as linear function of the fitting parameters

$$\begin{aligned} F(x) &= \alpha_1 f_1(x) + \alpha_2 f_2(x) + ... + \alpha_M f_M(x) \\ &= \sum_{k=1}^{M} \alpha_k f_k(x) \end{aligned} \tag{56}$$

where the f_k are linearly independent functions of x with no unknown parameters. Note that the $f_k(x)$ may be nonlinear functions of x, e.g., x^2 or $\sin(x)$. The fitting function is still linear in the coefficients α_k as long as the α_k only appear as coefficients (amplitudes) of the $f_k(x)$.

The x_i, y_i pairs will be assumed to be statistically independent, i.e., have zero covariances. Although both x_i and y_i will often have random errors, we will first consider the simpler case where the x_i are known exactly and only the y_i have random errors. The expected (rms) size of the random error associated with each particular y_i value will be represented by σ_i. The σ_i are assumed known (the uncertainties of the y_i) and may be the same for all data points or they may vary from point to point.

The physical model is not really $y = F(x)$, but rather that $F(x_i)$ gives the mean μ_i of the parent Gaussian from which each y_i is a sample.

$$\mu_i = F(x_i) \tag{57}$$

The probability dP_i of obtaining a measured y_i in a small range dy_i can then be written

$$dP_i = \frac{dy_i}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - F(x_i))^2}{2\sigma_i^2}\right] \tag{58}$$

If there are N such pairs of data points, the probability of all of them occurring is a product of N such terms. The best estimates of the α_k's are those values a_k that maximize this product probability, i.e., that minimize the chi-square

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - F(x_i))^2}{\sigma_i^2} \tag{59}$$

This χ^2 is an implicit function of the α_k that would appear in $F(x_i)$ and is minimized by setting the M partial derivatives of χ^2 with respect to each α_k (evaluated at $\{\alpha\} = \{a\}$) equal to zero.

$$\left.\frac{\partial \chi^2}{\partial \alpha_k}\right|_{\{\alpha\}=\{a\}} = 0 \tag{60}$$

These M simultaneous equations can then be solved for the M unknown a_k's.

Let's see how this works for a two parameter fit: $F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$. The treatment is similar for more than two parameters.

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - (\alpha_1 f_1(x_i) + \alpha_2 f_2(x_i)))^2}{\sigma_i^2} \tag{61}$$

Setting the partial derivatives of χ^2 with respect to α_1 and α_2 equal to zero (at $\alpha_1 = a_1$ and $\alpha_2 = a_2$) leads to the following two equations, respectively

$$\sum_{i=1}^{N} \frac{y_i f_1(x_i)}{\sigma_i^2} = a_1 \sum_{i=1}^{N} \frac{f_1(x_i) f_1(x_i)}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{f_1(x_i) f_2(x_i)}{\sigma_i^2} \quad (62)$$

$$\sum_{i=1}^{N} \frac{y_i f_2(x_i)}{\sigma_i^2} = a_1 \sum_{i=1}^{N} \frac{f_1(x_i) f_2(x_i)}{\sigma_i^2} + a_2 \sum_{i=1}^{N} \frac{f_2(x_i) f_2(x_i)}{\sigma_i^2} \quad (63)$$

Defining the following quantities—all of which can be obtained from the known quantities x_i, y_i, σ_i, $i = 1...N$, and the known functions f_1 and f_2

$$Y_1 = \sum_{i=1}^{N} \frac{y_i f_1(x_i)}{\sigma_i^2} \quad (64)$$

$$Y_2 = \sum_{i=1}^{N} \frac{y_i f_2(x_i)}{\sigma_i^2} \quad (65)$$

$$x_{11} = \sum_{i=1}^{N} \frac{f_1(x_i) f_1(x_i)}{\sigma_i^2} \quad (66)$$

$$x_{12} = \sum_{i=1}^{N} \frac{f_1(x_i) f_2(x_i)}{\sigma_i^2} \quad (67)$$

$$x_{22} = \sum_{i=1}^{N} \frac{f_2(x_i) f_2(x_i)}{\sigma_i^2} \quad (68)$$

we can rewrite Eqs. 62 and 63 in matrix form

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (69)$$

In this form the solution is easily found by the method of determinants

$$a_1 = \frac{1}{D} \begin{vmatrix} Y_1 & x_{12} \\ Y_2 & x_{22} \end{vmatrix} \quad (70)$$

$$a_2 = \frac{1}{D} \begin{vmatrix} x_{11} & Y_1 \\ x_{12} & Y_2 \end{vmatrix} \quad (71)$$

where D is the determinant

$$D = \begin{vmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{vmatrix} \quad (72)$$

Alternatively, vector algebra can be used. The α's, a's, and Y's can be written as column vectors

