What does the result of an experiment tell you?

In the famous oil-drop experiment Millikan measured e to be, converting to SI units, 1.592×10^{-19} C. What does this tell you:

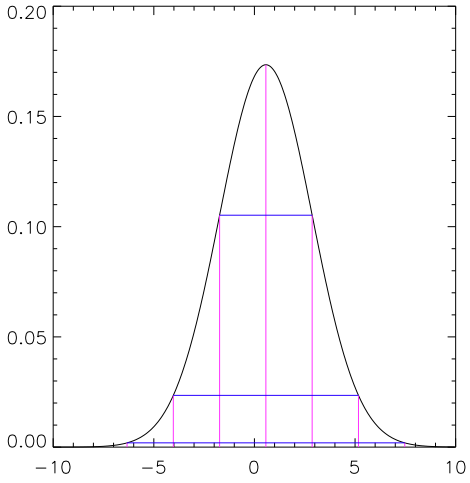
- $e = 1.59200000000 \times 10^{-19}$ C? Obviously not!
- 1.591 C$< 10^{19}e <$ 1.593 C? Well maybe. There's obviously some range of possible values, but can we assume that it's given by the number of digits? In general NO!! We distinguish between
  - the *precision* of a result: the least distinguishable change, given by the number of digits;
  - and the *accuracy*: the difference between the result and the true value.

In modern notation, Millikan gave the result as $e = 1.592(2) \times 10^{-19}$ C.
The figure in brackets is to be interpreted as a range on the last digit, so this implies a range of values between 1.590×10^{-19} C and 1.594×10^{-19} C.

A RESULT IS MEANINGLESS WITHOUT AN ERROR ESTIMATE.

So does Millikan's result tell us that e lies within that range? Still NO!

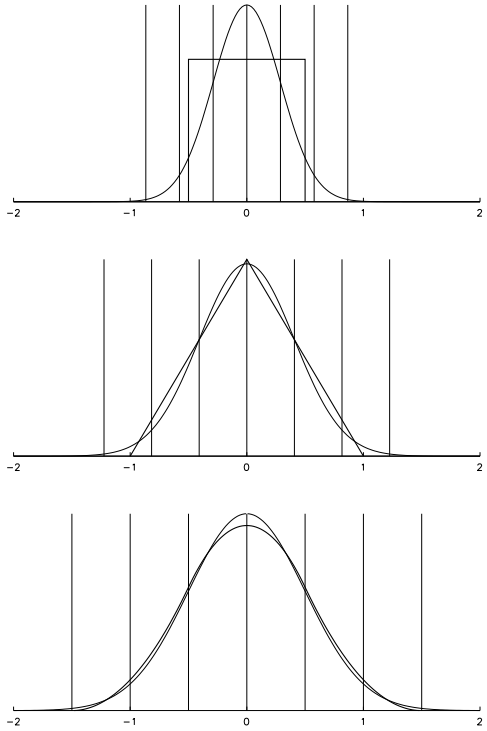
The result implies that e *probably* lies within that range, but there is a definite but non-zero probability that it lies outside that range. In order to attach numbers to these probabilities we need to know more than Millikan told us about the meaning of the error estimate.



The most complete information you could give is a *Probability Distribution Function (PDF)* of the result of an experiment, giving the probability that the result lies in any range. For the three marked ranges(±1, ±2, ±3 standard deviations) about the mean, the probability is the area between the ordinates (68%,95%,99.7% for this *normal distribution.*)

But however you interpret the error the result is INCONSISTENT with the accepted value of e of 1.602×10^{-19} C. But that's another story ...

Central Limit Theorem



The normal distribution is an important case because of the *central limit theorem:* the PDF of the result of an experiment approaches a normal distribution as the number of different error contributions increases.

This is illustrated here for a uniform distribution, appropriate to a single digitisation error. The three plots show how the PDF of the combination of one, two or three uniform distributions rapidly approaches the normal distribution.

How important is the error?

Over the century since Millikan's experiment, the importance of data anysis and the presentation of errors has grown steadily. The error is now *at least* as important as the result, and in some types of experiment much more so: null experiments, secular variation.

