Treatment of Experimental Data in the Physical Chemistry Laboratory

George C. McBane
Department of Chemistry
The Ohio State University

January 2, 2001

Preface and Acknowledgements

These notes are used in the Physical Chemistry Laboratory course at Ohio State University. They are intended to provide a concise description of the statistical and numerical techniques required to treat data obtained in physical science experiments at the advanced undergraduate level.

The notes have evolved from a handwritten set prepared by Isaiah Shavitt. Students in the course now use the computer program Mathcad[1] for a large fraction of the necessary analysis and data treatment; many of the Mathcad examples were provided by Jim Coe, who also has had substantial influence on other parts of the notes. Russell Pitzer made useful suggestions. Regina Ragins first converted the notes to electronic form.

Much of the recent material comes from one or more of four books: *Numerical Recipes*, by Press et al. [1]; *Data Reduction and Error Analysis for the Physical Sciences*, by Bevington and Robinson [2], called B&R in these notes; *Experiments in Physical Chemistry*, by Shoemaker et al. [3], called SGN here; and *Statistical Treatment of Experimental Data*, by Young [4].

George McBane
January 2, 2001

[1] Mathcad is a registered trademark of Mathsoft, Inc.

Contents

Chemistry 541 should teach you how to obtain good experimental data, how to analyze it quantitatively, and how to present it in written form.

When your future boss asks you to determine the boiling point of some new product, she doesn't really care what result you get; she wants to know what result the *customer* will get when he measures it. Since your crystal ball is cloudy, you must do the best you can to predict. Much of 541 is dedicated to exactly how you do that.

1 Types of experimental error

Several kinds of errors are usually present in experimental data. Their effects on the desired results can range from insignificant to disastrous, depending on how well they are understood and accounted for.

Some general characteristics of errors are described by two words with very specific meanings in quantitative work: *precision* and *accuracy*. Precision describes the tendency of several measurements in a set to have values close to one another; accuracy describes whether the measurements are close to a "true" or accepted value. A basketball player whose shots always pass exactly two feet to the right of the hoop shows excellent precision but suffers from poor accuracy.

1.1 Blunders, mistakes, screwups

These mistakes correspond to the common-English usage of the term "error". Some examples are

- using the wrong material or concentration,
- transposing digits in recording scale readings,
- arithmetic errors.

There are no fancy techniques I can teach which will save you from these sorts of errors; you just have to be careful and keep your wits about you. In general, you should read a scale, write down the result in your notebook, then read the scale again, to prevent mistakes. All recorded numbers should go directly into your notebook; that helps find and fix some kinds of errors such as arithmetic ones.

1.2 Systematic error

Systematic errors are consistent effects which change the system under study or the measurements you make on it. They have *signs*.

- Uncalibrated instruments (balances, etc.)
- Impure reagents
- Leaks
- Temperature effects not accounted for
- Biases in using equipment (even numbers in reading scales, seeing hoped-for small effects, etc.)
- Pressure differences between barometer and experiment caused by air conditioning

Systematic error affects the accuracy of an experiment but not the precision. Repeated trials and statistical analysis are of no use in eliminating its effects.

Careful experimental design and execution is the sole approach to reducing systematic error. Sometimes systematic errors can be corrected for in a simple way; for example, the thermal expansion of a metal scale can easily be accounted for if the temperature is known. Other errors, such as those caused by impure reagents, are harder to deal with. The most dangerous systematic errors are those that are unrecognized, and therefore can affect the results in completely unknown ways. The whole field of quantitative experimentation is dependent on workers' ability to recognize and eliminate systematic errors.

1.3 Random error

Random error arises from mechanical vibrations in the apparatus, electrical noise, uncertainty in reading of scale pointers, and other "fluctuations". It can be characterized, and sometimes reduced, by repeated (at least three) trials of an experiment. Its treatment is the subject of much of these notes.

Note that random error affects the precision of an experiment, and to a lesser extent its accuracy. Systematic error affects the accuracy only. Precision is easy to assess; accuracy is difficult.

2 Error distributions and distribution functions

If you measure some quantity (e.g., the barometric pressure) several times (say, twenty), you will probably get several different answers. A histogram of one student's results is shown in Figure 1:

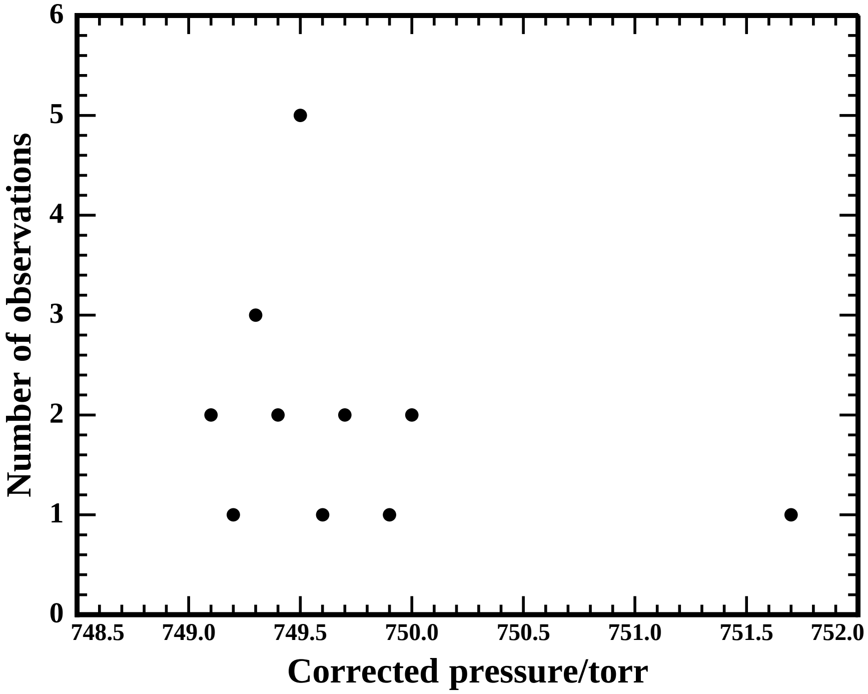


Figure 1: One set of barometric pressure measurements.

You can calculate several parameters from your observed collection of values. Some of the important ones are listed in Table 1, where N is the number of measurements and the x_i are the individual values obtained.

Those are parameters that apply to your particular sample. If you repeat the experiment, their new values will probably not be the same, even if the conditions (temperature in room, atmospheric conditions, etc.) were the same.

What relation do those parameters have to the actual barometric pres-

Table 1 Statistical characteristics of a simple data set.

mean

$$\overline{x} = \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

median The median x_{med} is the "middle" value in a dataset with an odd number of observations, and the average of the two middle values in a dataset with an even number of observations. So if the set of N data points x_i is sorted from lowest to highest, then

$$x_{\text{med}} = \begin{cases} x_{(N+1)/2} & (N \text{ odd,}) \\ \frac{1}{2}(x_{N/2} + x_{(N/2)+1}) & (N \text{ even.}) \end{cases} \tag{2}$$

variance

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{3}$$

standard deviation

$$S = \sqrt{S^2} = \frac{1}{\sqrt{N-1}} \left[\sum_{i=1}^{N} (x_i - \overline{x})^2 \right]^{\frac{1}{2}} \tag{4}$$

average absolute deviation from the mean

$$\text{ADev} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|. \tag{5}$$

average absolute deviation from the median

$$\text{ADev}_{\text{med}} = \frac{1}{N} \sum_{i=1}^{N} |x_i - x_{\text{med}}| \tag{6}$$

sure? (We assume that there is an "actual barometric pressure", though we have no way to know what it is.) The mean and median each gives an estimate of the true pressure; other parameters are indicators of the uncertainty in the true pressure caused by random errors. To understand the proper use of these estimates, we need some statistical tools.

If each measurement is just like any other so far as you know, then some unknown mechanism is making changes in your values before you get them. We assume that the probability of obtaining a certain value x_i in any one trial is given by a probability distribution, $P(x_i)$. This distribution is not known to us, but might be expected to look something like Figure 2. The data from Figure 1 are shown as open circles, scaled to fit on the plot.

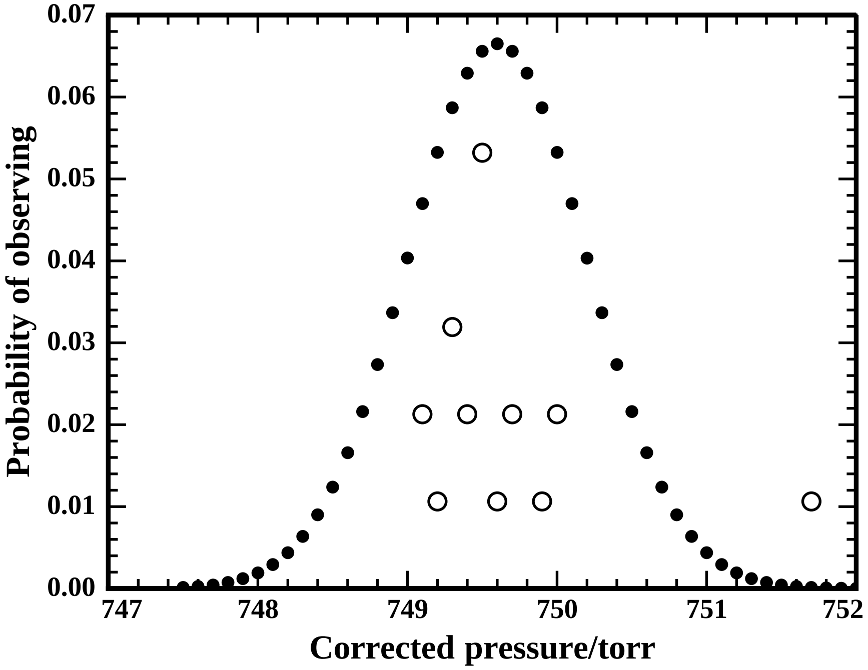


Figure 2: Possible parent distribution of pressure readings.

This is a discrete distribution, since each measurement is made only to the first decimal place. Such a distribution gives the (finite) probability of obtaining each possible result on any one trial of the experiment.

If we do the experiment, we must get *some* answer, so the distribution must be *normalized*:

$$\sum_{i=1}^{N} P(x_i) = 1. \tag{7}$$

While the pressure measurements can give only discrete answers (because of limitations of our eyes), the actual pressure has no such limitation. Presumably the distribution $P(x_i)$ is a "condensed" version of a *continuous* probability distribution (or, in strict parlance, probability density function) $P(x)$. The continuous distribution $P(x)$ has the definition that $P(x)\,dx$ gives the probability that a measurement will give a result in the range $[x,\ x + dx]$. So, the probability that a measured value will be between x_1 and x_2 is

$$P(x \in [x_1,\, x_2]) = \int_{x_1}^{x_2} P(x)\,dx. \tag{8}$$

We do not know $P(x)$. However, we do know that it must be normalized (unit probability of getting some answer):

$$\int_{-\infty}^{\infty} P(x)\,dx = 1. \tag{9}$$

Table 2 summarizes important properties of discrete and continuous probability distributions.

The distribution $P(x)$, or, in the case of measurements which are truly discrete, $P(x_i)$, which controls the probability of getting a particular answer on any one experimental trial, is called the *parent distribution*. The distribution actually obtained by the experimenter is called the *sample distribution*. Greek letters are usually used to represent parameters of the parent distribution (mean μ, standard deviation σ, etc.) and Roman letters used for the sample distribution ($\overline{x}$, S, etc.)

The set of values actually obtained in an experiment is simply one of very many possible sets. That idea underlies all statistical analysis of data. The "likelihood" of each possible set is controlled by the parent distribution. The sample distribution will become more and more similar to the parent distribution as the number of samples become larger, approaching equality as $N \to \infty$. If we knew the parent distribution $P(x)$, and there were no systematic errors, we would know the true answer. Instead we must try to guess the parent distribution from the available samples. Before I describe how to do that, I want to give some examples of calculations with probability distributions and discuss several important distributions.

Table 2 Summary of important properties of probability distributions.

- Normalization:

$$1 = \begin{cases} \int_{-\infty}^{\infty} P(x)\, dx & \text{(continuous)} \\ \sum_{i=1}^{N} P(x_i) & \text{(discrete)} \end{cases} \tag{10}$$

- Probability that a single result will lie in a specified interval:

$$P(x \in [x_1,\, x_2]) = \int_{x_1}^{x_2} P(x)\, dx \text{ (continuous).} \tag{11}$$

- Average value of x:

$$\langle x \rangle = \bar{x} = \begin{cases} \int_{-\infty}^{\infty} xP(x)\, dx & \text{(continuous)} \\ \sum_{i=1}^{N} x_i P(x_i) & \text{(discrete)} \end{cases} \tag{12}$$

- Average values (expectation values) of functions of x:

$$\langle f(x) \rangle = \overline{f(x)} = \begin{cases} \int_{-\infty}^{\infty} f(x)P(x)\, dx & \text{(continuous)} \\ \sum_{i=1}^{N} f(x_i)P(x_i) & \text{(discrete)} \end{cases} \tag{13}$$

- Standard deviation:

$$\sigma = \left(\int_{-\infty}^{\infty} (x - \bar{x})^2 P(x)\, dx \right)^{\frac{1}{2}} \text{ (continuous)} \tag{14}$$

$$S = \left(\frac{N}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 P(x_i) \right)^{\frac{1}{2}} \text{ (discrete)} \tag{15}$$

3 Examples of probability distribution calculations

3.1 A discrete distribution: possible results in throws of 2 dice

All the possible outcomes of a toss of two dice are enumerated in Table 3. Note that the distribution is normalized; the sum of the fractions in the right column is 1. The average throw, from equation 12, is $2 \times \frac{1}{36} + 3 \times \frac{1}{18} + 4 \times \frac{1}{12} + 5 \times \frac{1}{9} + 6 \times \frac{5}{36} + 7 \times \frac{1}{6} + 8 \times \frac{5}{36} + 9 \times \frac{1}{9} + 10 \times \frac{1}{12} + 11 \times \frac{1}{18} + 12 \times \frac{1}{36} = 7$. The average *squared* throw is $4 \times \frac{1}{36} + 9 \times \frac{1}{18} + 16 \times \frac{1}{12} + 25 \times \frac{1}{9} + 36 \times \frac{5}{36} + 49 \times \frac{1}{6} + 64 \times \frac{5}{36} + 81 \times \frac{1}{9} + 100 \times \frac{1}{12} + 121 \times \frac{1}{18} + 144 \times \frac{1}{36} = 54.8\overline{3}$. Note that the average squared throw is not 49.

Table 3 Possible outcomes in throws of 2 dice.

Result	Combinations	# Comb.	Prob.
2	1+1	1	1/36
3	1+2, 2+1	2	1/18
4	1+3, 2+2, 3+1	3	1/12
5	1+4, 2+3, 3+2, 4+1	4	1/9
6	1+5, 2+4, 3+3, 4+2, 5+1	5	5/36
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	6	1/6
8	2+6, 3+5, 4+4, 5+3, 6+2	5	5/36
9	3+6, 4+5, 5+4, 6+3	4	1/9
10	4+6, 5+5, 6+4	3	1/12
11	5+6, 6+5	2	1/18
12	6+6	1	1/36

3.2 A continuous distribution: lifetimes of nuclei

The lifetimes of ^{241}Am nuclei are governed by the distribution

$$P(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}, \tag{16}$$

where τ has the value 407 yr. We can check that the distribution is normalized:

$$\frac{1}{\tau} \int_0^\infty e^{-\frac{t}{\tau}}\, dt = -\int_0^{-\infty} e^z\, dz = 1, \tag{17}$$

where I made the change of variable $z = -t/\tau$, $dz = -\frac{1}{\tau}\, dt$, $dt = -\tau\, dz$. Note that since it makes no sense for a nucleus to have a negative lifetime, the distribution must have value zero for all $t < 0$. The normalization integral therefore goes only from 0 to ∞.

The mean lifetime of a nucleus is

$$\begin{aligned} \langle t \rangle &= \frac{1}{\tau} \int_0^\infty t e^{-\frac{t}{\tau}} \, dt \\ &= \frac{1}{\tau} \left[\frac{e^{-t/\tau}}{(-1/\tau)^2} \left(-\frac{t}{\tau} - 1\right) \right]_0^\infty \\ &= \tau. \end{aligned} \tag{18}$$

The probability that a nucleus will live longer than 3τ is

$$P(t \geq 3\tau) = \int_{3\tau}^\infty P(t) \, dt = -\int_{-3}^{-\infty} e^z \, dz = e^{-3} = 0.0498 \tag{19}$$

after the same change of variable as before.

4 Some important distributions

4.1 Binomial distribution

The binomial distribution is a discrete distribution which may be used to answer questions such as "if a regular six-sided die is tossed twenty times, what is the probability that it will land with the 4-spotted side up exactly seven times?". In general it describes repeated events which have two possible outcomes (either a toss gives 4 or it does not), which are often termed *success* and *failure*. In n events, if the probability of success on each event is p, the probability of having exactly x successes is

$$P_{\mathrm{B}}(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}, \tag{20}$$

where the exclamation point indicates factorial. The notation $\binom{n}{x}$ for the *binomial coefficient* is standard.

The answer to my question about tossing seven 4s out of twenty throws is therefore

$$P = \frac{20!}{(7!)(13!)} \left(\frac{1}{6}\right)^7 \left(1 - \frac{1}{6}\right)^{13} = 0.026. \tag{21}$$

4.2 Poisson distribution

The Poisson distribution is a limiting case of the binomial distribution for small p and large n. It arises most often in counting experiments. If the average number of events (cosmic rays detected by an apparatus, the number of cases of rabies observed in a city's racoon population, etc.) expected in a given interval is μ, then the probability of observing exactly x events is

$$P_{\mathrm{P}}(x;\mu) = \frac{\mu^x}{x!}e^{-\mu} \tag{22}$$

When $\mu \ll 1$, the Poisson distribution looks like a decaying exponential; when $\mu \gg 1$, it is peaked and looks like the normal distribution to be described next. More information on the binomial and Poisson distributions is given in Bevington and Robinson [2].

4.3 The normal distribution

A particularly important continuous distribution is the normal, or Gaussian, distribution, given by

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \tag{23}$$

where x is the independent variable and μ and σ are parameters describing the distribution. The normal distribution is shown in Figure 3.

The normal distribution is important in physical science largely because of the *central limit theorem*, which states that under certain conditions if a large number of small fluctuations are added together, their sum will be approximately described by the normal distribution no matter what their individual parent distributions are. Many experiments, whose sources of small random errors are many, do appear to have parent distributions well approximated by the normal distribution.