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \quad (73)$$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (74)$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad (75)$$

and a matrix $[\mathbf{X}]$ can be defined

$$[\mathbf{X}] = \begin{pmatrix} x_{11} & x_{12} \\ x_{12} & x_{22} \end{pmatrix} \quad (76)$$

so that Eq. 69 can be put in the matrix form

$$\mathbf{Y} = [\mathbf{X}]\mathbf{a} \quad (77)$$

This equation is solved by finding the inverse matrix $[\mathbf{X}]^{-1}$ and multiplying the equation above by $[\mathbf{X}]^{-1}$ on the left

$$\mathbf{a} = [\mathbf{X}]^{-1}\mathbf{Y} \quad (78)$$

The a's are random variables. If we were to take a whole new set of measurements (samples) of the y_i (at the previous values for the x_i) and then recalculate the a's for this set, we should expect them to change. If we did this repeatedly, we would find the a's occur with probabilities governed by their own pdf's. For

a single data set, we don't expect the a's to be exactly equal to the α's, but we should expect this to be true on average, i.e.,

$$\langle \mathbf{a} \rangle = \boldsymbol{\alpha} \tag{79}$$

The proof can be found in the addendum.

Since the a's can vary (were repeated data sets taken) we should also be interested in the variances of the pdf's for the a's. The variances are defined by

$$\sigma_{kk} = \sigma_k^2 = \langle (a_k - \alpha_k)^2 \rangle \tag{80}$$

(since the mean value of $a_k = \alpha_k$). Since the a's will generally be correlated, we should also be interested in the covariance between the a's, which are defined by

$$\sigma_{km} = \langle (a_k - \alpha_k)(a_m - \alpha_m) \rangle \tag{81}$$

for $k \neq m$.

The covariance matrix is used to express these quantities

$$[\boldsymbol{\sigma}] = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \tag{82}$$

It can be shown that

$$[\boldsymbol{\sigma}] = [\mathbf{X}]^{-1} \tag{83}$$

In other words, $[\mathbf{X}]^{-1}$ is the covariance matrix and gives the parameter variances (diagonal elements) and covariances (off-diagonal elements). The proof can also be found in the addendum.

Take careful note of the fact that the off-diagonal elements will normally be non-zero. Non-zero covariances imply the parameters will be correlated, and correlated parameters imply the propagated uncertainty in the value of any expression that includes more than one of the fitted parameters will need to take into account the covariance terms.

Uncertainties in Both Variables

When there are uncertainties σ_{x_i} and σ_{y_i} in both variables, a reasonable procedure is to use the following σ_i^2 in the fitting procedure

$$\sigma_i^2 = \sigma_{y_i}^2 + \left(\frac{dF}{dx}\right)^2 \sigma_{x_i}^2 \tag{84}$$

where dF/dx is evaluated at x_i using the best fit parameters from the regression. Consider that σ_{y_i} specifies the rms size of the random y-variations that can be expected for the point i, with a similar interpretation for σ_{x_i}. Because dF/dx is the slope of the fitting function at the ith data point, a random variation in the x-direction of the amount σ_{x_i} would cause an effective random variation in the y-direction of the amount $(dF/dx)\sigma_{x_i}$. The actual and effective y-variations would be independent and are thus added in quadrature to get the net expected y-variations σ_i.

Of course, the fitting parameters of F would need to be known to evaluate the derivatives dF/dx. Thus an iterative approach would be used, perhaps starting with a fit neglecting the x-uncertainties. After an initial parameter set is obtained, the fit would be repeated and new a_k would be found using the net σ_i as given by Eq. 84.

Equally weighted linear regression

Excel's regression program does not let the user provide the σ_i values that would ordinarily be needed to apply the linear regression equations (Eqs. 64-83). The Excel regression program assumes the σ_i are the same for all data points and thus its results are only valid to the extent this assumption is satisfied. Such a regression would be called equally weighted.

Exercise 10 *A data set $(x_i, y_i), i = 1...N$ will be used in a linear regression. Assume the σ_i*

are the same for all y_i and that the x_i are known exactly. (Again, we use $\sigma_i = \sigma_y$ when they are all equal.) Use the linear regression equations to explain (a rigorous proof is not required): (a) why the fitting parameters a_k are independent of the value for σ_y and (b) why the parameter standard deviations σ_k are proportional to σ_y. Hint: Refer to Eq. 78 for part (a) and to Eq. 83 for part (b). Also refer to the defining equations for the elements of $[\mathbf{X}]$ *and* $\mathbf{Y}$ *(Eqs. 64-68) to explain the dependence (or independence) on σ_y.*

Property (a) implies that the data point uncertainties need not be known to determine the fit parameters. This is why Excel does not need σ_y to get the a_k. And while property (b) implies σ_y must be known in order to get the fitting parameter uncertainties σ_k, the deviations between the data and the fit can be used to get a best estimate of σ_y. Excel uses this best estimate for σ_y in the regression equations to get best estimates of the σ_k.