In 1571 the Dominican friar Egnatio Danti collected the following measurements of the Obliquity of the Ecliptic, ε:

Observer	Date	Value of ε
Ptolemy	*circa* 150	$23°51'20''$
Albategni	880	$23°35'00''$
Arabel	1070	$23°34'00''$
Almeone	1140	$23°33'00''$
Danti	1570	$23°29'00''$

(Quoted in The Sun in the Church, *J L Heilbron, Harvard, 1999, p.135)*

These raise the reasonable conjecture that ε is decreasing. But all these values depended on naked-eye astronomy (remember Galileo's telescope was 1609), for which a minute of arc is a pretty small angle. And none of the authors prior to Danti had quoted errors.

Suppose we have a set of measurements x_i of some well-defined quantity.
We define the true value of the quantity to be X. This is, of course, *and remains* unknown.
But we can write

$$x_i = X + \epsilon_i$$

where ϵ_i is the ERROR in the i'th measurement.

Why do measurements have errors?

There is always some fundamental *stochastic* or *random* process that limits a measurement.
This could be due to (for example):

Thermal fluctuations

Fundamental quantum-mechanical uncertainty

Seismic noise

Atmospheric turbulence.

But because these processes are random we can get a more accurate answer by averaging: $\langle \epsilon_i \rangle = 0$ and $\langle \epsilon_i \epsilon_j \rangle = 0$ — in words, the errors average to zero and are uncorrelated. (Hence if the repeated measurements are all *identical*, the measurement is non-ideal because information is being thrown away.)

But not all errors are so well-behaved. The problem with Millikan's result for the oil-drop experiment was that the formula for e included the viscosity of air, and he used the wrong value.
So although the measurements he took were correct, every result for e was wrong.

If $\langle \epsilon_i \rangle \neq 0$ then it is called the *bias* or the *systematic error* in the result.

Systematic errors are many and various:

Parallax
Calibration of meters
Zero offset
Backlash
Temperature drift
...

The systematic errors ultimately limit the accuracy we can obtain by averaging, so we normally keep a separate account of them:

$$\boxed{x_i = X + \epsilon_{\text{sys}} + \epsilon_i \quad \text{or} \quad x_i = x + \epsilon_i \quad \text{where} \quad x = X + \epsilon_{\text{sys}}.}$$

We now consider a set of measurements x_i of the quantity X made with the same equipment.

The Basic Questions:

(1) How accurate are the measurements?

(2) What's the best result?

(3) How accurate is that?

A useful fiction is to consider a much larger population of possible measurements we could have taken of which the N actual ones are a sample. (Obvious where that idea came from!). Then x is the mean of that larger population. x, like X, is and remains, unknown.

We can then assume that the ϵ_i are random:

$$\langle \epsilon_i \rangle = 0 \qquad \text{and} \qquad \langle \epsilon_i \epsilon_j \rangle = 0 \quad \text{for} \quad i \neq j$$

where the angle brackets denote the population average.
In the case $i = j$ we have

$$\langle \epsilon_i^2 \rangle = V_x \quad \text{the } \textit{Population Variance;} \text{ (the usual notation is var}(x))$$

and $\sqrt{V_x} = \sigma_x$, the Population Standard Deviation, the basic measure of the width of the distribution of measurements (some more, some less).

The Answers:

So now we are dealing with N repeated measurements x_i subject to random errors obtained by an identical process (so there is *no objective reason* to prefer one result to any other).

We return to the questions posed above:

(1) How accurate are the measurements?
Our best guess at the population variance/standard deviation is

$$V_x = \frac{\sum_i (x_i - \bar{x})^2}{N-1} \qquad \sigma_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$$

The puzzle is why $N-1$? This is because we only have $N-1$ measures of the spread of the distribution.