4.3.1 Basic properties of the normal distribution

If we insert $x = \mu \pm \sigma$ into the normal distribution, we find that the height of the curve at positions $\pm\sigma$ away from the center is $e^{-1/2}/(\sigma\sqrt{2\pi})$. The height at the maximum ($x = \mu$)is $1/(\sigma\sqrt{2\pi})$, so

$$\frac{P(\mu \pm \sigma)}{P(\mu)} = e^{-\frac{1}{2}} = 0.6065. \tag{24}$$

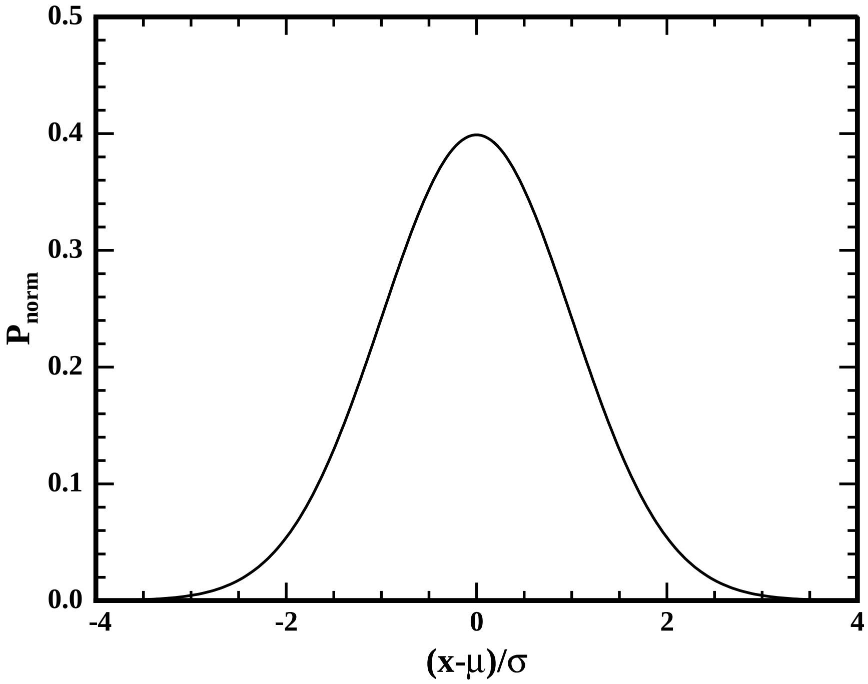


Figure 3: The normal distribution.

A more useful quantity is the fraction of measurements expected to be within one σ of μ, which we get from Eq. (8) as

$$\begin{aligned} P(x \in [\mu - \sigma,\, \mu + \sigma]) &= \int_{\mu-\sigma}^{\mu+\sigma} P(x)\, dx \qquad (25) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-\sigma}^{\mu+\sigma} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right] dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{1}{2}u^2}\, du, \end{aligned}$$

where the change of variable $u = (x - \mu)/\sigma$, $du = dx/\sigma$ has been made. This integral cannot be done analytically for limits other than $[0, \pm\infty]$ and $[-\infty, \infty]$. Its value can be looked up in tables or calculated numerically (more on this in a moment), and is 0.68269. Since the total area is 1, about

68% of the area is contained between $\mu - \sigma$ and $\mu + \sigma$. So 68% of the samples taken in a normally distributed experiment should lie in that range about μ.

We now have a first, simple example of a *confidence interval*: if samples are taken from a probability distribution which is normal with a mean μ and standard deviation σ, we expect that measured values will fall within σ of μ about 68% of the time. What if we want to give an interval that will contain more of the measurements? We need a way to evaluate integrals like Eq. (25) for more general values of the limits, to give larger values of the integrated probability. That is, we want to do the integral for limits $\mu \pm z\sigma$, and we will adjust z to give integral values of 0.9 if we want 90% probability, 0.95 if we want 95%, and so on.

4.3.2 Definite integrals of the normal distribution

There are two popular ways to obtain values for definite integrals of the normal distribution:

Use a table Tables of integrals of the normal distribution for different limits and for $\mu = 0, \sigma = 1$ are given in many places, including Bevington and Robinson, p. 253, Young, p. 161, and the Chemical Rubber Company's *Standard Mathematical Tables* [5]. Sometimes the *cumulative normal distribution* $F(t)$ is tabulated, corresponding to

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}u^2}\, du, \tag{26}$$

sometimes the symmetric integral values

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{1}{2}u^2}\, du, \tag{27}$$

and sometimes the integral from 0 to t,

$$\frac{1}{\sqrt{2\pi}} \int_{0}^{t} e^{-\frac{1}{2}u^2}\, du. \tag{28}$$

To use these tables with your own values of μ, σ, and z, you change variables: $u = (x - \mu)/\sigma)$. Don't forget to adjust the limits of integration too: if x ran from $\mu - z\sigma$ to $\mu + z\sigma$, then u will run from $-z$ to z. Then, depending on which integration limits your table gives, you might need to do some simple tricks with the symmetry to get the limits you wanted. For example, if your table gives the integral from 0 to t and you wanted the integral from $-t$ to t, you must multiply the value given in the table by 2. If your table gives the cumulative normal distribution $F(t)$ described in Eq. (26) and you want the integral from $-t$ to t, you must subtract off the part from $-\infty$ to $-t$; since the integral from $-\infty$ to ∞ is 1, you can take $F(t) - (1 - F(t)) = 2F(t) - 1$ to get your result.

Use a computer program Many computer programs can calculate values of the normal distribution and its integrals. For example, Mathcad has a function pnorm(x, μ, σ) which returns the cumulative normal distribution

$$\text{pnorm}(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt. \tag{29}$$

Worksheet 1 shows a use of Mathcad's `pnorm`(x, μ, σ) function to evaluate an integral of the normal distribution.

4.3.3 Simple confidence intervals

Now you can do confidence intervals of arbitrary probability for quantities described by the normal distribution. If you want to find the interval which will include 99% of the measurements from a normally distributed sampling process with mean μ and standard deviation σ, you keep evaluating integrals of the normal distribution with ever-widening limits until the value of the integral equals 0.99. Then the limits you used become the upper and lower confidence limits.

The Mathcad inverse cumulative probability distribution functions, such as `qnorm`(p, μ, σ), can help you do this calculation. Worksheet 2 shows an example.

5 Extracting information from limited data sets

Finally, we're back to the point we left on page 6, namely trying to estimate the characteristics of the parent distribution from the observed data. Our sample distribution is simply one of many possible ones, though we happened not to find the others. But it's all we have.

At this point we have several options. Sometimes the best is to use computer-based techniques known collectively as Monte Carlo methods, which I discuss in Section 7.2. Classical statistical analysis is based on a

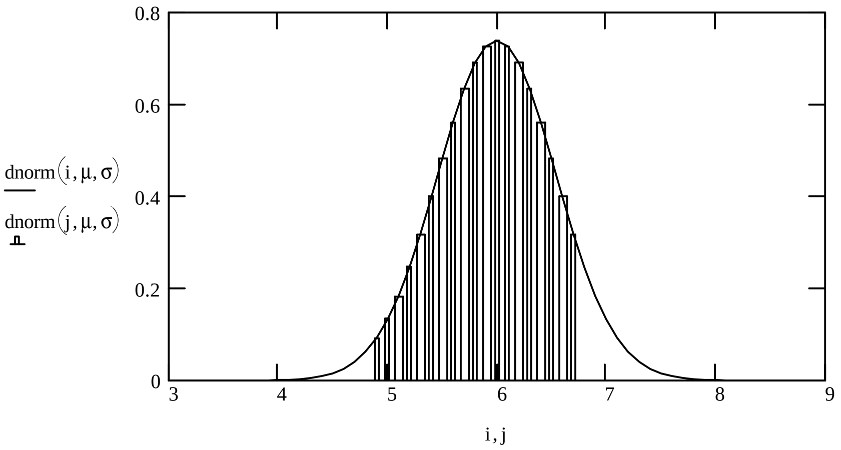
Worksheet 1 Integrals of the normal distribution.

Evaluate integral of normal distribution with μ=6, σ=0.54 from 4.9 to 6.7

$\mu := 6 \qquad \sigma := 0.54 \qquad \text{lowlim} := 4.9 \qquad \text{highlim} := 6.7$

$\text{int} := \text{pnorm}(\text{highlim}, \mu, \sigma) - \text{pnorm}(\text{lowlim}, \mu, \sigma) \qquad \text{int} = 0.882$

$i := 3, 3.1 .. 9 \qquad j := \text{lowlim}, \text{lowlim} + .1 .. \text{highlim}$

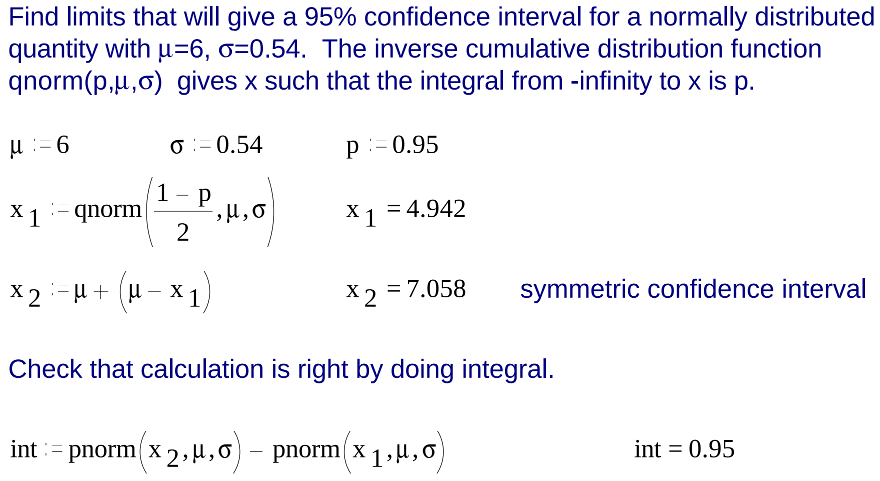


different procedure: we *assume* a particular form of parent distribution, and then use the "maximum likelihood principle" to find the parameters of the parent distribution which give the largest probability that the observed data set would occur.

Sometimes the form of parent distribution one should assume is well known from the type of experiment. For example, in experiments which involve counting the number of events (photons arriving at a detector, decays observed from a radioactive sample, etc.) occurring in different time intervals, the Poisson distribution is usually appropriate. More often, the sources of random error in the experiment are not well known, an implicit appeal to the central limit theorem is made, and the normal distribution is assumed to apply.

If we assume that the parent distribution is normal, then it is easy to show (Bevington and Robinson pp. 53–55) that the best estimate of the mean μ of the parent distribution is just $\overline{x}$, the mean of the sample distri-

Worksheet 2 Confidence limits from the normal distribution.



Find limits that will give a 95% confidence interval for a normally distributed quantity with μ=6, σ=0.54. The inverse cumulative distribution function qnorm(p,μ,σ) gives x such that the integral from -infinity to x is p.

$\mu := 6 \qquad \sigma := 0.54 \qquad p := 0.95$

$x_1 := \mathrm{qnorm}\left(\frac{1-p}{2}, \mu, \sigma\right) \qquad x_1 = 4.942$

$x_2 := \mu + (\mu - x_1) \qquad x_2 = 7.058$ symmetric confidence interval

Check that calculation is right by doing integral.

$\mathrm{int} := \mathrm{pnorm}(x_2, \mu, \sigma) - \mathrm{pnorm}(x_1, \mu, \sigma) \qquad \mathrm{int} = 0.95$

bution. Now we want to ask the question: How certain are we of the true value?

If the sample distribution looks like Figure 4, then it seems like we know the value of the mean to considerably better precision than ±S at 68% confidence. The sample standard deviation S gives the spread of *individual measurements*, so that approximately 68% of them will be within S of the mean. That is true no matter how many measurements we take. However, the value of the mean becomes better and better determined as more measurements are made.

5.1 Uncertainty in the mean value

The standard deviation S (called the *sample standard deviation* in some texts) gives an estimate of σ for the assumed parent distribution, and therefore describes the uncertainty in any individual measurement. The mean $\overline{x}$, which we want to use as an estimate of the parent mean μ, is calculated from the N individual x_i. We will show in the next section how to estimate the uncertainty in a quantity calculated from several directly measured quantities. The results for this case, the estimated variance and standard

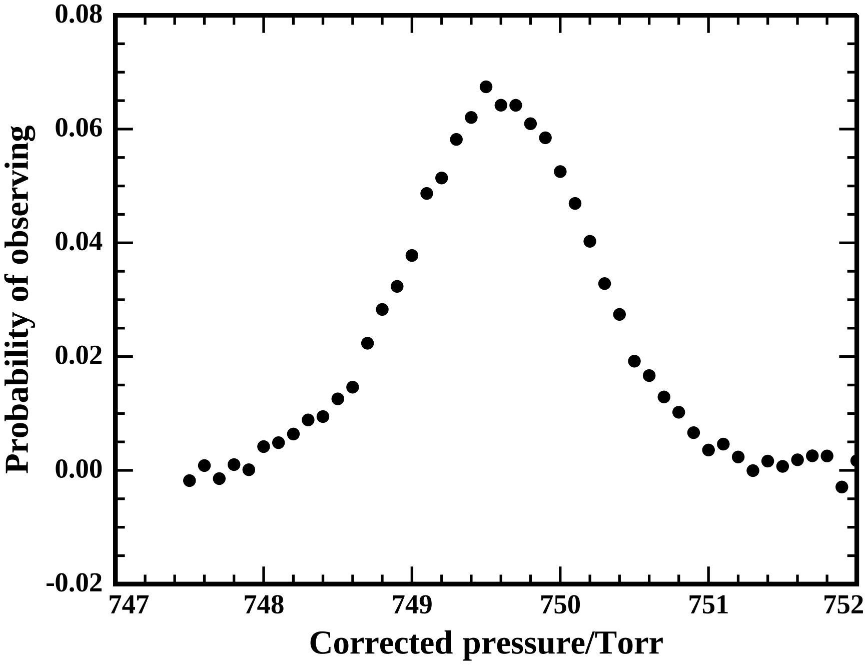


Figure 4: Possible sample distribution.

deviation *of the mean*, are

$$S_{\mathrm{m}}^2 = \frac{S^2}{N}, \text{ or} \tag{30}$$

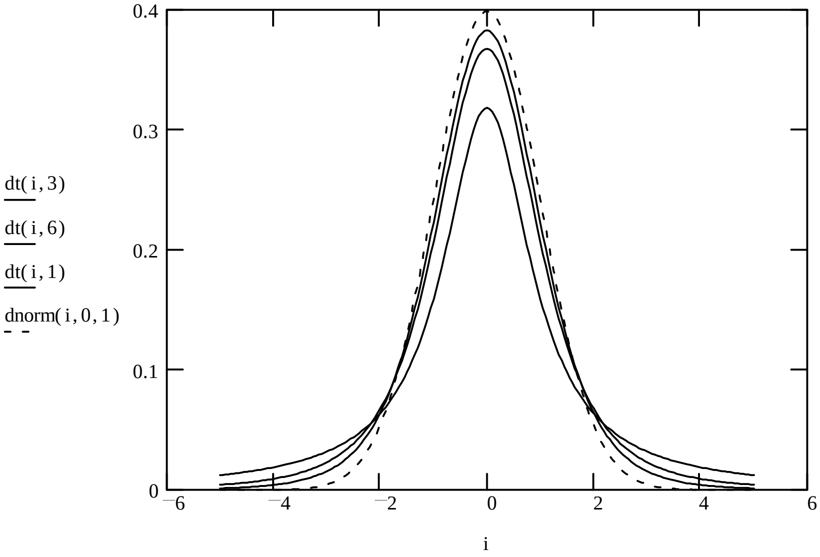
$$S_{\mathrm{m}} = \frac{S}{\sqrt{N}}. \tag{31}$$

Think of a probability distribution which describes the likely deviation of the measured $\overline{x}$ from the true mean μ. Unless we have a very large number of measurements N, that distribution is not Gaussian but is given by a different distribution called the Student-t distribution. If N is small, this distribution is rather wide; our sample might in fact have a mean $\overline{x}$ rather far from the true mean μ. As N becomes large, the Student-t distribution looks more and more like the normal distribution. Worksheet 3 shows plots of the Student-t distribution for several different N.

Worksheet 3 The Student-*t* distribution.

Show Student-t distribution for different degrees of freedom
Normal distribution is shown as a dashed line for comparison

$i := -5, -4.95 .. 5$



The Student-*t* distribution is given by

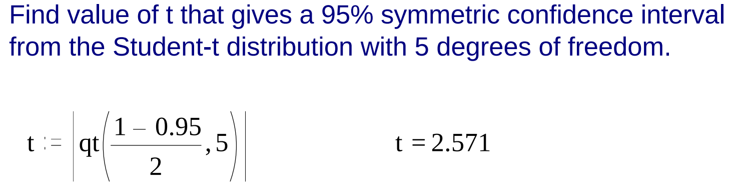
$$P(\tau) = k_{\text{norm}} \left(1 + \frac{\tau^2}{N-1}\right)^{-\frac{N}{2}}, \tag{32}$$

where $\tau = \frac{\bar{x}-\mu}{S_{\text{m}}}$, and k_{norm} is a normalization constant. (The expression for k_{norm} is given in equation (31) on p. 47 of SGN.) Confidence limits are found with this distribution the same way they are found for the normal distribution. With a table of integrals of the distribution, integration limits are chosen which make the integral equal to the desired fraction (0.9 for 90%, and so on.) Table 4 gives values of *t* which give various confidence intervals:

$$0.95 = \int_{\bar{x}-tS_{\text{m}}}^{\bar{x}+tS_{\text{m}}} P(\tau)\, d\tau \tag{33}$$

for a 95% confidence limit, and so on. The table was generated with a series of Mathcad calculations like that shown in Worksheet 4.

Worksheet 4 Confidence intervals from the Student-*t* distribution.



Find value of t that gives a 95% symmetric confidence interval
from the Student-t distribution with 5 degrees of freedom.

$$t := \left| qt\left(\frac{1-0.95}{2}, 5\right) \right| \qquad t = 2.571$$

Keep in mind what this 95% confidence limit actually means: if you were to repeat the experiment many times, and each time claim that the true mean μ was within the limits $\overline{x} \pm tS_{\mathrm{m}}$ you obtained on that particular repetition, you expect to be right 95% of the time.