The best estimate of σ_y will again be denoted s_y and it can be obtained using

$$s_y^2 = \frac{1}{N-M}\sum_{i=1}^{N}\left(y_i - F(x_i)\right)^2 \qquad (85)$$

Compare this with Eq. 30 for a sample set from a single pdf. In both cases s_y is an rms deviation between the data point y_i-values and their corresponding best estimates ($F(x_i)$ for Eq. 85, $\overline{y}$ for Eq. 30). And in both cases the divisor is the number of degrees of freedom ($N - M$ for Eq. 85, $N - 1$ for Eq. 30). Thus, the s_y of Eq. 85 is a sample standard deviation and a best estimate of σ_y were σ_y unknown. One can even show that $\langle s_y^2 \rangle = \sigma_y^2$.

In effect, a regression like Excel's first determines the fitting parameters a_k, then uses the a_k, $f_k(x_i)$, and y_i to determine s_y, and then uses s_y as the best estimate of σ_y to determine the fitting parameter uncertainties. Because they are obtained using a best estimate

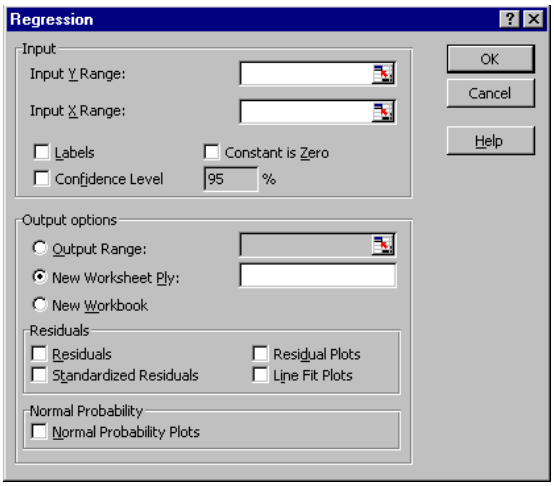


Figure 4: Excel's linear regression dialog box

s_y instead of the true σ_y, the fitting parameter uncertainties are really best estimates s_k of the true σ_k. They are sample standard deviations rather than true standard deviations. Accordingly, the probability that an interval $a_k \pm z\ s_k$ will include the true value α_k is governed by Student-T probabilities with $N - M$ degrees of freedom.

Linear regression with Excel

The starting point for using the Excel linear regression program is to set up columns for the y_i and the $f_k(x_i)$. The steps will be illustrated for a quadratic fit: $F(x_i) = \alpha_1 + \alpha_2 x_i + \alpha_3 x_i^2$. Thus, $f_1(x_i) = 1$, $f_2(x_i) = x_i$, and $f_3(x_i) = x_i^2$. In turns out to be unnecessary to have a column for the constant term; Excel can handle the constant function without using a column. There will need to be columns for y_i, x_i and x_i^2.

Excel's linear regression is found in the Tools|Data Analysis|Regression menu. The dialog box for this procedure appears as in Fig. 4.

Select the column containing the y_i-values for the Input Y-Range. Select what must be

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

Table 1: Data for Exercise 11

a contiguous rectangular block containing all $f_k(x_i)$ values (two columns for the quadratic fit). Leave the Constant is Zero box unchecked. (It would be checked if the fitting function did not include a constant term.) Leave the Labels box unchecked unless your x- and y-ranges include labels at the top of the columns. If you would like, check the Confidence Level box and supply a probability for a Student-T interval next to it. (Intervals for the 95% confidence level are provided automatically.) Select the New Worksheet Ply radio button or the Output Range. For the latter, also specify (in the edit box to its right) the upper left corner of an empty spreadsheet area for the results. Select any of the various options in the lower part of the dialog box and then click OK.

The upper *Regression Statistics* area contains the sample standard deviation s_y (Eq. 85) (referred to as the Standard Error). The lower area contains information about the constant (labeled *Intercept*) and the fitting parameters a_k (labeled X Variable k. Next to the best fit values (labeled *Coefficients*) are the parameter sample standard deviations s_k (labeled *Standard Error*). Then there are columns for the *t Stat* and *P-value* for each parameter. And lastly two double columns for the lower and upper limits of intervals at confidence levels of 95% and the user specified percentage.

Exercise 11 *Fit the data in Table 1 to a quadratic formula ($y = a + bx + cx^2$) using the Excel linear regression program and submit a printout of the appropriate worksheets. Save this spreadsheet. It will be used again for Exercises 12 and 13. (a) What does Excel use as a best estimate for the value of σ_i appearing in the linear regression equations? Describe and give the formula for this quantity. Circle and label its value on the worksheet. (b) The parameter standard deviations given on the worksheet are sample standard deviations s_k. Circle and label the parameters a_k and their sample standard deviations s_k. (c) Circle the program's 95% confidence interval (for the quadratic coefficient only) and show how it can be obtained from a_k, s_k and the Student-T table.*

Considerations if σ_y is known

In this lab, you will typically be expected to make an estimate of σ_y based on the measurement procedures and equipment. So suppose σ_y is known. How would this affect the results and interpretation of a regression, such as Excel's, that uses s_y for σ_y?

According to property (a) of Exercise 10, no adjustments are needed for the parameters a_k because their values do not depend on the value used for σ_y. Consequently, the fit and the rms deviation s_y would be unaffected.