(2) What's the best result?
Our best guess at the population mean is $\bar{x} = \dfrac{\sum_i x_i}{N}$

(3) How accurate is that?
Our best guess at the variance/standard deviation of the mean is the *standard error*:

$$V_{\bar{x}} = \frac{V_x}{N} = \frac{\sum_i (x_i - \bar{x})^2}{N(N-1)} \qquad \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N(N-1)}}.$$

Millikan's Second Method for h/e

After completing the oil-drop experiment Robert Millikan turned to testing Einstein's photoelectric equation:

$$eV = h\nu - \Phi = h\nu - eV_0 \quad \rightarrow V = \frac{h}{e}\nu - V_0.$$

for the stopping potential V of photoelectrons liberated by light frequency ν. One way he measured h/e was to take a pair of frequencies A and B. With a sandwich of data ABA he could extract the slope as $\Delta V/\Delta\nu$. He repeated this nine times with the same A frequency and different B frequencies:

Slope: 4.11, 4.14, 4.10, 4.12, 4.24, 3.98, 4.04, 4.24, 4.21, $\times 10^{-15}$ V/Hz.

These give:

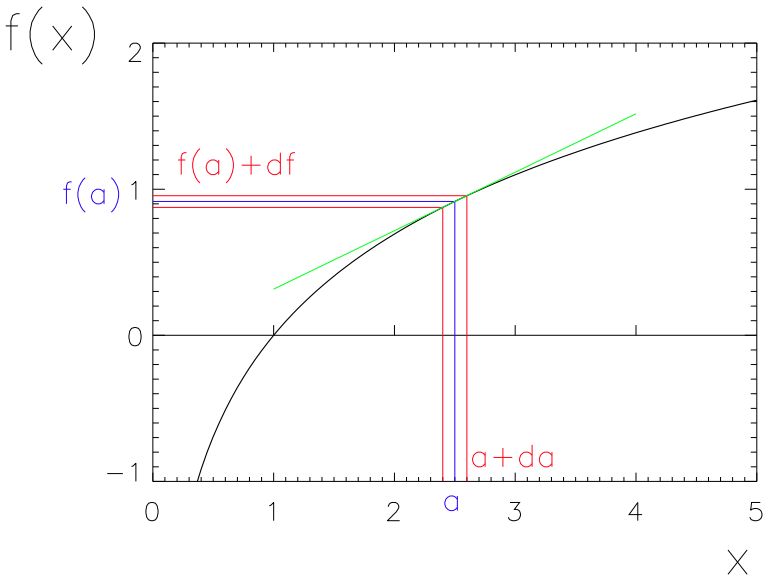
Mean 4.131 $\times 10^{-15}$ V/Hz

Population Standard Deviation 0.089 $\times 10^{-15}$ V/Hz

Standard Error 0.030 $\times 10^{-15}$ V/Hz

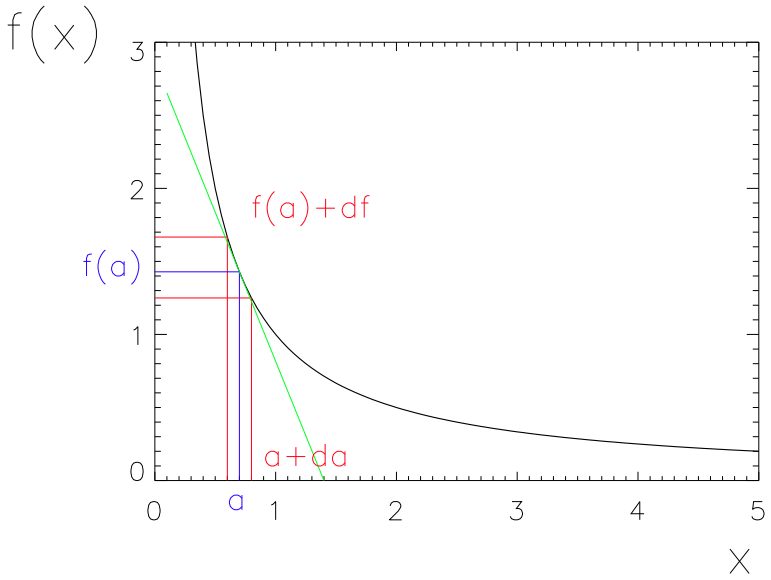
Thus the final result is 4.13(3) $\times 10^{-15}$ V/Hz. The fractional error is the error divided by the result, which is 0.007, or 0.7%. (The currently accepted value is $h/e = 4.135\,667\,33(10) \times 10^{-15}$ V/Hz, a fractional error of 2.5×10^{-8}.)