5.2 Reporting measurements of a single quantity

5.2.1 General guidelines

When you report a value x, you must also report some estimate of its uncertainty u. No matter how you arrived at x and u (some suggestions are given in the following sections), you should follow some general rules.

1. The value and the uncertainty should be reported to the same number of decimal places: (75.63 ± 0.06) kg, not (75.6347 ± 0.06) kg or (75.63 ± 0.0629) kg.

   I suggest that you report the uncertainty only to one significant figure if that figure is 3 or larger, and to two if the uncertain digits are 25 or less. The result is given to the same number of decimal places as its uncertainty. So if the calculated result had been 752.2083 Torr ± 0.0143 Torr, you would report (752.208 ± 0.014) Torr. Other conventions are used in some fields.

2. The value and the uncertainty should have the same power of 10 if you are using scientific notation. One good format is $(4.73 \pm 0.08) \times 10^{-5}$ J. You confuse your reader with $(4.732 \times 10^{-5} \pm 8.5 \times 10^{-7})$ J.

3. The units should be clear. If you are using SI units (if you aren't, why aren't you?), use the accepted SI symbols; in particular, the symbol for second is s, not sec, and that for gram is g, not gm. Symbols do not take an 's' to become plural: 12 kg, not 12 kgs. They do not need periods: 1.2 g, not 1.2 g., unless the symbol ends a sentence. They are always set in upright type, not italicized.

> In the examples above, I have used parentheses to make it clear that the unit applies both to x and to u: (752.208 ± 0.014) Torr. It is also acceptable to place the unit explicitly on both value and uncertainty: 752.208 Torr ± 0.014 Torr. The form 752.208 ± 0.014 Torr is widely used but is discouraged by NIST, the U.S. agency with responsibility for physical measurement standards.

5.2.2 Simplest case: standard deviation only

If you measure some quantity several (say N) times, and the parent distribution of the measurements is Gaussian as far as you know, the best estimate of the true value of that quantity is probably the mean of the N measured values. The simplest measure of the uncertainty in the true value obtained in that way is the estimated standard deviation of the mean S_m. So at the very least, quantities obtained in this way should be reported as

$$X = \overline{x} \pm S_\mathrm{m} \text{ (1 e.s.d. error limit).} \tag{34}$$

For example, if you made 6 measurements of the barometric pressure and obtained the values 758.23, 757.98, 757.92, 758.09, 758.17, and 758.14 Torr, the calculated mean is 758.088 Torr and the estimated standard deviation of the mean is 0.0488 Torr. You would report the value (758.09 ± 0.05) Torr, and specify that the uncertainty was one estimated standard deviation of the mean.

5.2.3 Confidence limits from the Student-*t* distribution

It is much better to report confidence limits, which carry information about the number of measurements. In this case, you report

$$X = \overline{x} \pm tS_\mathrm{m} \text{ (95\% confidence, } \nu = 7) \tag{35}$$

if you have 8 measurements and choose a 95% confidence limit. The value of t is found in Table 4 below (or in Table 3 on p. 49 of SGN.) Choose your desired confidence interval from the row labeled P, and choose $\nu = N - 1$. For $N = 6$ at 95% confidence, $t = 2.57$. For the example above, the reported result would be

$$(758.09 \pm 0.12) \text{ Torr (95\% confidence, } \nu = 5).$$

Table 4 Critical values of t. These values make the integral of the Student-t distribution from $-t$ to t or from $-\infty$ to t equal to the fractions given at the top in the rows labeled P and P', respectively. The values were generated with Mathcad.

	P	0.80	0.90	0.95	0.99
ν	P'	0.90	0.95	0.975	0.995
2		1.89	2.92	4.30	9.92
3		1.64	2.35	3.18	5.84
4		1.53	2.13	2.78	4.60
5		1.48	2.02	2.57	4.03
6		1.44	1.94	2.45	3.71
7		1.41	1.89	2.36	3.50
8		1.40	1.86	2.31	3.36
9		1.38	1.83	2.26	3.25
10		1.37	1.81	2.23	3.17
11		1.36	1.80	2.20	3.11
15		1.34	1.75	2.13	2.95
20		1.33	1.72	2.09	2.85
30		1.31	1.70	2.04	2.75
∞		1.28	1.64	1.96	2.58

5.2.4 Suspected nonnormal parent distributions

If you suspect that the parent distribution of your experiment is not Gaussian, for instance because you have several points that seem unreasonably far from the others, it is safer to use the median as the estimator of the central value. In that case, the absolute average deviation from the median is the preferred estimator of the uncertainty. So for the barometer data listed above, you would calculate

$$x_{\mathrm{med}} = \frac{1}{2}(758.14 + 758.09) = 758.115, \tag{36}$$

and

$$\mathrm{ADev}_{\mathrm{med}} = \frac{1}{N}\sum_{i=1}^{N} |x_i - x_{\mathrm{med}}| = 0.092, \tag{37}$$

and would therefore report the value as (758.12 ± 0.09) Torr, stating that the reported values are the median of 6 measurements and average absolute deviation from the median. There is no simple way to estimate confidence limits in this case; you are using the median instead of the mean because you are admittedly ignorant of the parent distribution. You therefore have no way to calculate confidence intervals, and are simply giving your reader the best information you can.

6 Rejection of data

What do you do if one measurement in a set of supposedly identical ones seems quite different from the others, so that you suspect that you made a mistake somehow?

First, you look to see if there is evidence of a problem — an error in arithmetic done in the notebook, for example. Barring that, you have two choices to make.

1. In the absence of any real evidence to indicate a problem with the "outlier", you might decide that it simply should be kept. You may want to report the median instead of, or in addition to, the mean as your measurement of central value. The median is much less suceptible to the influence of outlier points.

2. You may apply a statistical test, discarding the outlier value if there is less than some critical probability that it came from the same parent distribution as the others.

The second option – applying a statistical test – is much more palatable if you actually know something about the parent distribution for your experiment. If you have only a few experimental points, and you don't have some good theoretical reason to postulate a particular form of parent distribution, it is almost certainly safer to retain the outlier point and report both the mean and median.

I recommend two statistical tests.

6.1 The Q test

This test is extremely easy to perform, and works well for samples with small numbers of points. It *assumes* that the experiment is governed by the normal distribution, but it does not assume that a good estimate of σ is available. It is not applicable to experiments with nonnormal error distributions, such as counting experiments with small numbers of counts per sample.

To perform a Q test, calculate the value of Q for your sample:

$$Q = \frac{|x_{\text{suspect}} - x_{\text{closest}}|}{|x_{\text{suspect}} - x_{\text{farthest}}|} \tag{38}$$

(The suspect value will, of course, be either the largest or smallest of the set.) If the value of Q is larger than the critical value given in Table 5 (from SGN) for your number N of measurements, you may discard the suspect value. You should then recalculate $\overline{x}$, S_{m}, and the confidence interval based on the new (smaller) number of observations.

Table 5 Critical values of Q at 90% confidence (from SGN).

N	3	4	5	6	7	8	9	10
Q_c	0.94	0.76	0.64	0.56	0.51	0.47	0.44	0.41

Note that the Q-test can be applied *only once* to a data set. If you have more than one screwball value, you must either live with them or redo the experiment.

In routine practice, the Q test at 90% confidence is acceptable. For particularly important observations, it is important to decide in advance what criterion will be used for data rejection. It is very easy to insert bias into the data analysis if the choice of rejection criterion is not made in advance.

Dean and Dixon [6] and Dixon and Massey [7] give background information on the Q test.

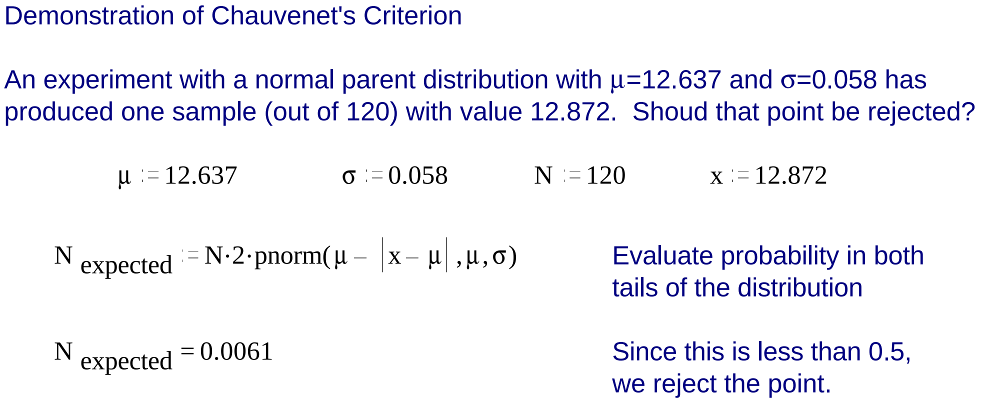
6.2 Chauvenet's criterion

A second statistical test is *Chauvenet's criterion*. It has the advantage that it can be used for any form of parent distribution, but the disadvantage that the parent distribution *and its parameters* must be known. (Note that the Q-test did not require a good value of σ.) The idea is simple: a data point should be rejected if the parent distribution predicts that fewer than half an event should appear which deviates as much from the mean as the questionable experimental point. (Fractions smaller than 1/2 may also be used, but should be agreed upon beforehand as discussed above.)

For example, suppose that you have performed a particular measurement in an automated way, and made 120 measurements on a sample. The histogram of your measurements indicates that the experiment does have a normal parent distribution. The mean of the measured values is 12.637 and the estimated standard deviation S, which should be a good estimate of σ since you have 120 samples, is 0.058. You have one point at 12.872 which you suspect to be erroneous. Should you reject that point?

You must calculate the number of measurements which are expected to lie at least as far afield as 12.872 on the basis of your knowledge of the parent distribution. Worksheet 5 shows a Mathcad worksheet which performs the calculation. It evaluates the probability that an individual measurement will lie as far away from the mean as 12.872, then multiplies by the number of measurements. Since the result is $< \frac{1}{2}$, the suspect point is rejected. Look carefully at the calculation of the probability to make sure you understand its use of the cumulative distribution function.

Worksheet 5 Mathcad worksheet illustrating Chauvenet's criterion.



Demonstration of Chauvenet's Criterion

An experiment with a normal parent distribution with μ=12.637 and σ=0.058 has produced one sample (out of 120) with value 12.872. Shoud that point be rejected?

$\mu := 12.637 \qquad \sigma := 0.058 \qquad N := 120 \qquad x := 12.872$

$N_{\text{expected}} := N \cdot 2 \cdot \text{pnorm}(\mu - |x - \mu|, \mu, \sigma)$

Evaluate probability in both tails of the distribution

$N_{\text{expected}} = 0.0061$

Since this is less than 0.5, we reject the point.

The criterion is used in exactly the same way with other distribution functions. You must know enough about your parent distribution to calculate (or look up) the relevant integral.

6.3 Choice of rejection criterion

For experiments where you have only a few measurements and expect the parent distribution to be approximately normal, the Q test is better. If you have at least 10 points, or you already know something about the parent distribution, Chauvenet's criterion is a good choice. If you don't know the form of the parent distribution and don't have very many samples, you have no statistical basis for rejecting the point and so must retain it.

7 Propagation of error

7.1 Formula approach

A quantity F is calculated from several measured quantities x, y, z:

$$F = F(x, y, z) \tag{39}$$

The total differential of F is

$$dF = \frac{\partial F}{\partial x}\,dx + \frac{\partial F}{\partial y}\,dy + \frac{\partial F}{\partial z}\,dz \tag{40}$$

(The partial derivative of F with respect to x, $\frac{\partial F}{\partial x}$, is calculated by taking the derivative of F with respect to x as though all the other variables were constants. See any introductory calculus text if you are not familiar with partial differentiation.) The total differential gives the infinitesimal change in F caused by infinitesimal changes in x, y, or z.

If we approximate the change in F brought about by small but finite changes in x, y, and z by a similar formula, we obtain

$$\Delta F = \frac{\partial F}{\partial x}\,\Delta x + \frac{\partial F}{\partial y}\,\Delta y + \frac{\partial F}{\partial z}\,\Delta z \tag{41}$$

This is equivalent to saying that the surface $F(x, y, z)$ is a plane over the region in space $[x \pm \Delta x, y \pm \Delta y, z \pm \Delta z]$; curvature over that small region is not important.

If the errors in x, y, and z are small and known (in both sign and magnitude), Eq. (41) can be used to propagate the errors and find the resulting error in F. On the other hand, in that case, it is both easier and more accurate to simply recalculate F using corrected values of x, y, and z. (Since you know the errors, you can do that.)

To handle random errors, we must perform averages. The average random error in F is zero, so we will calculate $\langle(\Delta F)^2\rangle$.

Square both sides of equation 41:

$$\begin{aligned}(\Delta F)^2 \quad = \quad & \left(\frac{\partial F}{\partial x}\right)^2 (\Delta x)^2 + \left(\frac{\partial F}{\partial y}\right)^2 (\Delta y)^2 + \left(\frac{\partial F}{\partial z}\right)^2 (\Delta z)^2 \\ & +2\left(\frac{\partial F}{\partial x}\right)\left(\frac{\partial F}{\partial y}\right)\Delta x\Delta y + 2\left(\frac{\partial F}{\partial x}\right)\left(\frac{\partial F}{\partial z}\right)\Delta x\Delta z \\ & +2\left(\frac{\partial F}{\partial y}\right)\left(\frac{\partial F}{\partial z}\right)\Delta y\Delta z \end{aligned} \tag{42}$$

Now we need to average $(\Delta F)^2$ over many determinations of F, each with different values of the errors in x, y, and z. If the the various errors in x, y, and z are

1. small,
2. symmetrically distributed about 0, and
3. uncorrelated,

the cross terms such as $\Delta y \Delta z$ are just as likely to be positive as negative in any one determination. Therefore, when we average the many determinations together, the cross terms will tend to add up to zero. The squared terms like $(\Delta x)^2$, though, will not. Therefore, when Eq. (42) is averaged over many data sets, we obtain the propagation of error equation:

$$(\Delta F)^2 = \left(\frac{\partial F}{\partial x}\right)^2 (\Delta x)^2 + \left(\frac{\partial F}{\partial y}\right)^2 (\Delta y)^2 + \left(\frac{\partial F}{\partial z}\right)^2 (\Delta z)^2. \tag{43}$$

In Eq. (43), the partial derivatives are evaluated at the mean values of x, y, and z. That equation is written for a result calculated from three independently measured quantities, but it should be obvious how it changes for other numbers of variables.

Note that you may use any form of error measure you like for the Δx; standard deviation, 95% confidence interval, etc., so long as you use the same form for each of the independent variables. The result ΔF will then be of that form.

7.1.1 Example

Say we need to know the number of moles of a gas from measurements of its pressure, volume, and temperature:

$$n = \frac{pV}{RT}. \tag{44}$$

We assume that the gas is ideal (a possible source of systematic error!) Perhaps our measurements of p, V, and T are

$$\begin{aligned} p &= 0.268 \pm 0.012\,\text{atm (e.s.d)} \\ V &= 1.26 \pm 0.05\,\text{L} \\ T &= 294.2 \pm 0.3\,\text{K}. \end{aligned}$$

We calculate

$$n = \frac{(0.268\,\mathrm{atm})(1.26\,\mathrm{L})}{(0.082058\,\frac{\mathrm{L\,atm}}{\mathrm{mol\,K}})(294.2\,\mathrm{K})} = 0.013988\,\mathrm{mol}. \tag{45}$$

For the propagation of errors formula, we must calculate partial derivatives:

$$\frac{\partial n}{\partial p} = \frac{V}{RT} \tag{46}$$

$$\frac{\partial n}{\partial V} = \frac{p}{RT} \tag{47}$$

$$\frac{\partial n}{\partial T} = -\frac{pV}{RT^2}. \tag{48}$$

So,

$$\begin{aligned}\Delta n &= \left[\left(\frac{V}{RT}\right)^2 (\Delta p)^2 + \left(\frac{p}{RT}\right)^2 (\Delta V)^2 + \left(\frac{pV}{RT^2}\right)^2 (\Delta T)^2\right]^{\frac{1}{2}} \\ &= \left[\left(\frac{1.26\,\mathrm{L}}{\left(0.082058\,\frac{\mathrm{L\,atm}}{\mathrm{mol\,K}}\right)(294.2\,\mathrm{K})}\right)^2 (0.012\,\mathrm{atm})^2\right. \\ &\quad + \left(\frac{0.268\,\mathrm{atm}}{\left(0.082058\,\frac{\mathrm{L\,atm}}{\mathrm{mol\,K}}\right)(294.2\,\mathrm{K})}\right)^2 (0.05\,\mathrm{L})^2 \\ &\quad \left. + \left(\frac{(0.268\,\mathrm{atm})(1.26\,\mathrm{L})}{\left(0.082058\,\frac{\mathrm{L\,atm}}{\mathrm{mol\,K}}\right)(294.2\,\mathrm{K})^2}\right)^2 (0.3\,\mathrm{K})^2\right]^{\frac{1}{2}} \\ &= 0.000837\,\mathrm{mol}.\end{aligned} \tag{49}$$

So we would report

$$n = (1.40 \pm 0.08) \times 10^{-2}\,\mathrm{mol}\ \text{(e.s.d).}$$

Always check the units to avoid mistakes. The uncertainty must have the same units as its quantity.

Note that I used a value of R, the gas constant, accurate to several more figures than my measurements. (Actually, six digits was overkill; four or five would have sufficed.) Accurate modern values for fundamental constants such as R can be found at the NIST Fundamental Physical Constants web site at `http://physics.nist.gov/constants`, and in *Journal of Physical and Chemical Reference Data*, volume 28, number 6 (1999). A new set of values, based on measurements done through the end of 1998, appeared in July 1999. Good tabulations (including the one in every August issue of *Physics Today*) give the uncertainties in the values, which are needed for precise work.

The propagation of error equation, Eq. (43), is useful only if the errors in the independent variables are uncorrelated; that is, if a positive error in x is just as likely to be associated with a negative error in y as a positive one. That will not be true if, for example, one is the slope and one is the intercept of a line fitted through a single data set. In that case, you may include the "covariance" terms, as discussed below in Section 7.1.3, or use the Monte Carlo methods we will discuss shortly.

7.1.2 Numerical differentiation

If the function $F(x,y,z)$ is complicated, but you have a programmable calculator or computer in front of you, you can often save time by evaluating the partial derivatives numerically rather than analytically. In general, evaluation of numerical derivatives is a dicey business, but error calculations don't need to be very precise so you can get away with pretty simple evaluations.