However, the parameter uncertainties would need to be adjusted. According to property (b) of Exercise 10, the fitting parameter uncertainties σ_k are proportional to the data point uncertainties σ_y. Since the linear regression uses s_y when the proper value is σ_y, the scaling property implies they must be multiplied by σ_y/s_y.

Moreover, after scaling them this way, they should no longer be considered best estimates or sample standard deviations s_k. They

should be considered true standard deviations, σ_k. Consequently, the confidence intervals $a_k \pm zs_k$ calculated according to the Student-T probabilities become invalid and should be replaced by intervals $a_k \pm z\sigma_k$ calculated according to Gaussian probabilities.

Exercise 12 *Return to the spreadsheet for Exercise 11. (a) Create a column for* $(y_i - F(x_i))^2$ *(using the fitting parameters) and then use this column to evaluate* s_y*. Show that it agrees with Excel's value. (b) Suppose the uncertainties in the* y_i *values for that exercise are known to be* $\sigma_y = 0.5$*. Use the scaling rule to determine the parameter uncertainties in this case. Give the 95% confidence interval for the quadratic coefficient. Should you use Student-T or Gaussian probabilities?*

Nonlinear Regression

Linear regression techniques can only be used when the fitting function can be put in the form of Eq. 56, where $dF/d\alpha_k$ is independent of the entire set of fitting parameters α. Many fitting functions cannot be so expressed. For example, a common nonlinear fitting function is an exponential decay with a constant offset.

$$F(x) = \alpha_1 + \alpha_2 \exp(-\alpha_3 x) \qquad (86)$$

Nonlinear regression techniques would be needed to find the three fitting parameters in this function for a given data set x_i, y_i, $i = 1...N$.

Nonlinear regression is similar to linear regression in that the fitting parameters are still taken to be those which minimize the χ^2 of the data set and thus maximize the data set probability; nonlinear regression is still a least squares method based on the principle of maximum likelihood.

Nonlinear regression programs repeatedly "try out" various possible sets of fitting parameters, evaluating the χ^2 value each parameter set produces. Consult the references to learn about the various "search" algorithms used to find the fitting parameters which minimize the χ^2. We will not be concerned with that aspect of the programs. Spreadsheet programs and other data analysis packages often have such algorithms (Solver in Excel). Here, we would like to describe how to use the Excel Solver to accomplish the task of finding the best estimates of the fitting parameters and, if needed, their variances and covariances.

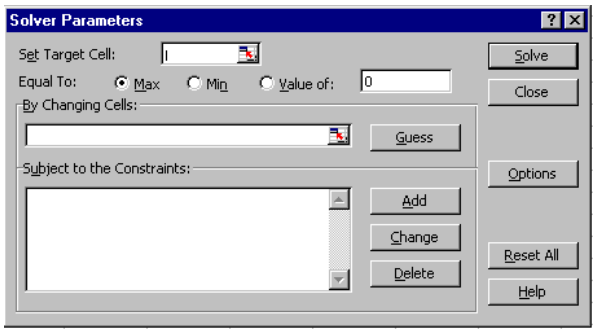


Figure 5: Excel's Solver dialog box

To use Solver:

1. Set up an area of the spreadsheet for the fitting parameters. The area need not be contiguous, but using Solver will be somewhat easier if parameters are confined to a single block of cells. Enter initial guesses for the values of each fitting parameter.

2. Enter your data in columns, including x_i and y_i and their uncertainties σ_{x_i} and σ_{y_i}. If either σ_{x_i} and/or σ_{y_i} is the same for all data points, the value can be placed in a single cell.

3. Create a column for $F(x_i)$. This will be the fitted y-values based on the theoretical formula using the addresses for the x_i and the a_k fitting parameters.

4. Construct a column for σ_i. These are the data point uncertainties constructed according to Eq. 84, based on the derivative of the theoretical formula, and using the cell addresses of the x_i-values, the a_k fitting parameters, and the measurement uncertainties σ_{x_i} and σ_{y_i}. If σ_i is the same for all data points, it can be assigned to a single spreadsheet cell.

5. Make a column for $(y_i - F(x_i))^2/\sigma_i^2$. These are the data point contributions to the chi-square. Sum this column in a separate cell to provide the χ^2 value needed for the fitting procedure.

6. Invoke the Solver from the Tools menu. The dialog box is shown in Fig. 5.

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

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

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

10. Accept the final, changed, fitting parameters.

Parameter variances and covariances are not provided by the Excel Solver program, but can be calculated in subsequent steps. The guiding principle is quite simple. If you offset a fitting parameter by its uncertainty, i.e., from a_k to $a_k + \sigma_k$, and if all other fitting parameters are then reoptimized, the χ^2 will change by unity. Based on this principle, the following procedure can be used to find the fitting parameter uncertainties.