We have an experimental value a, with an error (standard deviation) σ_a.
Suppose we want to work out the error in a function $f(a)$.
This is called propagating the error.



Provided we can approximate the function over the small range by a straight line then it appears from the diagram that $\sigma_f = \left(\frac{df}{dx}\Big|_a\right)\sigma_a$.

However the slope might be negative:



So the correct formula must be

$$\boxed{\sigma_f = \left| \left(\left. \frac{df}{dx} \right|_a \right) \right| \sigma_a .}$$

Simple examples of Error Propagation

$f(a)$	σ_f
ka	$\lvert k \rvert \sigma_a$
k/a	$\dfrac{\lvert k \rvert \sigma_a}{a^2}$
$\ln(ka)$	$\dfrac{\sigma_a}{a}$
$\exp(ka)$	$\exp(ka)\,\lvert k \rvert \sigma_a$

The interesting example here is the logarithm:

- The fractional error σ_a/a is also the error in the logarithm;
- Multiplying the argument of the log by the constant k does *not* affect the error.

Suppose we want to use, in a calculation, *two* experimental values a and b, each with errors σ_a, σ_b:

$$a = \mathsf{a} + \epsilon_a \quad \text{where} \quad \langle \epsilon_a^2 \rangle = V_a = (\sigma_a)^2$$

and similarly for b:

$$b = \mathsf{b} + \epsilon_b \quad \text{where} \quad \langle \epsilon_b^2 \rangle = V_b = (\sigma_b)^2$$

The simplest case is where we wish to calculate $\mathsf{c} = \mathsf{a} + \mathsf{b}$:

$$a + b = \mathsf{a} + \mathsf{b} + \epsilon_a + \epsilon_b = \mathsf{c} + \epsilon_a + \epsilon_b.$$

Thus in each individual case the error in using our measured values a and b to calculate c is simply the sum of the errors. But what is it on average — because the errors could add, if we're unlucky, or cancel?

The variance of the result is, as usual,

$$V_c = \langle \epsilon_c^2 \rangle = \langle (\epsilon_a + \epsilon_b)^2 \rangle = \langle \epsilon_a^2 \rangle + \langle \epsilon_b^2 \rangle + 2\langle \epsilon_a \epsilon_b \rangle = V_a + V_b + 2C_{ab}$$

The final term here is called the *covariance* of a and b, (usual notation $\mathrm{cov}(a, b)$).
It measures the extent to which the errors in the two variables are coupled.

Covariance and Independence

The simplest case is when the errors in a and b are *independent*, so that positive and negative errors each occur randomly. In these circumstances the covariance is zero: $C_{ab} = 0$.

The opposite case would be when the value of b is actually calculated from a, so that the error in b is *perfectly correlated* with the error in a:

$$\text{Perfect Correlation:} \quad \epsilon_b = \alpha \epsilon_a \quad \rightarrow \quad C_{ab} = \alpha V_a \quad \text{and} \quad V_b = \alpha^2 V_a \quad \rightarrow \quad C_{ab} = \pm\sqrt{V_a V_b}.$$

In fact there is a very general inequality (the Schwarz or Cauchy-Schwarz inequality) which states

$$\left(\langle \epsilon_a \epsilon_b \rangle\right)^2 \leq \langle \epsilon_a^2 \rangle \langle \epsilon_b^2 \rangle \qquad \text{where the equality occurs when } \epsilon_b = \alpha \epsilon_a$$

so that we can write

$$C_{ab} = r\sqrt{V_a V_b} = r\sigma_a \sigma_b \qquad \text{where} \quad -1 \leq r \leq 1.$$

The coefficient r is the correlation coefficient, with $r = 0$ representing independence, and $r = \pm 1$ representing perfect linear correlation.