First: what you do *not* do is to calculate the upper and lower error limits on F as follows:

$$\begin{aligned} \Delta_+ F &= F(x+\Delta x, y+\Delta y, z+\Delta z) \\ \Delta_- F &= F(x-\Delta x, y-\Delta y, z-\Delta z) \end{aligned}$$

and report that the correct value of F lies between $\Delta_+ F$ and $\Delta_- F$. That procedure will give all sorts of ridiculous error estimates, depending on the particular functional form of $F(x,y,z)$. (Actually, it isn't so bad if F depends only on one variable, but I still don't recommend it. It's completely wrong if F depends on two or more variables.)

The correct technique is to evaluate the partial derivatives numerically, and use the regular propagation of error formula. Program your calculator to evaluate F given input variables x, y, z. Then evaluate the derivatives:

$$\frac{\partial F}{\partial x} \approx \frac{F(x+\Delta x, y, z) - F(x-\Delta x, y, z)}{2\Delta x} \tag{50}$$

(with similar formulas for y and z.) For each derivative you evaluate F twice, so if you have three independent variables you will evaluate it six times. (It's therefore almost imperative that you have an easy way to evaluate $F(x,y,z)$; thus the suggestion of a programmable calculator.) Then insert the numerical values of the partial derivatives you found into the propagation of errors formula, and calculate the uncertainty in F the normal way.

This technique saves you the trouble of doing all those analytic derivatives. On the other hand, if you found F several times with different values of the independent variables, you must recalculate the partials for each different value. In that case, it's often more efficient to just work out the analytic formula.

7.1.3 Full covariant propagation

If the independent variables you are using did not result directly from different independent measurements, then the assumption of equation 43 that the errors are uncorrelated may be violated. For instance, if the "independent" variables are parameters which resulted from a fit of a model function to some (x,y) data, the errors in the values will be correlated. (You might calculate some result from the slope and intercept of a least-squares line, for example.) In that case the cross terms will not cancel. Bevington and Robinson show on pp. 42–43 that the formula for propagation of errors is then

$$\begin{aligned}\sigma_F^2 &= \left(\frac{\partial F}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial F}{\partial y}\right)^2 \sigma_y^2 + \cdots \\ &\quad +2\left(\frac{\partial F}{\partial x}\right)\left(\frac{\partial F}{\partial y}\right)\sigma_{xy}^2 + \cdots, \qquad (51)\end{aligned}$$

where the errors in x and y are expressed as variances σ^2 and their correlation is expressed by the *covariance*

$$\sigma_{xy}^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}[(x_i-\overline{x})(y_i-\overline{y})]\right] \qquad (52)$$

Very often, the error in a calculated result will be far too large if the covariances are not included. For example, one student's data from Experiment 8, Heat of Vaporization of Water, gives a result for the heat of vaporization of 35.3 ± 0.4 kJ/mol at 350 K if the covariances are properly included, while the result is 35 ± 18 kJ/mol if they are left out of the calculation.

A good least-squares fitting routine will provide the covariance between x and y as one of its outputs. The Mathcad template we provide, `genlinls.mcd`, reports the covariances. In fact, the covariances are a natural result of a linear least-squares fitting procedure, and can therefore be provided with no additional computational effort. Far too many fitting programs which otherwise are perfectly fine for scientific use do not provide covariances. If you need to calculate a result involving more than one parameter obtained from the same fit, and you do not have the covariances available, you should use Monte Carlo estimation to do the propagation of error.

7.2 Monte Carlo approach

A second, and in fact superior, way to evaluate errors in a calculated quantity is to use a *Monte Carlo simulation* of your experiment and analysis. The idea is simple. Use a computer to generate many (perhaps 100–1000) synthetic data sets, just as we averaged over many hypothetical data sets in the section above. The data sets should have values of the independent variables drawn as closely as possible from the same parent distribution that applied in your experiment. Then, for each synthetic data set, calculate a value of F the same way you did for your real data set. Look at the list of resulting values of F. Find two values in that list which enclose 90%, 95%, or whatever of all the values. Those two values give your confidence limits.

The Monte Carlo method has several advantages over classical propagation of error.

- Often the errors in F are not normally distributed, even though the raw data x, y, z are normally distributed. The Monte Carlo method gives the correct distribution for F. That means it works correctly even when the assumption that the errors are small is violated.

- Even if the errors in the independent variables are not symmetrically distributed about 0, this method works correctly as long as the simulations are done with the correct parent distribution.

- Monte Carlo analysis can be done even when the evaluation of F from the data involves very complicated calculations such as nonlinear fits and Fourier transforms, so long as the calculations are automated.

The disadvantage is that one must both (1) have a computer with a good random number generator available, and (2) be pretty handy with it. It is possible to do Monte Carlo analysis “by hand”, but I’ve never known anyone to actually do it.

7.2.1 Generating synthetic data sets

The idea is to use the computer to simulate many experiments. Perhaps you measured the pressure, temperature, and volume of a gas sample in order to calculate the number of moles. You perform each of the measurements several times, so that you have some idea of the parent distribution for each quantity. If you have made many measurements with your apparatus, and know that the results are normally distributed and their standard deviations, you can use the normal distribution for generating your synthetic data sets. If you are new to the apparatus, and have only a few measurements of each quantity, then you can still use the normal distribution but you must use your estimated standard deviations of the mean S_{m}. (Occasionally, other parent distributions are needed because of the nature of the experiments.) From your measurements, you calculate the average values of p, V, and T, their estimated standard deviations, and the e.s.d. of the mean values, just as you have done in the past.

Then you need to use the computer to generate many (p, V, T) triples to subject to analysis. Each generated value should be drawn randomly from the parent distribution that controls your experimental results. Usually you use a normal distribution with the appropriate mean (the mean value $\bar{x}$ from your experiment) and standard deviation σ_{m} (the known standard deviation of the mean for your experiment).

The generation of "random" numbers from specified distributions belongs to an interesting branch of computer science called "seminumerical algorithms." I will not discuss the rather deep mathematics behind construction of good random number generators, but will assume that you have one available. Mathcad 6.0 has a built-in function rnorm(m, μ, σ) that generates m random numbers drawn from the normal distribution with mean μ and standard deviation σ.

Many other programs provide only a function that provides uniformly distributed random numbers on the domain $[0, 1]$. In that case, is not too difficult to obtain numbers drawn from other distributions. *Numerical Recipes* and Bevington and Robinson discuss the techniques. In particular, it is easy to obtain normally distributed numbers from uniformly distributed ones by using the *Box-Muller method*, discussed in both the above references.

If you have random numbers r_i from a normal distribution with mean zero and standard deviation one, it is easy to get numbers s_i from a distribution with mean μ and standard deviation σ; just evaluate $s_i = \sigma r_i + \mu$.

Now we have enough information to generate synthetic data sets with appropriately distributed errors. For each independent variable, which in your experiment had mean $\overline{p}$ and estimated standard deviation of the mean $S_{\mathrm{m}p}$, generate lots of random numbers r_i from the normal distribution (mean $\overline{p}$, s.d. $S_{\mathrm{m}p}$). Do this for each of your independent variables, and you now have many computer-generated data sets.

If the relevant uncertainties were given to you by someone else as fractional uncertainties ("the pressure measurement is uncertain by 0.4%"), then calculate your synthetic data sets as

$$p_i = \overline{p}(1 + r_i(0.004)), \tag{53}$$

where the r_i are random numbers from the normal distribution with mean 0 and standard deviation 1.

7.2.2 Analyzing synthetic data

This is easy: you just do exactly the same things to each synthetic data set that you did to your real data set. Since you have hundreds or thousands of synthetic data sets, clearly the analysis procedure should be automated. In the end you should have a long list of values of F.

7.2.3 Analyzing resulting collection of Fs

There are at least two ways of using the list of F values you obtain to get confidence limits. I recommend you do both. One is good at giving you a feeling for the importance of errors in your experiment and the probability distribution of the result, and the other is more convenient for getting out numerical error limits.

Histogram method You now have a collection of calculated Fs. You can make a histogram of them, by making a plot of the number of values which fall within various intervals. (Many programs can do this for you; it's usually called a "histogram" or a "frequency plot.") Often a choice of interval width which gives a total of 15-20 bins works well. That histogram gives you a visual description of the probability distribution of F. If the histogram looks like a normal distribution, you can crudely estimate σ_F just by picking values off the x axis where the histogram has about 60% of its maximum height. More accurately, you can directly evaluate the standard distribution of your collection of Fs in the same way you do for regular experimental data. The real reason for making the histogram, though, is to see whether F really is distributed normally; often it isn't.

Sorting method This method is quick and easy, and gives good confidence limits, but doesn't give as good a feel for the F distribution as the histogram method. Get the computer to sort the collection of Fs from lowest to highest. Then just read confidence limits from that list. If you want 95% confidence limits, for example, and you have 1000 simulated points, the 25th value from each end of the list (values 25 and 975) give the lower and upper confidence limits. You can also do this graphically; plot the F values on the y axis vs. their positions in the sorted list, and find the positions on the x axis which correspond to the desired confidence interval. The corresponding y values give the upper and lower limits.

7.2.4 Example

I will treat the same example I used in the analytic case, the calculation of number of moles in a gas sample from its pressure, volume, and temperature. Look back at page 25; there we had the experimental results

$$\begin{aligned} p &= 0.268 \pm 0.012\,\mathrm{atm(e.s.d)} \\ V &= 1.26 \pm 0.05\,\mathrm{L} \\ T &= 294.2 \pm 0.3\,\mathrm{K}. \end{aligned}$$

Worksheet 6 shows a Monte Carlo error propagation on these data with Mathcad.

Most data analysis and plotting programs will be able to do simulations in a style very similar to that I used in the example. Modern programmable calculators can also do a nice job; you need to be able to accumulate the simulated results in an array, and the calculator may or may not be able to perform sorts and histograms, but you can usually evaluate the standard deviation of the array of F values to get an error estimate.

In a spreadsheet, it is easiest to make each synthetic experiment a single row; you will then have as many rows as you have simulated experiments. Most spreadsheets offer a uniform random number generator; only a few have a built-in normal distribution generator, so you must use the Box-Muller method mentioned above. Nonetheless, Monte Carlo simulations in spreadsheets are fairly easy.

I encourage you to use Monte Carlo error analysis on at least one of your laboratory experiments. It's a good skill to have.

Worksheet 6 Monte Carlo calculation in Mathcad.

Monte Carlo Error Analysis

Define index and constants $npts := 1000$ $i := 0 .. npts - 1$ $R := 0.0821$

Establish means and std. devs from expt $pbar := 0.268$ $\sigma pbar := 0.012$

$Vbar := 1.26$ $\sigma Vbar := 0.05$

$Tbar := 292.4$ $\sigma Tbar := 0.3$

Set up vectors of synthetic data

$p := rnorm(npts, pbar, \sigma pbar)$

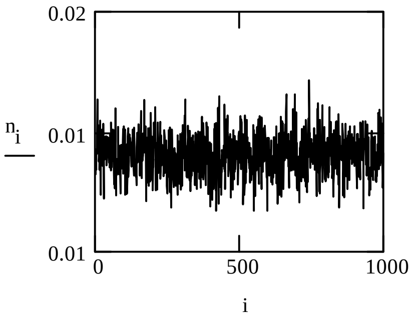
$V := rnorm(npts, Vbar, \sigma Vbar)$

$T := rnorm(npts, Tbar, \sigma Tbar)$

the number of moles is $n_i := \dfrac{p_i \cdot V_i}{T_i \cdot R}$

calculate the mean and standard deviation of the number of moles

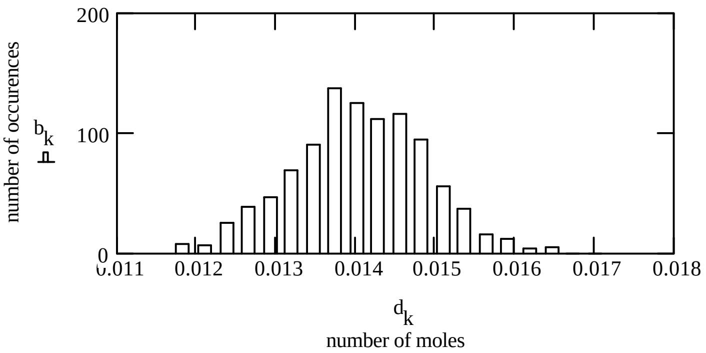
$nbar := mean(n)$ $\sigma nbar := stdev(n)$



Make a histogram

$nbins := 20$ $binwidth := \dfrac{max(n) - min(n)}{nbins}$

index for number of bins	define bins	count occurences in n between c(j) and c(j+1)	make x-index at center of bin
$jk := 0 .. nbins - 1$	$c_j := min(n) + j \cdot binwidth$	$b := hist(c, n)$	$d_k := c_k + \dfrac{binwidth}{2}$



GCM January 2, 2001

Worksheet 6 (continued)

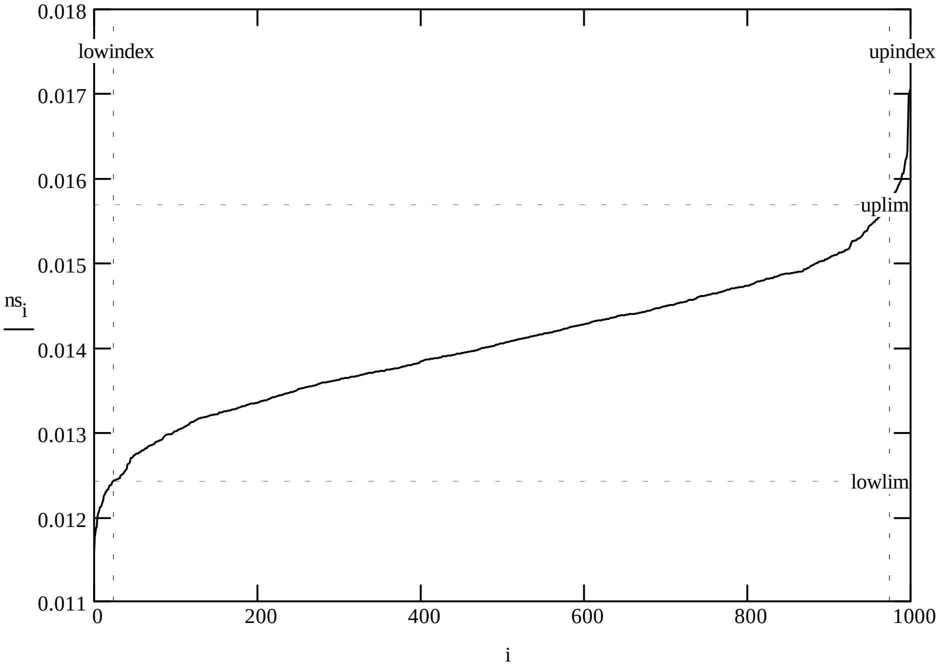
Sorting method

$ns := \mathrm{sort}(n)$ $\qquad conf := .95$

$upindex := \mathrm{ceil}\left[npts\cdot\left(1-\dfrac{1-conf}{2}\right)\right]$ $\qquad lowindex := \mathrm{floor}\left[npts\cdot\left(\dfrac{1-conf}{2}\right)\right]$

$lowindex = 25$ $\qquad lowlim := ns_{lowindex}$ $\qquad lowlim = 0.012$

$upindex = 975$ $\qquad uplim := ns_{upindex}$ $\qquad uplim = 0.016$



7.3 Reporting computed quantities

The mean, standard deviation of the mean, and confidence limits are the standard language for reporting of calculated quantities. If you have done analytic or numerical propagation of error where you used the standard deviations of x, y, etc. as the uncertainties, then you must report the uncertainty in the calculated result as the estimated standard deviation. (There is no simple way to get confidence limits then, since it's not clear what number of degrees of freedom, ν, to use.) If you used confidence limits as the uncertainty inputs to the propagation of error, report the resulting uncertainty in F as its confidence limit.

If you used Monte Carlo estimation, then look at the histogram; if the distribution of F is approximately Gaussian, you can report its standard deviation or confidence limits from the sorting method, as you choose. If the distribution of F does not look Gaussian, then at least you should use the sorting method to provide confidence limits. You might also want to report something about the shape of the F distribution (you could even include a plot.). In any case, you should report that your error estimates were obtained by Monte Carlo simulation of your experiment.

8 Significance tests for differences

Often it is useful to know whether two independent results are "significantly different". In some cases one result is known to be of high accuracy, and can be regarded as correct. Then the second result is generally compared to determine whether some more convenient measurement method is free of systematic error. In that case the "significance test for the difference" is just what you expect: you calculate the confidence interval at the desired confidence level (95% or whatever) $\overline{x} \pm \Delta$ on the unsure result, and if it contains the accurate value then no significant difference has been shown to exist.

In the second case, two results are being compared which may both be "unsure". Then a value D, the difference between the two, is calculated, with its own confidence interval ΔD. If the confidence interval includes the value 0, then the observed difference is not significant.

To find ΔD, use propagation of error:

$$D = \overline{x}_1 - \overline{x}_2 \tag{54}$$

$$S_D^2 = \left(\frac{\partial D}{\partial \overline{x}_1}\right)^2 S_{\overline{x}_1}^2 + \left(\frac{\partial D}{\partial \overline{x}_2}\right)^2 S_{\overline{x}_2}^2 \tag{55}$$

$$= S_{\mathrm{m1}}^2 + S_{\mathrm{m2}}^2 \tag{56}$$

Now choose a confidence level P; 95% is typical, and the appropriate $\nu = N_1 + N_2 - 2$. Find t in Table 4, or evaluate it with the Mathcad function $\mathrm{qt}(p, d)$:

$$t = -\mathrm{qt}(\frac{1-P}{2}, \nu). \tag{57}$$

Then $\Delta D = tS_D$, and

$$D = (\overline{x}_1 - \overline{x}_2) \pm tS_D \tag{58}$$

Therefore, if $tS_D \geq |D|$, the confidence interval on the difference includes 0 and the difference is not significant.

SGN and *Numerical Recipes* give a somewhat more complicated formula for S_D, which is better if N_1 and N_2 are not nearly the same. If $N_1 \approx N_2$, the two formulas give nearly the same result.