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

12. Change the value in the cell containing one of the fitting parameters, say a_k, by a bit—to a new (unoptimized) value we will call a'_k. The χ^2 will increase because it was originally at a minimum. The cell containing a_k is then removed from the list of adjustable parameters and the Solver is rerun. The other parameters might change a bit and the χ^2 might go down a bit, but it will still be larger than the original χ^2 for the fit.

13. If the change in χ^2 is more (or less) than unity, the tested change in a_k is too big (or too small). Repeat the procedure with a value of a'_k which is closer to (or farther from) the best value a_k.

14. Repeat until the final χ^2 has increased by unity at which point the amount by which a_k was changed is its uncertainty σ_k. That is, when you have found the a'_k for which the χ^2 increases by one, you then have $\sigma_k = |a'_k - a_k|$.

15. If you will want the covariances between parameters, record also the new optimized values for the other parameters which we will label a^k_m indicating the new values for the a_m as optimized for $a'_k = a_k + \sigma_k$.

16. Repeat the whole procedure for all fitting parameters. The covariances can then be

calculated from the following

$$|\sigma_{km}| = \sqrt{|(\sigma_k \sigma_m (a_m^k - a_m)(a_k^m - a_k)|} \tag{87}$$

The sign of σ_{km} is the sign of $(a_k' - a_k)(a_m^k - a_m)$, i.e., if after increasing a_k one finds that a_m increases (decreases) upon re-minimizing the χ^2, σ_{km} is positive (negative).

Cautions

One must be wary in case the search algorithm finds its way into a local χ^2 minimum rather than an absolute minimum. The algorithm may even fail completely if the initial guesses for the fitting parameters are not close enough to the optimum values, and one may have to adjust the initial parameter guesses to get a solution.

Depending on how the algorithm varies the fitting parameters in its search to find the χ^2 minimum, it may be important to have the magnitude of all parameters near unity. This is often a good practice in any case. Thus, if the amplitude of an exponential decay is of order 10^6, the decay constant is of order 10^{-3}, and the background is of order 10^3, rather than perform the fit directly to Eq. 86 it would be wiser to fit to

$$F(x) = 10^3 \cdot \alpha_1 + 10^6 \cdot \alpha_2 \exp(-10^{-3} \cdot \alpha_3 x)$$

so that all fitting parameters are of order unity.

Exercise 13 *While linear regression can only be used for linear models, nonlinear techniques can be used for both linear and nonlinear models. In this exercise, you are to use the nonlinear regression techniques just described to fit the data of Table 1 to a quadratic formula. This is the same data and model used for Exercises 11 and 12 and you will be asked to show that the same results are obtained.*

Start with the spreadsheet from Exercise 12. Add a cell for σ_y and reference this cell in the spreadsheet (anywhere σ_i is needed) so you can change the value in that cell and have it updated throughout the spreadsheet.

(a) Start by setting σ_y to 2.0 and then to 3.0 running the solver for each value. Demonstrate that the a_k do not depend on the value used for σ_y; that in both cases the optimized a_k are the same as those from Exercise 11. This demonstrates Exercise 10 property (a). Can you explain why the algorithm should behave this way? (b) Now, assume σ_y is unknown and use the value of s_y for σ_y. Recall, this is what the Excel regression program does. What value of χ^2 does this produce? What value should be expected? Use the numerical method described in this section for determining parameter uncertainties to determine the standard deviation of the quadratic coefficient only. Show it is the same as that obtained by Excel's linear regression. (c) Now assume $\sigma_y = 0.5$. This value was used in Exercise 12, but in that exercise you were asked to use scaling arguments to determine how this affected the uncertainty in a_3. Here you will show that that scaling was correct. Use the numerical method to redetermine the uncertainty of a_3 with $\sigma_y = 0.5$. Now you have two values for the uncertainty in a_3, one using $\sigma_y = s_y$ and one using $\sigma_y = 0.5$. Show that they scale in proportion to the value used for σ_y. This demonstrates Exercise 10 property (b).

Evaluation of a Fit

The first evaluations of the agreement between a fitting function and a data set should be performed graphically. Typically, the best fit function $y = F(x)$ is plotted vs. x as a smooth curve (no markers) and overlayed with the x_i, y_i data points (markers, no curve). It may also be helpful to include error bars on these

graphs—lines extending $\pm\sigma$ about each point in either or both the x- and y-directions. The fitting function should pass through roughly 68% of the data points (including their error bars).

Deviations between the fit $F(x_i)$ and the data y_i should be random. Deviations which vary smoothly with x would be indicative of systematic errors. Graphing the deviations (also called residuals) $y_i - F(x_i)$ vs. x_i can be helpful in clarifying whether the deviations are random or systematic.

The chi-square test

In this section, it will be important to distinguish between the true theoretical function $F_\alpha(x)$ using the true parameters α_k and the best fit $F_a(x)$ using the best estimates a_k obtained from the fit.

Recall σ_i is the expected rms deviation between the measured y_i value and the mean of the distribution from which y_i is a sample. According to the theory, this mean is $F_\alpha(x_i)$ and thus the expectation can be expressed

$$\Big\langle (y_i - F_\alpha(x_i))^2 \Big\rangle = \sigma_i^2 \qquad (88)$$

Of course, any particular value of $(y_i - F_\alpha(x_i))^2$ might be larger or smaller than σ_i^2; it should only be σ_i^2 "on average."

In order to treat the general case where the σ_i are not the same for all points, this equation needs to be rewritten:

$$\left\langle \frac{(y_i - F_\alpha(x_i))^2}{\sigma_i^2} \right\rangle = 1 \qquad (89)$$

Equation 89 is true for every data point and so a sum of this quantity over all N data points should have an expectation value of N.

$$\left\langle \sum_{i=1}^{N} \frac{(y_i - F_\alpha(x_i))^2}{\sigma_i^2} \right\rangle = N \qquad (90)$$

Note that the term inside the expectation values is the chi-square of Eq. 59.

Normally, $F_\alpha(x)$ is unknown and it is only $F_a(x)$ that is determined by the fit. How does this affect things? Is Eq. 90 valid if $F_a(x_i)$ is substituted for $F_\alpha(x_i)$? Recall that the fitted values for a_k always minimize the chi-square. Thus, for the data set that produced it, $F_a(x_i)$ will always give a lower value of χ^2 than would $F_\alpha(x_i)$. How much lower? Here is the "on average" answer:

$$\left\langle \sum_{i=1}^{N} \frac{(y_i - F_a(x_i))^2}{\sigma_i^2} \right\rangle = N - M \qquad (91)$$

The expectation value of the χ^2 falls by the number of fit parameters M.

The proof of Eq. 91 is tedious (it is a more difficult version of Exercise 6b) and so we will simply illustrate its validity for one particular case. Suppose experimental x_i, y_i data points are predicted to lie on a straight line. If the line (slope and intercept, $M = 2$) is known ahead of time (i.e., $F_\alpha(x)$ is known), the points would not be expected to fall exactly on that line. On average, random error would scatter each point one-σ_i from the line. Were there only two points in the complete data set (a rather small data set, but perfect to make our argument) the χ^2 obtained using $F_\alpha(x)$ would have an expectation value of 2. Obviously, if the two points are fit to a straight line, the line will pass exactly through those points. That is, $y_i = F_a(x_i)$, and thus $\chi^2_\nu = 0$—in agreement with Eq. 91.

One can show that the quantity

$$\chi^2_\nu = \frac{1}{N-M} \sum_{i=1}^{N} \frac{(y_i - F_a(x_i))^2}{\sigma_i^2} \qquad (92)$$

is a reduced chi-square random variable with

$N - M$ degrees of freedom.[4] and Eq. 91 implies that it has the proper expectation value. Of course, the actual χ^2_ν value obtained from a fit should have a probability of occurrence governed by a reduced chi-square pdf, and the chi-square test is a simple check to see whether this χ^2_ν value is reasonable.

Appreciating the meaning of a given χ^2_ν is often more important than the chi-square test itself. A value of χ^2_ν near one would indicate that the data points on average (in an rms sense) miss the fit by the expected amount—one-σ_i. A value of $\chi^2_\nu = 4$ would indicate the data points on average miss the fit by $2\sigma_i$. A value of $\chi^2_\nu = 0.25$ would indicate the data points on average miss the fit by $0.5\sigma_i$ In other words, the square root of the reduced chi-square gives the rms deviation in units of σ_i.

Larger values of χ^2_ν indicate poorer agreement between the data and the fit. Thus, one looks at the reduced chi-square table for the probability of getting a value as large or larger than the actual χ^2_ν value from the fit. If all such values are too improbable to be accepted as a chance occurrence, one must conclude that the data miss the fit by more than can be expected. Either the σ_i are underestimated or the fitting formula does not model the data. The latter conclusion might be appropriate if the data appear to deviate systematically from the best fit, while the former might be more appropriate if the scatter in the data points simply appears to be larger than expected.

In rare cases, the χ^2_ν value may come out too small—well under the expected value of one and indicating that the data agree with the fit too well. In this case, use the table to evaluate the probability of getting any reduced chi-square value this small or smaller. If such small values are too unlikely to be accepted as a chance occurrence, the only logical conclusion is that the σ_i are incorrect and overestimated.

Considerations if σ_i is unknown

If the values of the σ_i are only roughly known, the χ^2_ν value from a fit would not be very useful for a chi-square test. It is not unusual for experimental estimates of σ_i to be uncertain at the factor-of-two of level in which case the χ^2_ν is uncertain at the factor-of-four level and the chi-square test is unusable. One might then take the χ^2_ν as an indication of how well the σ_i were estimated. A $\chi^2_\nu = 0.25$ would then be taken to mean that the σ_i were overestimated by a factor of two; a $\chi^2_\nu = 4$ would mean the σ_i were underestimated by a factor of 2. However, if even the largest σ_i's one might reasonably assign still give an unreasonably large χ^2_ν, one would then conclude the fit and data are not in agreement.