There is a subtlety in this testing of differences. If the question is simply one of difference ("Did these two students get significantly different results for the barometric pressure?"), then in finding the value of t in Table 4, use the confidence intervals in the row labeled P, or the Mathcad formula given above. If, though, the question "has a direction" ("Is Sue's result *smaller* than John's result?"), then choose from the row labeled P′, or use

$$t = \mathrm{qt}(P', d). \tag{59}$$

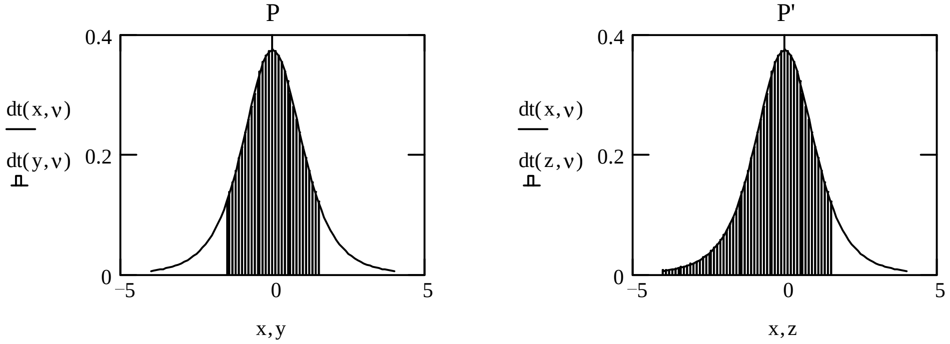
The different values of t come from different areas being evaluated under the Student-t distribution. The first question uses the integral of the distribution from $-t$ to t, and gives a so-called "double-tailed confidence interval"; the second uses the integral from $-\infty$ to t, and gives the "single-tailed" interval. See Worksheet 7.

9 Modeling of data

(The material in this section comes largely from *Numerical Recipes* chap. 15, SGN (6th ed.) chapter XXII, and Bevington and Robinson. The notation is closest to that of SGN.)

Worksheet 7 Two types of difference tests.

$x := -4, -3.9 .. 4$ $\quad$ $y := -1.5, -1.4 .. 1.5$ $\quad$ $z := -4, -3.9 .. 1.5$ $\quad$ $\nu := 4$



We have a data set $\{y_i, \sigma_i, \mathbf{x}_i\}$. The y_i are values of a "dependent" variable which we measured at N different values of the "independent" variable(s) $\mathbf{x}$. The σ_i are the standard deviations of the y_i; I will assume throughout this discussion that the $\mathbf{x}_i$ contain no error.[2]

We also have (usually) a physical model of the system, implying a functional relationship between y and $\mathbf{x}$:

$$y(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\alpha}) \tag{60}$$

In Eq. (60) the independent (measured or controlled) variables are represented by $\mathbf{x}$; the M components of $\boldsymbol{\alpha}$, that is, $\alpha_0, \alpha_1, \cdots \alpha_{M-1}$, are *adjustable parameters*.

An example is the integrated Clausius-Clapeyron equation describing the variation of the vapor pressure of a pure liquid with T:

$$\ln\left(\frac{p}{p_0}\right) = C - \frac{\Delta H_{\mathrm{vap,m}}}{RT} \tag{61}$$

Here $\mathbf{x}$ corresponds simply to T, and the model parameters are C and $\Delta H_{\mathrm{vap,m}}$.

In a modeling problem we want to find

[2] *Numerical Recipes* discusses fitting a straight line with errors in both coordinates; for the general case, see Jefferys, *Astronomical Journal* **85**, 177 (1980), and **86**, 149 (1981), and the errata in **95**, 1299–1300.

1. the "best" values of the model parameters $\alpha_1 \cdots \alpha_{M-1}$
2. uncertainties (confidence limits) of those parameters
3. an indication of whether the model actually "fits" the data at all.

We can't calculate from a data set the probability that any particular parameter set is correct. The approach is to calculate the probability of obtaining our actual data given any set of parameters, and choose the parameter set which gives the greatest probability of finding that data set. That corresponds again to the principle of maximum likelihood: we proceed by assuming that our data set is the *most probable* one.

If the error distributions of the y_i are normal, there is no or negligible error in the values of the $\mathbf{x}_i$, and we know that the model is correct, the probability density for obtaining a single one of the y_i is

$$P(y_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{[y_i - f(\mathbf{x}_i; \boldsymbol{\alpha})]^2}{\sigma_i^2}\right\}. \tag{62}$$

Since the different y_i measurements are independent, calculating the probability density of the entire data set given the parameters is straightforward: it's just the product of the probability densities for obtaining each y_i independently.

$$\begin{aligned} P(\{y_i\}) &= \prod_{i=1}^{N} P(y_i) \qquad (63) \\ &= \left(\prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}}\right) \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{[y_i - f(\mathbf{x}_i; \boldsymbol{\alpha})]^2}{\sigma_i^2}\right\} \end{aligned}$$

We want to maximize this probability by varying the components of $\boldsymbol{\alpha}$. Since the part outside the exponential is independent of the values of the α_j, we can just maximize the exponential, which is the same as minimizing the value of χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(\mathbf{x}_i; \boldsymbol{\alpha})]^2}{\sigma_i^2} \tag{64}$$

This is called least-squares fitting, or χ^2 fitting. Its algorithm is to adjust the parameters to minimize the sum of the squared vertical deviations of the data from the model. Some things to be noted:

- If your data are not really described by a normal error distribution, a least-squares fit will always pay too much attention to the "outlier" points. This can be a real problem.

- If the errors are not normally distributed but you know the error distribution, you can still use the principle of maximum likelihood. This will give much better results for non-Gaussian systems (for instance, "counting" experiments with small numbers of counts.) Chapter 10 in Bevington and Robinson discusses the technique.

9.1 Linear least squares

Least squares fitting problems generally fall into two groups, the *linear* and the *nonlinear* problems, which require somewhat different numerical approaches. I'll deal with the linear problems first. In a linear fitting problem, the model function (Eq. (60) above) has the special form

$$y(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\alpha}) = \sum_{j=0}^{M-1} \alpha_j f_j(\mathbf{x}). \tag{65}$$

Note that in the linear problem the f_j "basis functions" do not depend on $\boldsymbol{\alpha}$.

The most familiar example of such a model function is the straight-line model, where there is only one independent variable x, and the model function is

$$y = \alpha_0 + \alpha_1 x. \tag{66}$$

The basis functions are $f_0(x) = 1$ and $f_1(x) = x$. It is easy to see how to extend this example to fitting functions which are polynomials of arbitrary order, by adding terms to the model which look like $\alpha_j x^j$.

Another example which occurs often in physical chemistry is

$$y = \alpha_0 + \frac{\alpha_1}{x}, \tag{67}$$

where the basis functions are $f_0(x) = 1$ and $f_1(x) = 1/x$.

9.1.1 Finding $\boldsymbol{\alpha}$

We need to minimize χ^2 by varying the α_j. The standard way to do that is to take its derivative with respect to each of the α_j and set it equal to 0. In the linear least squares problem, we can solve the resulting M equations to get $\alpha_0 \cdots \alpha_{M-1}$. That solves the first part of the problem (finding the best values of the model parameters.)

Inserting the linear model function (65) into the definition of χ^2 (64), we obtain

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - \sum_{j=0}^{M-1} \alpha_j f_j(\mathbf{x}_i)]^2}{\sigma_i^2} \tag{68}$$

Taking the partial derivative with respect to a particular adjustable parameter α_k gives

$$\frac{\partial \chi^2}{\partial \alpha_k} = \sum_{i=1}^{N} \left(\frac{1}{\sigma_i^2}\right) (2) \left[y_i - \sum_{j=0}^{M-1} \alpha_j f_j(\mathbf{x}_i)\right] (-f_k(\mathbf{x}_i)) \tag{69}$$

Setting that derivative equal to 0 and dividing both sides by -2 gives

$$\begin{aligned} 0 &= \sum_{i=1}^{N} \frac{f_k(\mathbf{x}_i)}{\sigma_i^2} \left[y_i - \sum_{j=0}^{M-1} \alpha_j f_j(\mathbf{x}_i)\right] && (70) \\ &= \sum_{i=1}^{N} \frac{y_i f_k(\mathbf{x}_i)}{\sigma_i^2} - \sum_{i=1}^{N} \frac{f_k(\mathbf{x}_i)}{\sigma_i^2} \sum_{j=0}^{M-1} \alpha_j f_j(\mathbf{x}_i). && (71) \end{aligned}$$

Now, separating the sum which involves the y_i, and moving the $\frac{f_k(\mathbf{x}_i)}{\sigma_i^2}$ term inside the sum over j, we obtain

$$\sum_{i=1}^{N} \sum_{j=0}^{M-1} \frac{f_k(\mathbf{x}_i)}{\sigma_i^2} \alpha_j f_j(\mathbf{x}_i) = \sum_{i=1}^{N} \frac{y_i f_k(\mathbf{x}_i)}{\sigma_i^2}. \tag{72}$$

Interchanging the order of summations (which is legal since both sums are over finite numbers of terms) and rearranging slightly gives

$$\sum_{j=0}^{M-1} \sum_{i=1}^{N} \frac{f_k(\mathbf{x}_i) f_j(\mathbf{x}_i)}{\sigma_i^2} \alpha_j = \sum_{i=1}^{N} \frac{y_i f_k(\mathbf{x}_i)}{\sigma_i^2}. \tag{73}$$

We have one equation like Eq. (73) for each value of k, that is, for each basis function. We therefore have M of these equations for the M unknown parameters α_k. They are called the *normal equations* of the least-squares problem.

The normal equations can be written in a compact form by defining the matrix A and the vector $\mathbf{h}$ as follows:

$$\mathsf{A}_{kj} = \sum_{i=1}^{N} \frac{f_k(\mathbf{x}_i) f_j(\mathbf{x}_i)}{\sigma_i^2}, \tag{74}$$

$$\mathbf{h}_k = \sum_{i=1}^{N} \frac{y_i f_k(\mathbf{x}_i)}{\sigma_i^2}. \tag{75}$$

Note that all the components of A and $\mathbf{h}$ can be calculated directly from the data. Now the normal equations are simply

$$\mathsf{A}\boldsymbol{\alpha} = \mathbf{h}. \tag{76}$$

Many methods are available for solving Eq. (76). For our purposes, it is best to *invert* the matrix A with a technique such as Gauss-Jordan elimination (which is built into Mathcad) and then multiply both sides of Eq. (76) by the inverse matrix $\mathsf{B} = \mathsf{A}^{-1}$, obtaining

$$\boldsymbol{\alpha} = \mathsf{B}\mathbf{h} \tag{77}$$

Thus we obtain the values of the best-fit model parameters $\alpha_0 \cdots \alpha_{M-1}$.

9.1.2 Uncertainties in the parameters

The value of a single parameter α_k is given by Eq. (77) as

$$\alpha_k = \sum_{j=0}^{M-1} \mathsf{B}_{kj}\mathbf{h}_j. \tag{78}$$

Only the y_i, which are used to compute $\mathbf{h}$, have errors. We therefore find the standard deviation in α_k by propagation of error, using the y_i as our "independent variables":

$$\sigma_{\alpha_k}^2 = \sum_{i=1}^{N} \left(\frac{\partial \alpha_k}{\partial y_i}\right)^2 \sigma_i^2 \tag{79}$$

It takes about one page of work to show that

$$\sigma_{\alpha_k}^2 = \mathsf{B}_{kk}, \tag{80}$$

and similarly, the covariances are

$$\sigma_{\alpha_j\alpha_k}^2 = \mathsf{B}_{jk}. \tag{81}$$

(I will spare you the details; Bevington and Robinson do the calculation explicitly on p. 123.) The matrix B is therefore called the *variance-covariance matrix*: its diagonal elements give the variances (squared standard deviations) of the individual fitting parameters, and its off-diagonal elements give the covariances between pairs of parameters.

Note that B does not depend at all on the y_i; it comes from inverting A which is calculated just from the $\mathbf{x}_i$ and σ_i. Its information about the uncertainties in $\boldsymbol{\alpha}$ comes therefore purely from the stated uncertainties in y_i, namely the σ_i. If the σ_i were misstated, the calculated uncertainties in the parameters will be wrong. If the model function is known to provide a good description of the data, then the calculated uncertainties can be "corrected" by estimating the standard deviation in y from the deviations of the y_i from the best-fit model. This correction will be described in Section 9.1.4 below. Using true standard deviations in the fit, though, has a very important advantage: it permits you to determine whether the model function actually describes the data. That is the next topic.

9.1.3 Goodness of fit

We need a goodness-of-fit measure, in order to have some indication of whether the model describes the data in a reasonable way at all. It is possible to apply the formulas above to any set of (x, y) data, whether or not they resemble the model function! We obtain the goodness of fit by determining the probability that we would get a χ^2 as bad as the one we have even though the model was correct.

Mathcad provides a cumulative chi-squared distribution function called $\mathtt{pchisq}(x, d)$. To estimate a goodness-of-fit parameter Q, evaluate

$$Q = 1 - \mathtt{pchisq}(\chi^2, N - M). \tag{82}$$

If the errors are normally distributed, the model is correct, and your estimates of the measurement errors σ_i are good, you should get $Q \geq 0.1$ or so. If you have underestimated the errors, or your experiment does not really have normally distributed errors, correct models can sometimes give values of Q as low as perhaps 10^{-3}. Genuinely wrong models often give $Q \ll 10^{-3}$. This method of checking goodness-of-fit works for all linear least squares problems, and works decently for nonlinear problems as well. If you don't have Mathcad handy, tables of the cumulative chi-squared distribution are given in many books including Young (p. 163) and Bevington and Robinson (Table C.4).

9.1.4 Fitting with unknown errors

Everything up to now has assumed that you know the correct standard deviations for your y_i, and have put them into the fitting program as the σ_i. Sometimes the standard deviations for all the measurements are the same. That's okay, as long as each y really does have that standard deviation.

If you do not know the true errors σ_i, then it is not possible to obtain an estimate of goodness of fit, and you have no statistical arguments available to help you claim that your model function is a realistic one. However, from the deviations of the individual points from the fitted function, it is still possible to estimate the uncertainties in the parameters, *assuming* that the model is correct for the data. To do that, set all $\sigma_i = 1$, do the fit described above, then multiply each element of the resulting variance-covariance matrix by $\chi^2/(N-M)$. The standard deviations of the individual fitted parameters are then the square roots of the diagonal elements of the new variance-covariance matrix. An estimate of the standard deviation of the y_i is then

$$\sigma_y \approx \sqrt{\chi^2/(N-M)}. \tag{83}$$

Keep in mind that this estimate is only sensible if the model function is known in advance to be a good description of the data.

9.1.5 Fitting with relative errors

Occasionally you will not know the absolute standard deviations for individual y measurements, but you will know that some of the measurements have higher precision than others. You need to be able to "tell" the fitting algorithm to pay more attention to the more precise points and worry less about not passing close to the more uncertain points. In that case, you can generate a list of "relative standard deviations" for the points, which presumably are all related to the true standard deviations by multiplication by some unknown constant. Put the values of your relative σ_i into the fitting routine above. At the end of the fit, scale the variance-covariance matrix B as described above in section 9.1.4 to get error estimates for the parameters, and multiply your relative standard deviations by $\sqrt{\chi^2/(N-M)}$ to get estimates of the true standard deviations of the y_i. This procedure still does not allow you to make an independent assesment of goodness of fit, but it does give better values of the fitted parameters $\boldsymbol{\alpha}$ than you would get by setting all the $\sigma_i = 1$. If your assigned relative uncertainties were in fact close to the true ones, and the model is good, then $\sqrt{\chi^2/(N-M)}$ will be close to 1.

9.1.6 Mathcad implementation

Worksheet 8 shows the program `genlinls.mcd`, which performs almost exactly the calculations described here. It uses a single independent variable x rather than a vector $\mathbf{x}$, as is appropriate for most physical chemistry applications. It uses a variable `abserrs` to determine whether you have put in the absolute uncertainties. If `abserrs` is 1, it does not scale the variance-covariance matrix. If `abserrs` is zero, the errors in the parameters and their covariances are estimated from the deviations of the data from the fit, assuming the model is correct, as described in Section 9.1.4 and 9.1.5.

9.1.7 Pathologies

The normal equations approach to linear least squares is relatively easy to understand, and programs implementing it are compact and fast. However, it is suceptible to one common disease of least-squares procedures. If the data do not clearly distinguish between two of your basis functions, it is possible for the fitted coefficients of those functions (the fitted parameters α) to have nearly meaningless values even though the fit through the data looks reasonable. You should suspect this problem when some pair of your fitted parameters have very large covariances. In this case, it is necessary to use a different least-squares procedure based on *singular value decomposition*. See *Numerical Recipes* for details. Mathcad Plus has SVD built in, so it can be used for the task.