If the σ_i are so roughly known that the chi-square test is unusable, the fitting parameter uncertainties would also be unreliable. In this case, the σ_i can be determined or scaled so as to force the χ^2 value to take on its expectation value, $N - M$. If the σ_i so adjusted can be considered reasonable, the method described previosuly can then be used to determine the fitting parameter uncertainties. They would then become sample standard deviations s_k and Student-T confidence intervals would be applicable.

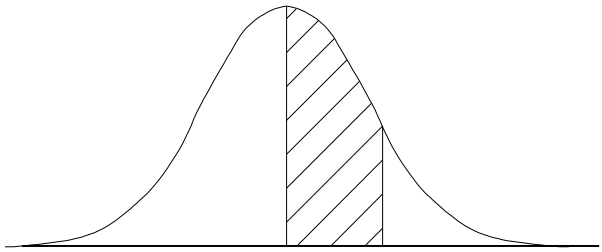
Exercise 14 *(a) What is the χ^2_ν for the fit of Exercise 13 assuming $\sigma_y = 0.5$? What is the probability it would have come out this big or*

[4] If $F_\alpha(x)$ is known, the quantity

$$\chi^2_\nu = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i - F_\alpha(x_i))^2}{\sigma_i^2}$$

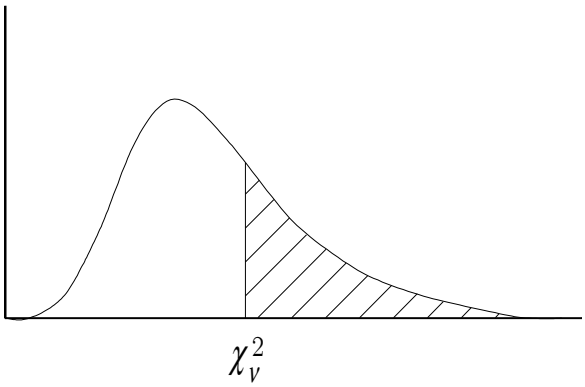
is a reduced chi-square random variable with N degrees of freedom.

bigger? (b) Suppose σ_y was not known. How small would it have to be before the deviations from the fit would have to be deemed (at the 99% level) too big to be in agreement with the quadratic fitting formula? (c) How big would σ_y have to be before one would have to conclude it is too big to be reasonable (at the 99% level).



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.10	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.20	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.30	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.40	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.50	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.60	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.70	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.80	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.90	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.00	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.10	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.20	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.30	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.40	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.50	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.60	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.70	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.80	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.90	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.00	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.10	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.20	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.30	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.40	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.50	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.60	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.70	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.80	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.90	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.00	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900

Table 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.



P / ν	0.99	0.98	0.95	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.000	0.001	0.004	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.010	0.020	0.051	0.105	0.223	0.357	0.511	0.693	0.916	1.204	1.609	2.303	2.996	3.912	4.605	6.908
3	0.038	0.062	0.117	0.195	0.335	0.475	0.623	0.789	0.982	1.222	1.547	2.084	2.605	3.279	3.782	5.422
4	0.074	0.107	0.178	0.266	0.412	0.549	0.688	0.839	1.011	1.220	1.497	1.945	2.372	2.917	3.319	4.617
5	0.111	0.150	0.229	0.322	0.469	0.600	0.731	0.870	1.026	1.213	1.458	1.847	2.214	2.678	3.017	4.103
6	0.145	0.189	0.273	0.367	0.512	0.638	0.762	0.891	1.035	1.205	1.426	1.774	2.099	2.506	2.802	3.743
7	0.177	0.223	0.310	0.405	0.546	0.667	0.785	0.907	1.040	1.198	1.400	1.717	2.010	2.375	2.639	3.474
8	0.206	0.254	0.342	0.436	0.574	0.691	0.803	0.918	1.044	1.191	1.379	1.670	1.938	2.271	2.511	3.265
9	0.232	0.281	0.369	0.463	0.598	0.710	0.817	0.927	1.046	1.184	1.360	1.632	1.880	2.187	2.407	3.097
10	0.256	0.306	0.394	0.487	0.618	0.727	0.830	0.934	1.047	1.178	1.344	1.599	1.831	2.116	2.321	2.959
11	0.278	0.328	0.416	0.507	0.635	0.741	0.840	0.940	1.048	1.173	1.330	1.570	1.789	2.056	2.248	2.842
12	0.298	0.348	0.436	0.525	0.651	0.753	0.848	0.945	1.049	1.168	1.318	1.546	1.752	2.004	2.185	2.742
13	0.316	0.367	0.453	0.542	0.664	0.764	0.856	0.949	1.049	1.163	1.307	1.524	1.720	1.959	2.130	2.656
14	0.333	0.383	0.469	0.556	0.676	0.773	0.863	0.953	1.049	1.159	1.296	1.505	1.692	1.919	2.082	2.580
15	0.349	0.399	0.484	0.570	0.687	0.781	0.869	0.956	1.049	1.155	1.287	1.487	1.666	1.884	2.039	2.513
16	0.363	0.413	0.498	0.582	0.697	0.789	0.874	0.959	1.049	1.151	1.279	1.471	1.644	1.852	2.000	2.453
17	0.377	0.427	0.510	0.593	0.706	0.796	0.879	0.961	1.048	1.148	1.271	1.457	1.623	1.823	1.965	2.399
18	0.390	0.439	0.522	0.604	0.714	0.802	0.883	0.963	1.048	1.145	1.264	1.444	1.604	1.797	1.934	2.351
19	0.402	0.451	0.532	0.613	0.722	0.808	0.887	0.965	1.048	1.142	1.258	1.432	1.587	1.773	1.905	2.306
20	0.413	0.462	0.543	0.622	0.729	0.813	0.890	0.967	1.048	1.139	1.252	1.421	1.571	1.751	1.878	2.266
22	0.434	0.482	0.561	0.638	0.742	0.823	0.897	0.970	1.047	1.134	1.241	1.401	1.542	1.712	1.831	2.194
24	0.452	0.500	0.577	0.652	0.753	0.831	0.902	0.972	1.046	1.129	1.231	1.383	1.517	1.678	1.791	2.132
26	0.469	0.516	0.592	0.665	0.762	0.838	0.907	0.974	1.045	1.125	1.223	1.368	1.496	1.648	1.755	2.079
28	0.484	0.530	0.605	0.676	0.771	0.845	0.911	0.976	1.045	1.121	1.215	1.354	1.476	1.622	1.724	2.032
30	0.498	0.544	0.616	0.687	0.779	0.850	0.915	0.978	1.044	1.118	1.208	1.342	1.459	1.599	1.696	1.990
32	0.511	0.556	0.627	0.696	0.786	0.855	0.918	0.979	1.043	1.115	1.202	1.331	1.444	1.578	1.671	1.953
34	0.523	0.567	0.637	0.704	0.792	0.860	0.921	0.980	1.042	1.112	1.196	1.321	1.429	1.559	1.649	1.919
36	0.534	0.577	0.646	0.712	0.798	0.864	0.924	0.982	1.042	1.109	1.191	1.311	1.417	1.541	1.628	1.888
38	0.545	0.587	0.655	0.720	0.804	0.868	0.926	0.983	1.041	1.106	1.186	1.303	1.405	1.525	1.610	1.861
40	0.554	0.596	0.663	0.726	0.809	0.872	0.928	0.983	1.041	1.104	1.182	1.295	1.394	1.511	1.592	1.835
42	0.563	0.604	0.670	0.733	0.813	0.875	0.930	0.984	1.040	1.102	1.178	1.288	1.384	1.497	1.576	1.812
44	0.572	0.612	0.677	0.738	0.818	0.878	0.932	0.985	1.039	1.100	1.174	1.281	1.375	1.485	1.562	1.790
46	0.580	0.620	0.683	0.744	0.822	0.881	0.934	0.986	1.039	1.098	1.170	1.275	1.366	1.473	1.548	1.770
48	0.587	0.627	0.690	0.749	0.825	0.884	0.936	0.986	1.038	1.096	1.167	1.269	1.358	1.462	1.535	1.751
50	0.594	0.633	0.695	0.754	0.829	0.886	0.937	0.987	1.038	1.094	1.163	1.263	1.350	1.452	1.523	1.733
60	0.625	0.662	0.720	0.774	0.844	0.897	0.944	0.989	1.036	1.087	1.150	1.240	1.318	1.410	1.473	1.660
70	0.649	0.684	0.739	0.790	0.856	0.905	0.949	0.990	1.034	1.081	1.139	1.222	1.293	1.377	1.435	1.605
80	0.669	0.703	0.755	0.803	0.865	0.911	0.952	0.992	1.032	1.076	1.130	1.207	1.273	1.351	1.404	1.560
90	0.686	0.718	0.768	0.814	0.873	0.917	0.955	0.993	1.031	1.072	1.123	1.195	1.257	1.329	1.379	1.525
100	0.701	0.731	0.779	0.824	0.879	0.921	0.958	0.993	1.029	1.069	1.117	1.185	1.243	1.311	1.358	1.494
120	0.724	0.753	0.798	0.839	0.890	0.928	0.962	0.994	1.027	1.063	1.107	1.169	1.221	1.283	1.325	1.447
140	0.743	0.770	0.812	0.850	0.898	0.934	0.965	0.995	1.026	1.059	1.099	1.156	1.204	1.261	1.299	1.410
160	0.758	0.784	0.823	0.860	0.905	0.938	0.968	0.996	1.024	1.055	1.093	1.146	1.191	1.243	1.278	1.381
180	0.771	0.796	0.833	0.868	0.910	0.942	0.970	0.996	1.023	1.052	1.087	1.137	1.179	1.228	1.261	1.358
200	0.782	0.806	0.841	0.874	0.915	0.945	0.972	0.997	1.022	1.050	1.083	1.130	1.170	1.216	1.247	1.338

Table 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.

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

Table 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.