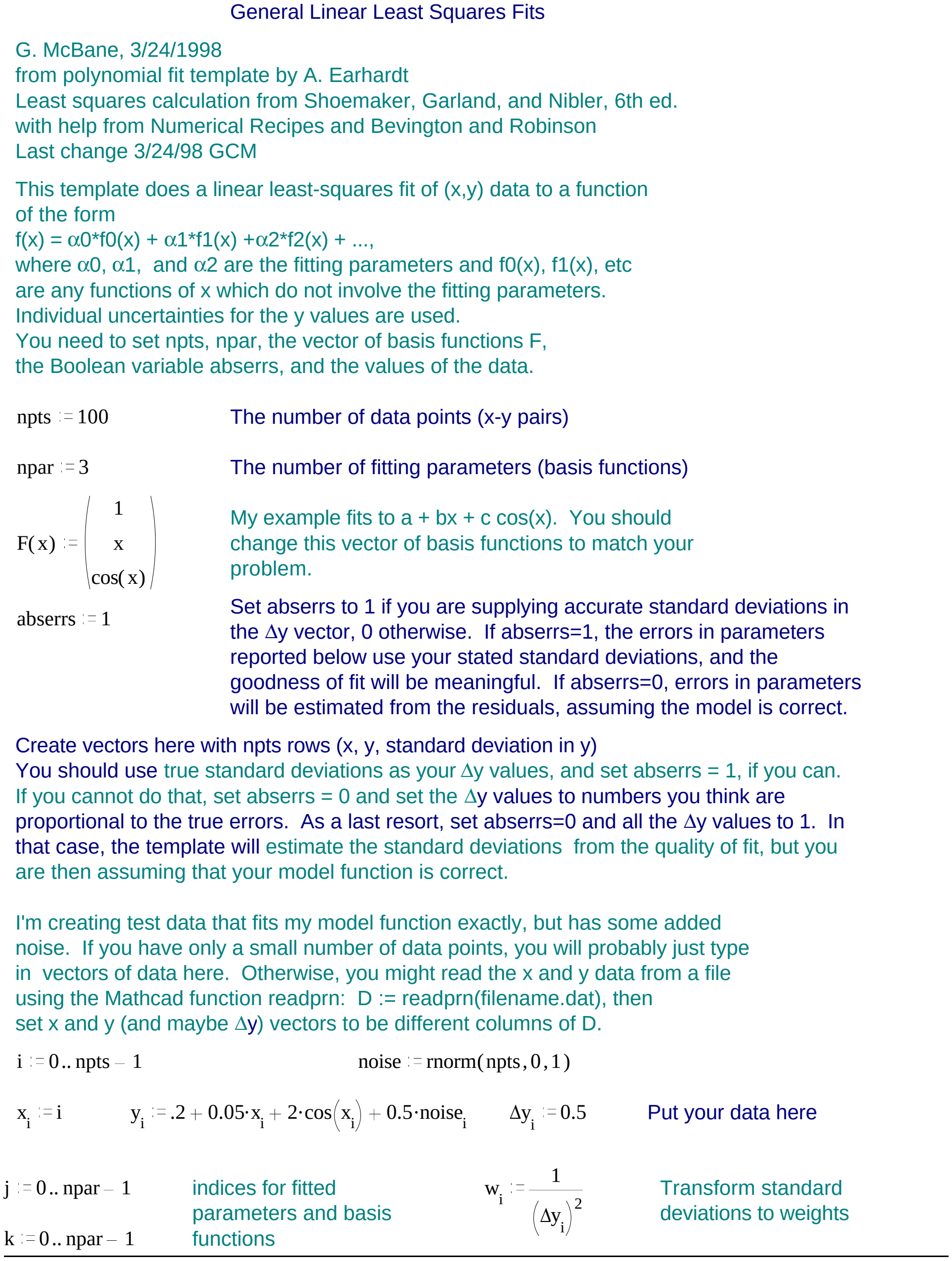
9.2 Nonlinear models

Often the physical model function will vary nonlinearly with one or more parameters; one common example is the exponential model,

$$y = ae^{-kx}, \tag{84}$$

where a and k are the parameters to be obtained by fitting a set of (x, y) data. Differentiating χ^2 with respect to a and k and setting the derivatives equal to 0 gives a set of nonlinear equations, which cannot be solved in a simple way. The algebraic methods useful for fitting linear models are therefore not applicable.

Worksheet 8 Linear least squares calculation in Mathcad.



General Linear Least Squares Fits

G. McBane, 3/24/1998
from polynomial fit template by A. Earhardt
Least squares calculation from Shoemaker, Garland, and Nibler, 6th ed.
with help from Numerical Recipes and Bevington and Robinson
Last change 3/24/98 GCM

This template does a linear least-squares fit of (x,y) data to a function of the form
f(x) = α0*f0(x) + α1*f1(x) +α2*f2(x) + ...,
where α0, α1, and α2 are the fitting parameters and f0(x), f1(x), etc are any functions of x which do not involve the fitting parameters.
Individual uncertainties for the y values are used.
You need to set npts, npar, the vector of basis functions F,
the Boolean variable abserrs, and the values of the data.

$\mathrm{npts} := 100$ The number of data points (x-y pairs)

$\mathrm{npar} := 3$ The number of fitting parameters (basis functions)

$$F(x) := \begin{pmatrix} 1 \\ x \\ \cos(x) \end{pmatrix}$$

My example fits to a + bx + c cos(x). You should change this vector of basis functions to match your problem.

$\mathrm{abserrs} := 1$

Set abserrs to 1 if you are supplying accurate standard deviations in the Δy vector, 0 otherwise. If abserrs=1, the errors in parameters reported below use your stated standard deviations, and the goodness of fit will be meaningful. If abserrs=0, errors in parameters will be estimated from the residuals, assuming the model is correct.

Create vectors here with npts rows (x, y, standard deviation in y)
You should use true standard deviations as your Δy values, and set abserrs = 1, if you can. If you cannot do that, set abserrs = 0 and set the Δy values to numbers you think are proportional to the true errors. As a last resort, set abserrs=0 and all the Δy values to 1. In that case, the template will estimate the standard deviations from the quality of fit, but you are then assuming that your model function is correct.

I'm creating test data that fits my model function exactly, but has some added noise. If you have only a small number of data points, you will probably just type in vectors of data here. Otherwise, you might read the x and y data from a file using the Mathcad function readprn: D := readprn(filename.dat), then set x and y (and maybe Δy) vectors to be different columns of D.

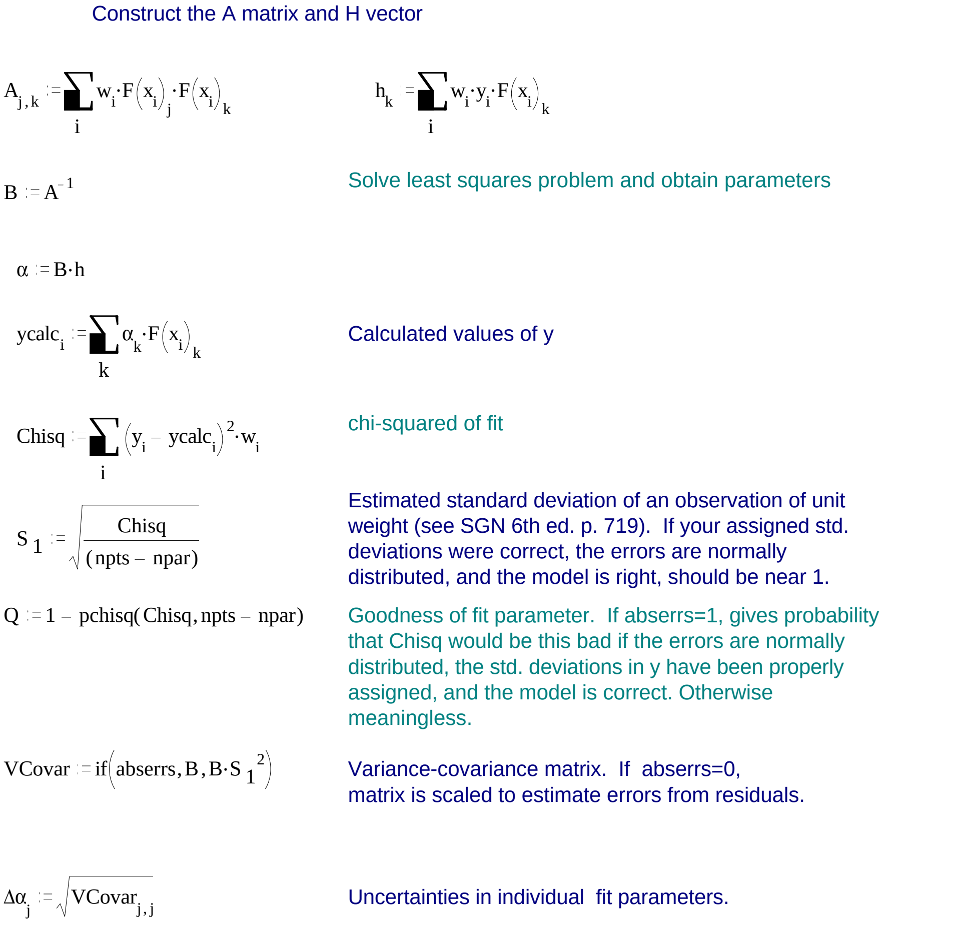
$i := 0 .. \mathrm{npts} - 1$ $\mathrm{noise} := \mathrm{rnorm}(\mathrm{npts}, 0, 1)$

$x_i := i$ $y_i := .2 + 0.05 \cdot x_i + 2 \cdot \cos\left(x_i\right) + 0.5 \cdot \mathrm{noise}_i$ $\Delta y_i := 0.5$ Put your data here

$j := 0 .. \mathrm{npar} - 1$ indices for fitted parameters and basis functions $w_i := \dfrac{1}{\left(\Delta y_i\right)^2}$ Transform standard deviations to weights

$k := 0 .. \mathrm{npar} - 1$

Worksheet 8 (continued)



Construct the A matrix and H vector

$A_{j,k} := \sum_i w_i \cdot F(x_i)_j \cdot F(x_i)_k$ $h_k := \sum_i w_i \cdot y_i \cdot F(x_i)_k$

$B := A^{-1}$ Solve least squares problem and obtain parameters

$\alpha := B \cdot h$

$ycalc_i := \sum_k \alpha_k \cdot F(x_i)_k$ Calculated values of y

$Chisq := \sum_i (y_i - ycalc_i)^2 \cdot w_i$ chi-squared of fit

$S_1 := \sqrt{\frac{Chisq}{(npts - npar)}}$ Estimated standard deviation of an observation of unit weight (see SGN 6th ed. p. 719). If your assigned std. deviations were correct, the errors are normally distributed, and the model is right, should be near 1.

$Q := 1 - pchisq(Chisq, npts - npar)$ Goodness of fit parameter. If abserrs=1, gives probability that Chisq would be this bad if the errors are normally distributed, the std. deviations in y have been properly assigned, and the model is correct. Otherwise meaningless.

$VCovar := if(abserrs, B, B \cdot S_1{}^2)$ Variance-covariance matrix. If abserrs=0, matrix is scaled to estimate errors from residuals.

$\Delta\alpha_j := \sqrt{VCovar_{j,j}}$ Uncertainties in individual fit parameters.

9.2.1 "Linearizable" models

In the example above, it is possible to transform the data in a way that gives a linear model:

$$\begin{aligned} \ln y &= \ln(a) - kx \qquad (85) \\ &= a' - kx, \end{aligned}$$

so that plotting $\ln(y)$ vs. x should give a straight line with intercept $\ln(a)$ and intercept $-k$. Linear least squares may be used to fit the transformed data. This is a completely legitimate trick, and is often useful for simple models. However, to use it correctly, you *must* use the appropriate weighting (uncertainties) for the "transformed" y values.

Worksheet 8 (continued)

Results

Parameters and individual uncertainties

Variances and covariances

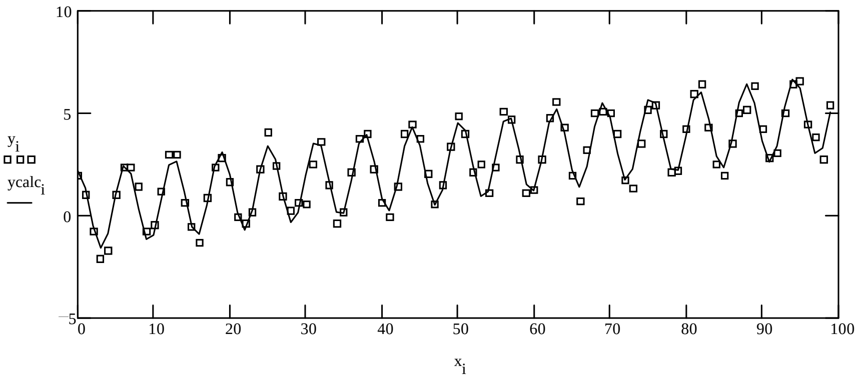
	α_j	$\Delta\alpha_j$
α0	0.231732	0.09929
α1	0.047838	$1.733159\cdot10^{-3}$
α2	1.97936	0.070762

$$\mathrm{VCovar} = \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 3.004\cdot10^{-6} & 4.211\cdot10^{-6} \\ 0 & 4.211\cdot10^{-6} & 0.005 \end{bmatrix}$$

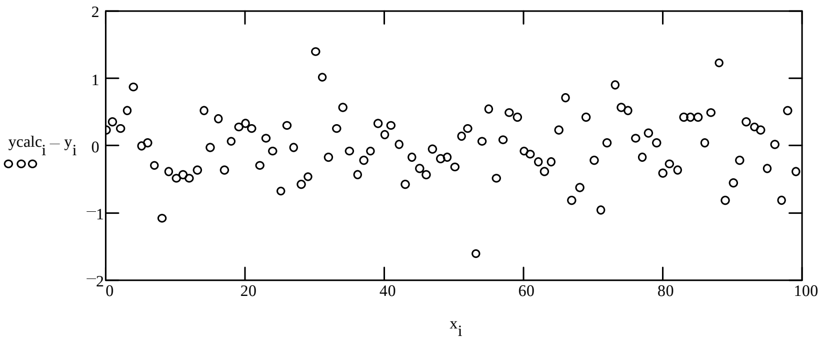
Goodness of fit parameter, Q $Q = 0.571$

Estimated σ of an observation of unit weight $S_1 = 0.98371$

Data and fit



Residuals



If you make a linear fit to transformed data, you must do a simple propagation of error calculation to find the appropriate uncertainties in the transformed data. Often, as far as you know, the uncertainties in the raw measurements are all the same. In that case, the uncertainties in the transformed data (the values you use in the fit) are usually not all the same.

For example, say you have a measured data set $\{x_i, d_i, \sigma_i\}$, where the σ_i give the (relative or absolute) uncertainties in the d_i. You expect the data to fit the exponential model

$$d = ae^{-kx}, \tag{86}$$

and you would like to estimate k and a by doing a linear least squares fit of $\ln(d_i)$ vs. x_i.

You must evaluate the uncertainty in $y_i = \ln(d_i)$:

$$\frac{\partial y}{\partial d} = \frac{1}{d} \tag{87}$$

$$\sigma_y^2 = \frac{1}{d^2}\sigma_d^2. \tag{88}$$

The correct uncertainties in a least-squares fit of $\ln(d)$ vs. x are therefore σ_i/d_i, rather than σ_i. If you think about it, that makes sense: the logarithm function is very sensitive (steep) to errors in d if d is small, but much less steep at large d.

Figure 5 shows a simulated set of data created from Eq. (84), with $a = 2.3$ and $k = 2.5$, and normally distributed noise with standard deviation 0.1 added to the y values. Also shown are the results of three least-squares fits: one nonlinear fit with uniform weighting, one linear fit to Eq. (85) with uniform weighting, and one linear fit with correct weighting ($w_i = y_i^2/\sigma_i^2$).

Note that the unweighted linear fit diverges from the other two at the early part of the decay. That's a crucial error, since those early points carry the most information about the k parameter (usually the one you're most interested in.) Since the weighted fit pays most attention to those, it does a much better job.

The weighted linear and nonlinear fits, while close together, are not exactly the same. There are two reasons. First, both methods are subject to numerical errors (accumulated roundoff errors, incomplete convergence, etc.) Second, and more important, the parent error distribution of the transformed data is not exactly normal, even though the parent distribution of the raw data was normal. The technique of χ^2 fitting is therefore not strictly applicable to the data once it has been transformed. Usually this subtlety is

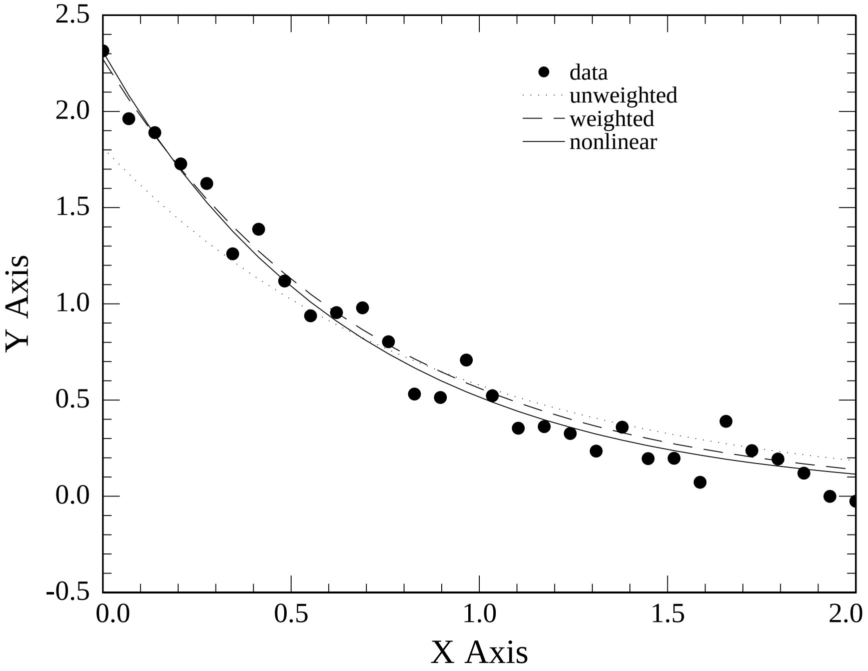


Figure 5: Fits of exponential function to noisy data.

not particularly important, since the transformed parent distribution is still close to Gaussian. It does, however, produce small errors in the parameters.

9.2.2 Nonlinear fitting

If there is no simple transformation of the model which produces a linear fitting problem (or even if there is but you want to avoid the problems already mentioned), then you must find a way to do the χ^2 minimization without the benefit of linear least squares. Generally you must resort to iterative minimization techniques. You imagine a "terrain" where the distance north-south and east-west correspond to the values of your parameters (the image only works for two parameters, but the math works for any number). The altitude corresponds to the value of χ^2. Your task is to find the lowest point in this terrain. To make the metaphor more realistic, you cannot “survey” the terrain by looking all around; you must be satisfied with evaluating the altitude at specific points. Several methods are available to you. They are outlined in Chapter 8 of B&R.

Grid search This is the most obvious (and usually one of the slowest) techniques. You hold all but one of your parameters fixed, and you vary that one, looking for the value which produces the minimum value of χ^2. You then fix that one at the optimum value you just found, choose another parameter, and vary it. When you have worked your way through all the parameters once, you usually find that the value of the first one is no longer optimum, so you start over. This method corresponds to limiting yourself to moving only north-south or east-west along your unknown terrain; you never permit yourself to move northeast, say. When you find that all your parameters seem to be optimum, you quit.

Gradient search, or steepest descent search In this technique you start at some position and evaluate the *gradient* (the vector of partial derivatives of χ^2 with respect to the parameters). The gradient shows you the direction which is steepest downhill. You move in that direction until you find yourself moving uphill again. You then reevaluate the gradient, and repeat. When you can find no downhill directions (the gradient is zero), you stop. This method is almost always faster than the grid search, but it can still be pretty slow, especially near the true minimum. It has the advantage that it always finds *some* minimum (assuming the surface has one.)

Expansion methods Here you assume that you are already pretty close to the minimum, and therefore a low-order Taylor expansion of the χ^2 function with respect to the parameters is likely to be pretty good. By evaluating several derivatives at your current position, you can obtain enough information to model the function as a paraboloid or ellipsoid. You then move to the position you calculate to be the minimum of that surface, and start over. Implementation of this technique is easy; it’s almost identical to a linear least squares calculation that just gets repeated over and over. (See SGN.) When you are in fact close to the minimum, this works great and is very fast. When you are far away, though, it is unstable and can carry you on awful wild goose chases.

Marquardt method The Marquardt method is a way of going smoothly from the gradient search method to the expansion method as you get closer to the minimum. It works very well, and is the industry standard. It's a little complicated to program (though not actually difficult), but there are lots of available programs which already include it. The shareware programs NONLIN (for DOS) and Datafit (for Windows) and the public domain programs Fudgit, Gnufit, and GLE (all for DOS, OS/2, and UNIX) all provide implementations. Most commercial data-treatment programs such as Igor, SigmaPlot, and PsiPlot also use the Marquardt method. Mathcad's function `minerr` uses Marquardt, but unfortunately provides absolutely no error information so it's largely useless by itself. (You can use `minerr` to find parameter values, and then construct and invert an A matrix to get uncertainties, though.) For those who can compile their own programs, the disk which comes with B&R has a Pascal program which does nonlinear fits, as does the disk which can be purchased to accompany *Numerical Recipes* in FORTRAN or C.

NONLIN is easy to use and versatile, and (unlike many other programs) will provide the variance-covariance matrix from the fit when it is needed. Its main weakness is that it does not allow you to specify the errors for the y_i; that is sometimes a problem, though not as bad a one for nonlinear models as for linear ones.

Worksheet 9 shows a NONLIN input file for fitting data to the function $y = axe^{-kx}$, Worksheet 10 shows part of the resulting output (.LST) file, and Figure 6 shows the data, the curve corresponding to the intial guesses for the parameters, and the final result of the fit.

Note the line "Stopped due to: Both parameter and relative function convergence" in the NONLIN output file. It's important to look for that; if it insteads says "Stopped due to: iteration limit reached", or something similar, then the algorithm has not converged and the output values are probably meaningless. You should choose better starting parameter values and start over, or you should set the ITERATIONS command in your NONLIN input file to a number higher than 50 to see whether more iterations will eventually help.

9.3 Recommendations

A Marquardt fit is almost always fastest and is easy to perform. Many programs will calculate the required derivatives $\partial\chi^2/\partial a_j$ numerically for you, leaving you to specify only the model function, the list of parameters to vary, and the data. Good implementations will let you specify individual uncertainties for the data points and will provide the variance-covariance matrix if you wish. If you have no Marquardt program available, grid-

Worksheet 9 Typical NONLIN input file.

```
title Nonlin example
register ! you should register NONLIN if you like it
variable X ! independent and dependent variables (observations)
variable Y
parameter a=5     ! parameters are what gets adjusted
parameter k=0.1 ! it’s best to give initial guesses
covariance ! ask for covariance matrix
function Y = a*x*exp(-k*X)
plot !ask NONLIN to produce a picture on the screen
data ! in order listed in Variables
             0    -0.02802838
              2       6.107835
              4       8.233952
              6       8.526069
              8       7.438673
             10       6.297892
             12       5.045212
             14       3.989115
             16       3.077397
             18       2.355075
             20       1.582216
```

search (easy but slow) and gradient-search methods work pretty well. The expansion method is easy to implement in Mathcad and works well if you are quite sure that you are starting with very good guesses for the parameter values (that is, you are already quite close to the minimum).

9.4 After the fit

Okay, you have now performed a Marquardt fit and your program has kicked out “optimum” values of the parameters and their standard deviations. What do you do now?

Don’t believe the program’s answer yet. It may have converged to a false minimum; your model may not actually fit the data; the stated standard deviations may be garbage because of correlations between different parameters.

First, make a plot of the *residuals*, the differences between the data points

Worksheet 10 Part of resulting NONLIN output file.

```
   ----  Final Results  ----

Nonlin version 2.5
Copyright (c) 1992-1993 (shareware) Phillip H. Sherrod.

Nonlin example
Function: Y = a*x*exp(-k*X)
Number of observations = 11
Maximum allowed number of iterations = 50
Convergence tolerance factor = 1.000000E-010
Stopped due to: Both parameter and relative function convergence.
Number of iterations performed = 8
Final sum of squared deviations = 6.23905E-002
Standard error of estimate = 0.0832603
Average deviation = 0.0568523
Maximum deviation for any observation = 0.16646
Proportion of variance explained (R^2) = 0.9992  (99.92%)
Adjusted coefficient of multiple determination (Ra^2) = 0.9991  (99.91%)
Durbin-Watson test for autocorrelation = 2.273



                   ----  Calculated Parameter Values  ----

 Parameter  Initial guess   Final estimate   Standard error      t      Prob(t)
----------  -------------  ----------------  --------------  ---------  -------
         a              5         4.59155865      0.04403474     104.27  0.00001
         k            0.1        0.198761219     0.001140971     174.20  0.00001


           ----  Variance-Covariance Matrix  ----

 Parameter              a              k
----------   -------------  -------------
         a:      0.0019391    4.4165E-005
         k:    4.4165E-005    1.3018E-006
```

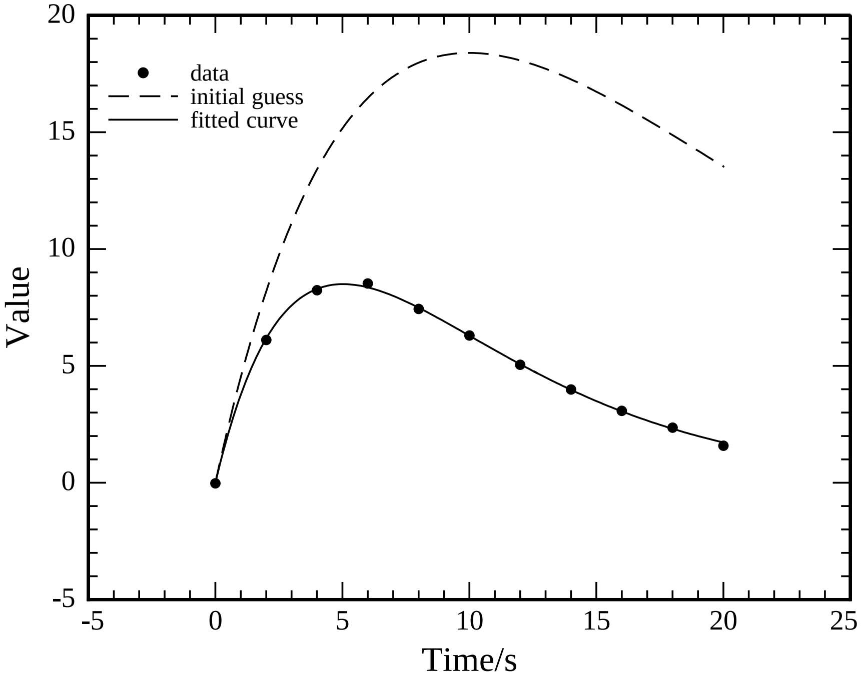


Figure 6: Data for Worksheet 9, with initial guess and final fit curves.

and your model: $r_i = y_i - f(\mathbf{x}_i, a_1, a_2, \cdots)$. The residuals should bounce up and down randomly about zero. If they have a big hump in them somewhere, then either your model does not describe the data well or you have not found the true minimum in χ^2.

Second, if the residuals look okay, do a goodness-of-fit test as described in Section 9.1.3. (You must have used true standard deviations in your weighting to have a useful absolute magnitude of χ^2). You should come up with a reasonably large value of Q, certainly at least 0.001.

Finally, you need to evaluate the uncertainties in the values of your fitted parameters. (If the first two tests fail, the uncertainties (and probably the fitted parameters too) will be meaningless.)

9.4.1 Uncertainties in fitted parameters

Once again, you have several options. Good fitting programs will provide you with estimated standard deviations for individual parameters, and should also provide you with covariances giving their correlations.

1. You can use the standard deviations for the parameters which your nonlinear fitting program provided. Those will be meaningful if the errors in your experiment were truly normal, your model function is an accurate description of the data, and the goodness of fit is okay. Then the calculated standard deviations are good for *individual* parameters, that is, are useful if you care about only one of the parameter values. To get confidence intervals on any one of the values, you can use the t-table as we did with simple averages.

2. You can use a projection method, described in SGN 6th ed. on page 726. That will give you more accurate confidence limits when the parameters are strongly correlated or the errors are slightly nonnormal.

3. You can do a Monte Carlo simulation, generating many synthetic data sets and subjecting each of them to your nonlinear fit. The resulting lists of parameter values can then be treated separately (as discussed before in Section 7.2) to obtain individual confidence intervals, or plotted together to get joint intervals. The latter treatment allows you to say "I am 95% confident that a lies between 1.1 and 1.5 units *and* that b lies between 3.2 and 4.1 units."

4. You can find the boundaries of the *error ellipsoid* for your data. This, too, will permit you to make statements about more than one of the values. See *Numerical Recipes* section 15.6 (second edition in C) for details.

10 Interpolation

Interpolation problems appear when you know the value of some function $f(x)$ only at some discrete values of x $(x_1, x_2, \cdots, x_n)$, and you want to know the value at some x not equal to one of the x_i. If $x_1 < x < x_n$ you have an interpolation problem; otherwise it's an extrapolation problem. You might have this problem because

- You measured some property only at discrete values of x;

- You have a table of values in a book, but no analytic formula for $f(x)$;
- You calculated values of f at several x_i, but each calculation is difficult or expensive and you don't want to do it any more.

The interpolation techniques I will describe in this section assume that you do not have any real, physical idea about the true form of $f(x)$, and therefore assume that it's just as well to use any convenient mathematical form. That convenient form is usually (though not always) a polynomial. Since many real functions are not well approximated by polynomials over finite distances, there's some danger in "blind" polynomial approximations. Therefore I must offer two important caveats:

1. If you have a physical model that gives an expected functional form for $f(x)$, it is always better to fit your data (experimental, looked-up, or calculated) to that form with least squares, and then calculate the interpolated values from your best-fit model parameters.
2. In the absence of such a physical model, *extrapolation* to x values more than about one x-spacing away from the available data is likely to give complete nonsense.

I will discuss two kinds of interpolation procedures: polynomial and spline. Polynomial interpolation is easiest to do with pencil and calculator. Spline interpolation is easy to do with a computer and is not as suceptible to noise in the data, but is not as flexible. Both assume that your function is tabulated at evenly spaced x_i. Neither works well when the underlying (unknown) function is not well approximated by polynomials.

10.1 Polynomial interpolation with difference tables

The idea behind polynomial interpolation is to find the polynomial of degree $n-1$ that goes exactly through n tabulated points near x, and then evaluate the resulting polynomial at x. In general, though, it's best not to find the coefficients of the interpolating polynomial and then evaluate $f(x)$ from them. Rather, one chooses one of the x_i which is close to the desired x, then "corrects" the value $f(x_i)$ by incorporating information from the function values at other nearby x_j.

You have a set of $\{x_i, y_i\}$ data points where i runs from 1 to N. I will "renumber" the points so that x_0 is one of the tabulated x_i close to your desired x; the y_i corresponding to x_0 becomes y_0, the preceding x_i will become x_{-1}, etc.

Now construct a *central difference table*, as shown in Table 6. The leftmost two columns are just your tabulated data points. Each δ is found by subtracting the value above and left of it from that below and left. For example,

$$\delta_{-\frac{1}{2}} = y_0 - y_{-1} \tag{89}$$

$$\delta_0^2 = \delta_{\frac{1}{2}} - \delta_{-\frac{1}{2}} \tag{90}$$

Table 6 Central difference table.

x_{-2}	y_{-2}			
		$\delta_{-1\frac{1}{2}}$		
x_{-1}	y_{-1}		δ^2_{-1}	
		$\delta_{-\frac{1}{2}}$		$\delta^3_{-\frac{1}{2}}$
x_0	y_0		δ^2_0	
		$\delta_{\frac{1}{2}}$		$\delta^3_{\frac{1}{2}}$
x_1	y_1		δ^2_1	
		$\delta_{1\frac{1}{2}}$		
x_2	y_2			

Table 6 shows only a small part of the difference table. If you need an interpolated value at only one x, you need only construct the table in that region. If you need interpolated values at many different x, it is very easy to generate the difference table in a spreadsheet or in Mathcad.

As an example, I'll construct a difference table from data in the CRC giving the velocity of sound in dry air at different temperatures, in meters per second. The data in the table are available at 10°C intervals. The difference table is shown in Table 7.

Note that the signs of the columns tend to alternate. In a precise tabulation such as this one, the differences remain pretty smooth even far out to the right (implying higher-order interpolating polynomials). In a tabulation of noisier data, the differences will rapidly stop being smooth. When you construct a difference table, you should stop going farther to the right as soon as the numbers in any column stop varying smoothly.

There are many different interpolation formulas which use the differ-

Table 7 Central difference table for velocity of sound

x	v	δ	δ^2	δ^3	δ^4
-30.00	312.72				
		6.37			
-20.00	319.09		-0.13		
		6.24		0.01	
-10.00	325.33		-0.12		0.00
		6.12		0.01	
0.00	331.45		-0.11		0.00
		6.01		0.01	
10.00	337.46		-0.10		-0.01
		5.91		0.00	
20.00	343.37		-0.10		0.00
		5.81		0.00	
30.00	349.18		-0.10		
		5.71			
40.00	354.89				

ences in Table 6. Two well-known ones are *Stirling's formula*,

$$f(x) \approx y_0 + \frac{1}{2}p(\delta_{\frac{1}{2}} + \delta_{-\frac{1}{2}}) + \frac{1}{2}p^2\delta_0^2 + \cdots, \tag{91}$$

and *Bessel's formula*,

$$f(x) \approx y_0 + p\delta_{\frac{1}{2}} + \frac{p(p-1)}{4}(\delta_0^2 + \delta_1^2) + \frac{p(p-\frac{1}{2})(p-1)}{6}\delta_{\frac{1}{2}}^3 + \cdots. \tag{92}$$

The variable p in those formulas is the fraction of an x-spacing your desired x is away from x_0:

$$p = \frac{x - x_0}{\Delta x}, \tag{93}$$

where Δx is the spacing between adjacent x_i. In Stirling's formula, it's best to choose x_0 to be the tabulated x_i closest to your desired x. Then $-\frac{1}{2} < p < \frac{1}{2}$. In Bessel's formula, you usually choose x_0 to be the largest tabulated x_i less than your desired x, so that $0 < p < 1$. In either formula you can use as many terms as you want; as you go out, the successive corrections should get smaller and smaller. You should stop adding terms when the next term is smaller than you care about, or when the differences in the next column do not vary smoothly. Note that interpolation cannot add precision to your original data: hoping to get more significant figures than the tabulated points have is not reasonable!

Either formula works well for quick work with a calculator and pencil. Note that if you just take the first two terms in Bessel's formula, you are doing an ordinary linear interpolation. In general Bessel's formula is slightly better if your desired x lies nearly halfway between two tabulated x_i, while Stirling's formula is better if x is very close to one of the x_i. In most cases there's almost no difference.

Say I need to know the velocity of sound at 17 °C. The closest temperature in the table is 20°C, so for Stirling's formula I choose that as x_0. Then $p = \frac{x-x_0}{\Delta x} = \frac{17-20}{10} = -0.3$. From Eq. (91), I find

$$\begin{aligned} v_{17} &= 343.37 + \frac{1}{2}(-0.3)(5.81 + 5.91) + \frac{1}{2}(-0.3)^2(-0.10) \qquad (94) \\ &= 343.37 - 1.758 + -0.005 \\ &= 341.61 \text{ m/s} \end{aligned}$$

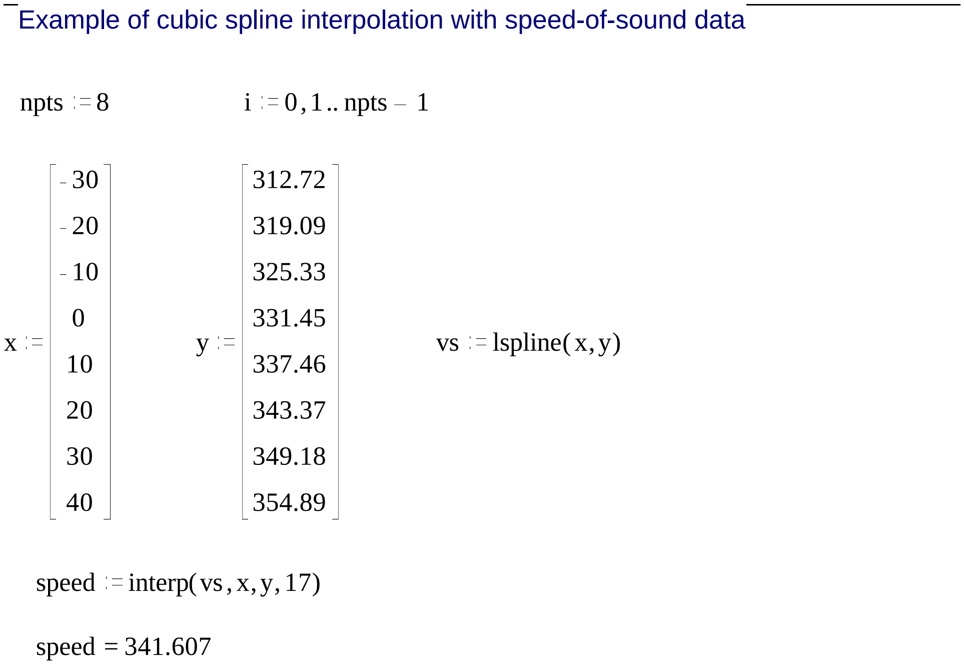
If I want to use Bessel's formula instead, I choose $x_0 = 10$, $p = \frac{x-x_0}{\Delta x} = \frac{17-10}{10} = 0.7$ and get

$$\begin{aligned} v_{17} &= 337.46 + (0.7)(5.91) + \frac{(0.7)(-0.3)}{4}(-0.10 + -0.10) \qquad (95) \\ &= 337.46 + 4.137 + 0.011 \\ &= 341.61 \text{ m/s} \end{aligned}$$

In fact, the table in the CRC is given at intervals of 1°C, so I can check whether these interpolations are accurate: the value for 17°C is exactly 341.61.

Higher terms in these two formulas, and many more formulas, can be found in the Chemical Rubber Company's *Standard Mathematical Tables*. The formulas there extend to quite high orders, meaning that high-degree polynomials are being fitted through more and more points of your data table. That is a little dangerous. Small bits of noise in the tabulated data can make a high-order polynomial take large "unphysical" swings, in order to go exactly through the data points. You can tell how high is safe by looking at the difference table itself: once the numbers in successive columns have stopped varying smoothly, noise has taken over and it is wrong to carry an interpolation formula out that far. If you need something more sophisticated than the quadratic or cubic formulas above, then have a look at the CRC math tables or at *Numerical Recipes*, which gives programs for polynomial interpolation of arbitrary order.

Worksheet 11 Cubic spline interpolation in Mathcad.



Example of cubic spline interpolation with speed-of-sound data

$\mathrm{npts} := 8 \qquad i := 0, 1 .. \mathrm{npts} - 1$

$$x := \begin{bmatrix} -30 \\ -20 \\ -10 \\ 0 \\ 10 \\ 20 \\ 30 \\ 40 \end{bmatrix} \qquad y := \begin{bmatrix} 312.72 \\ 319.09 \\ 325.33 \\ 331.45 \\ 337.46 \\ 343.37 \\ 349.18 \\ 354.89 \end{bmatrix} \qquad \mathrm{vs} := \mathrm{lspline}(x, y)$$

$\mathrm{speed} := \mathrm{interp}(\mathrm{vs}, x, y, 17)$

$\mathrm{speed} = 341.607$

10.2 Cubic spline interpolation

Spline interpolation is more complicated than simple polynomial interpolation, but it does not have so strong a tendency to oscillate wildly between data points and is very efficient for large data sets. Splines, unlike the interpolating polynomials used in the last section, have continuous derivatives up to some given order (continuous first and second derivatives for cubic splines.) The most popular splines are the cubic ones, and that is the version that Mathcad provides. There's no convenient way to go to higher order; you just have to take what you get. On the other hand, what you get is often pretty good.

To use the cubic spline routines in Mathcad, you first call a function `lspline` that sets up the *spline coefficients*; you do this just once for your entire data set. Then for each value of x where you want an interpolated y value, you call a second function `interp`. Worksheet 11 demonstrates the procedure for the same data set we used above.

My recommendations are

1. Use Bessel's or Stirling's formulas for simple linear or quadratic interpolations with a calculator.

2. Use cubic splines when you need lots of interpolated points. notes-8

3. Use routines from *Numerical Recipes* for higher order interpolation should you need them.

4. For *extrapolation*, stick with low-order ($n \leq 2$) fits, unless you have a good physical reason to believe that the underlying function really is a polynomial whose order you know.

I'll say it again: completely ignore this section, and interpolate with a least-squares fitted model function, if you have a sensible physical model. That's much better, especially for extrapolation.

11 Smoothing of data

The purpose of "smoothing" is to remove some random fluctuations in data to make later analyses more robust. It depends on the assumption that some underlying "real" quantity is varying slowly, so that any rapid fluctuations observed in the data are due to random error. Smoothing techniques try to remove the rapid fluctuations while preserving the slowly-varying signal. At least two procedures are better than smoothing:

1. Least-squares fitting of a model function to the raw data, followed by analysis using the fitted parameters;

2. Digital filtering, a technique which uses knowledge of the different characteristics of the "signal" and the "noise" to systematically remove the noise and retain the signal. See *Numerical Recipes* for leading references.

Most "smoothing" algorithms amount to naive applications of filtering, which make assumptions about the relative characteristics of signal and noise, and which typically let the user adjust the assumptions until he gets something he likes. This should strike you as somewhat ad-hoc and dicey, and subject to bias. It is; that's why the techniques listed above are preferable. Nonetheless, smoothing is sometimes useful as a preliminary to noise-sensitive procedures such as numerical differentiation or interpolation of a data set, or as a visual guide to help you pick the slow trend out of a graph of a noisy data set.

One of the best generic smoothing techniques is the class of moving-average filters know as *Savitsky-Golay* filters. These filters amount to doing least-squares fits of polynomials through each segment of the data set. A polynomial of degree n is fit through more than $n+1$ points, so it does not go through all the points but tends to reduce variations among them. Because of the special characteristics of evenly-spaced data, it is not necessary to perform an actual least-squares fit in each data segment. Instead, the "new" values of the data points are calculated by special weighted averages of the "old" values. The formula which comes from fitting a quadratic to 5 points is

$$y'_i = \frac{1}{35}[17y_i + 12(y_{i+1} + y_{i-1}) - 3(y_{i+2} + y_{i-2})]. \tag{96}$$

Other formulas of higher order are given in SGN and *Numerical Recipes*. However, you can get a long way just on this one, because the smooth is “stable”, that is, you can apply it more than once to get more and more smoothing. So if you have things set up so you can do a smooth, plot a graph, and do another smooth, you can keep applying this one formula until you are happy. Eventually, if you keep going, you’ll smooth the whole data set into an uninteresting blob; the technique really has no understanding of what is “signal” and what is “noise”.

Worksheet 12 shows a Mathcad implementation of this Savitsky-Golay smoothing function and its application to a noisy test data set. The model function is $y = \cos(x^2)$, corrupted with Gaussian noise of standard deviation 0.3. The upper plot shows the corrupted data as dots and the result of applying Eq. (96) once as a smooth line; the lower plot shows the underlying test function and the result of applying the smoothing function six times in succession. You can see that as you smooth more and more, the noise does decrease, but the quickly-changing parts of the true signal are lost.

Mathcad contains a built-in function to do smoothing called `medsmooth`. It replaces each data point with the median of itself and some surrounding points (you choose how many surrounding points). For data with normally distributed noise it is inferior to Savitsky-Golay, though it is quite reliable and can be used easily for quick projects. It only works once on a data set; successive applications don’t do anything. Mathcad Plus has a more sophisticated smoothing routine, `supsmooth`, similar to Savitsky-Golay.

Worksheet 12 Mathcad version of Savitsky-Golay smoothing.

Savitsky-Golay smoothing

$\text{npts} := 200 \qquad j := 0, 1 .. \text{npts} - 1 \qquad \text{noiselevel} := 0.3$

$\text{sginrange}(v, i) := (i > 1) \cdot (i < (\text{length}(v) - 2))$

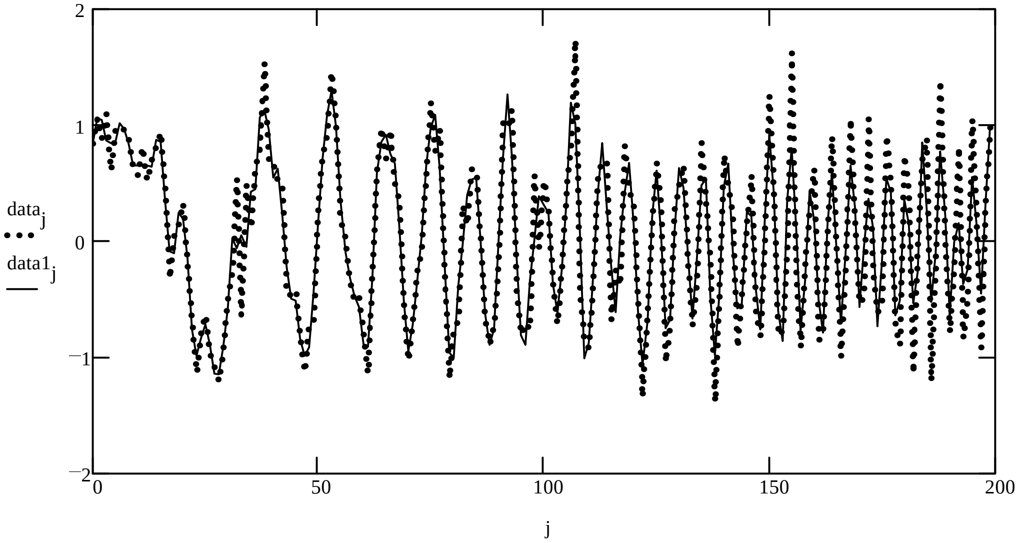
$\text{savgol}(v, i) := \frac{1}{35} \cdot \left[17 \cdot v_i + 12 \cdot \left(v_{i-1} + v_{i+1} \right) - 3 \cdot \left(v_{i-2} + v_{i+2} \right) \right]$

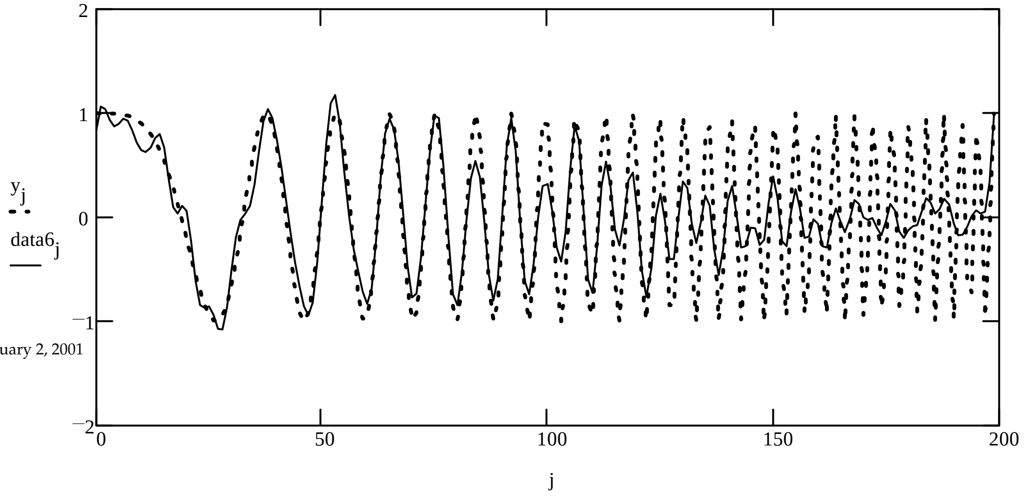
$\text{sgsmooth}(v, i) := \text{if}\left(\text{sginrange}(v, i), \text{savgol}(v, i), v_i \right)$

$y_j := \cos\left[\left(\frac{j}{15} \right)^2 \right] \qquad \text{data} := y + \text{noiselevel} \cdot \text{rnorm}(\text{npts}, 0, 1)$

$\text{data1}_j := \text{sgsmooth}(\text{data}, j) \qquad \text{data2}_j := \text{sgsmooth}(\text{data1}, j) \qquad \text{data3}_j := \text{sgsmooth}(\text{data2}, j)$

$\text{data4}_j := \text{sgsmooth}(\text{data3}, j) \qquad \text{data5}_j := \text{sgsmooth}(\text{data4}, j) \qquad \text{data6}_j := \text{sgsmooth}(\text{data5}, j)$





GCM January 2, 2001

A The laboratory notebook

Several sorts of errors can be more easily tracked and eliminated if careful and thorough records are kept. To quote from Wilson [8]: "An experimental scientist without his [or her] laboratory notebook is off duty."

You should purchase a bound, quadrille-ruled laboratory notebook with alternating white and yellow pages, identically numbered. At the end of each laboratory period you must get a teaching assistant to date and initial each notebook page you have written on. You should then hand in the yellow copies of the notebook pages.

All primary data should be recorded in ink in your lab notebook (*not* on paper towels, to be transferred to the notebook!) In addition, you should do preliminary analysis, such as graphing the points and eyeball-fitting a line, in the notebook *while you are doing the experiment*. This way you will be able to notice whether something is amiss right away, rather than after you have done the whole experiment or (worse) at midnight the night before your report is due. You'll also be much further along when you begin your detailed analysis.

When primary data are accumulated on a computer and stored in a file, it is important to be able to associate the notebook entry describing the experimental conditions with the data file. I recommend naming the file with a code giving the location (notebook, page, entry) of the notebook entry. The second entry on page 14 of my notebook (GM) would describe the data contained in computer file `GM014B.dat`, for example. If you are accumulating your data in computer files, it is very important to copy all the day's files to a separate disk at the end of each laboratory period, and to take the disk with you.

When you perform experiments with a partner, record primary data in both notebooks if you can. If that is impractical (if one student is holding a stopwatch and writing down readings while the other repeatedly reads a meter, for instance) then photocopies of the data pages from one notebook may be made immediately after the lab and pasted into the other. The students should each perform a complete analysis of the data (though they may discuss the proper way to do that between themselves), and the two reports must be prepared independently.

B Laboratory reports

Your reports will be read by the TAs and me, who generally know more about the experiments than you do. However, to write a good report, you must pretend that you are writing for someone else. The imaginary reader knows a fair amount of chemistry, but is unfamiliar with the particular experiments you do. He does have access to a good library and to the instructional handouts we use in the lab. He might be a professor at another institution or in another department at OSU, a group leader at Amoco, a colleague who needs your result for some of her own work, or whoever. The object of the report is to convey as concisely and clearly as possible

- what you were trying to measure,
- what experiments you performed,
- what results you saw,
- how you analyzed those results to reach a conclusion,
- how confident you are that other people measuring the same thing will get the same answer, and
- what (if any) things you think are interesting about your results.

Generally that list translates into a pretty standard technical report/article format.

In Chemistry 541 we will use two report formats. The informal one is designed to show that you did the experiment properly, analyzed its results correctly, and understand the technique you used. In preparing these short reports you will use many of the mathematical and statistical techniques discussed in lecture and in these notes.

The short format we use in 541 is a condensed version of that used in engineering and science research journals, though it is less demanding in terms of background citations and requires less discussion since you will not be obtaining "new" results.

A short report has the following parts:

Title and Abstract The abstract is a very concise summary. It gives, in about three sentences, the objective of the experiment, the experimental technique used, and the results. Important numerical results should be given, with uncertainties and units.

Results The Results section contains the data you obtained and the analysis you used. Probably your data will be presented in tables or graphs (or both). You should include an analysis of random errors in this section, though long derivations should be placed in an appendix and cited here. When your imaginary reader finishes the Results section, he should know what data you obtained and how you analyzed it.

Discussion Here you may put pretty much whatever you think is important that you have not yet said. In particular, I recommend the following:

- Discussion of any sources of systematic error in the experiment, how they might be eliminated in a better experiment, and your judgement about their importance.
- Comments about the likely sources of random error, and possible ways of reducing it.
- Your best values for the sought-after quantities and their uncertainties.
- Comparisons of your result with values obtained by others.
- Any other observations you want to make about the experiment, the analysis, or the results.

Questions Answers to any questions asked in the laboratory handout.

Bibliography Citations of other sources, including unpublished ones such as fellow students who helped you. Use a consistent citation style such as that used in Shoemaker, Garland, and Nibler (SGN).

Appendix This contains long, mostly nontextual, material which would unnecessarily break up the main report, and which is not important to a reader who just wants to know what you did and what results you got. Examples are long tables of raw data and long mathematical derivations. If you presented tables of calculated quantities in the Results section, you should show a sample calculation here.

Your report is an essay; it should read smoothly, and have the same quality of exposition that you would use in (say) a literature class. The first person is appropriate when you are describing actions you took. Please do not fall into the old trap of removing all the subjects from your sentences.

The report for one of the experiments you do should be prepared in the style used in the Journal of Physical Chemistry, including an introduction, a clear experimental section, and more extensive discussion.

C Tables, graphs, and units

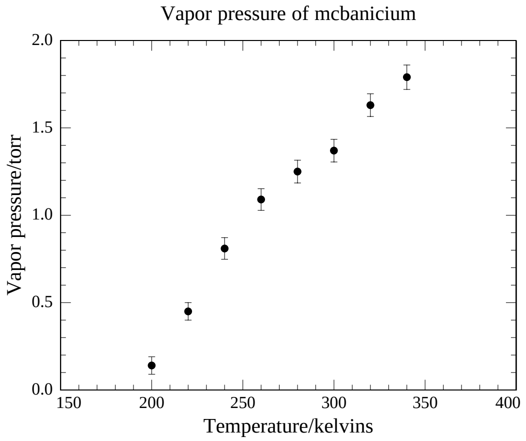
Tables and graphs should have identifying numbers and captions, and should give both values and units for all quantities. Two different ways of specifying units are acceptable.

1. Make the numbers or points in your table or graph dimensionless, and provide labels which show exactly how the dimensionless values were obtained. For example, a table column might be labeled T/K to indicate that the (dimensionless) numbers listed in the table were obtained by dividing the measured temperatures by 1 Kelvin. In a graph, the T/K should be the axis title, and should not be attached to a particular tick mark label on the axis. This is the style I prefer.

2. Tabulate or plot dimensioned quantities, and attach the appropriate units to the first item in the table column or to one of the tick-mark labels on the graph axis. The column heading or axis title should name only the quantity being plotted or tabulated and should have no reference to units.

Examples of these two styles are shown in Figure 7. Note that I do not recommend the common style of placing the axis units in parentheses after the axis label, as in "Concentration (M)". That notation becomes ambiguous when the units must be scaled. If a graph axis is labeled "Concentration (10^4 M)", and the numbers on the axis tick marks run from 0 to 4, it is not clear to the reader whether the true range runs from 0 to 4×10^4 M or from 0 to 4×10^{-4} M.

Temperature/K	Vapor pressure/Torr
200	0.14
220	0.45
240	0.81
260	1.09
280	1.25
300	1.37
320	1.63
340	1.79



Temperature	Vapor pressure
200 K	0.14 Torr
220	0.45
240	0.81
260	1.09
280	1.25
300	1.37
320	1.63
340	1.79

Figure 7: Examples of plot and table styles.

References

[1] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, New York, 2nd edition, 1992.

[2] Philip R. Bevington and D. Keith Robinson. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York, 2nd edition, 1992.

[3] David P. Shoemaker, Carl W. Garland, and Joseph W. Nibler. *Experiments in Physical Chemistry*. McGraw-Hill, New York, 6th edition, 1996.

[4] Hugh D. Young. *Statistical Treatment of Experimental Data*. Waveland Press, Prospect Heights, IL, 1992. (Originally published by McGraw-Hill.).

[5] William H. Beyer, editor. *Standard Mathematical Tables*. CRC Press, Boca Raton, FL, 27th edition, 1984.

[6] R. B. Dean and W. J. Dixon. *Analytical Chemistry*, 23:636, 1951.

[7] Dixon and Massey. *Introduction to Statistical Analysis*. McGraw-Hill, New York, 3rd edition, 1969.

[8] E. Bright Wilson, Jr. *An Introduction to Scientific Research*. McGraw-Hill, New York, 1952.

Colophon

These notes were typeset with LaTeX. The text is set in Palladio L by URW, with matching mathematical symbols from Diego Puga's Pazo Math set. Editing was done in GNU Emacs with the aucTeX enhancement package. The typesetting program was Christian Schenk's MikTeX. Most figures were generated with the Genplot program from Computer Graphic Service. Genplot and Mathcad figures and worksheets were saved as Encapsulated PostScript files and inserted into the final document by Thomas Rokicki's `dvips` program.