For a general $c = f(a, b)$ we simply propagate the error due to a and b through the function f:

$$c = \mathsf{c} + \epsilon_c = f(\mathsf{a} + \epsilon_a, \mathsf{b} + \epsilon_b) = f(\mathsf{a}, \mathsf{b}) + \left(\frac{\partial f}{\partial a}\epsilon_a + \frac{\partial f}{\partial b}\epsilon_b\right)$$

to first order. Thus the error in the computed value of c is ϵ_c given by

$$\epsilon_c = \left(\frac{\partial f}{\partial a}\epsilon_a + \frac{\partial f}{\partial b}\epsilon_b\right) = \mathbf{g}^T\mathbf{e} \text{ (or } \mathbf{e}^T\mathbf{g}) \quad \text{where } \mathbf{g} = \begin{pmatrix} \frac{\partial f}{\partial a} \\ \frac{\partial f}{\partial b} \end{pmatrix},\ \mathbf{e} = \begin{pmatrix} \epsilon_a \\ \epsilon_b \end{pmatrix}$$

implying that the variance is

$$\mathsf{V}_c = \mathbf{g}^T\langle\mathbf{e}\mathbf{e}^T\rangle\mathbf{g} = \mathbf{g}^T \begin{pmatrix} \langle\epsilon_a\epsilon_a\rangle & \langle\epsilon_a\epsilon_b\rangle \\ \langle\epsilon_b\epsilon_a\rangle & \langle\epsilon_b\epsilon_b\rangle \end{pmatrix} \mathbf{g} = \begin{pmatrix} \frac{\partial f}{\partial a} & \frac{\partial f}{\partial b} \end{pmatrix} \begin{pmatrix} \mathsf{V}_a & \mathsf{C}_{ab} \\ \mathsf{C}_{ab} & \mathsf{V}_b \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial a} \\ \frac{\partial f}{\partial b} \end{pmatrix}.$$

Combining Independent Errors

We now concentrate on the important case where the errors in a and b are independent. In the case we looked at above $c = a + b$ we found $\mathsf{V}_c = \mathsf{V}_a + \mathsf{V}_b$ or:

$$\boxed{\sigma_c = \sqrt{\sigma_a^2 + \sigma_b^2}.}$$

This is called *adding the errors in quadrature.*

Simple Examples

General Result	$c = f(a, b)$	$\sigma_c^2 = \left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 \sigma_b^2$
Sum	$c = a + b$	$\sigma_c^2 = \sigma_a^2 + \sigma_b^2$
Difference	$c = a - b$	$\sigma_c^2 = \sigma_a^2 + \sigma_b^2$
Product	$c = ab$	$\left(\frac{\sigma_c}{c}\right)^2 = \left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2$
Quotient	$c = \frac{a}{b}$	$\left(\frac{\sigma_c}{c}\right)^2 = \left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2$

The examples I am going to discuss are where we have a dataset of values of two variables x and y, *and* a theoretical expectation of a functional relation between them:

$$y = f(x; a) \text{ for some more-or less } \textit{known} \text{ function } f \text{ with parameters } a.$$

Examples include $y = kx$ or $y = c + mx$, where the parameters are k, m, c. These are *linear models*, where the parameters simply multiply known functions of x. But there could be more complicated relationships like $y = e^{ax}$.

[This *excludes* many classes of data analysis which are important in other fields, such as epidemiology. For example they might have data on the incidence in a sample group of a certain disease, together with lifestyle or genetic data on the same group. The incidence rate depends on *many* variables, and the question of whether a particular data item is or is not relevant is one of the unknowns.]

Then we have to do the following:

(1) Find the parameter values a that give the best fit.

(2) Consider whether the data are consistent with this functional form: deviations 'look random', and consistent with our error estimates.

(3) Assign error estimates to the parameters.

Most students skip question (2)!!

(1) Finding the Best Fit

The first thing to do is to find the best fit between the function and the data, taking into account the errors in the data. *This requires you to understand the errors in the data!* In all cases known to me the pre-programmed fitting routines assume that all the error is in y and that the values of x are precise. This is entirely for convenience: it produces a well-posed mathematical problem with a unique solution: trivially easy for a linear model, and not too hard for many non-linear models. Assuming errors in both x and y just makes the problem a great deal harder.

The consequence is that you have to decide which variable has the larger errors in comparison with the span of the data. That must be taken as y.

For the same reason the routines also assume that all the data have independent errors.

Then the best fit criterion is usually the least-squares one:

$$R(a) = \sum_i \Big(y_i - f(x_i; a)\Big)^2 \quad \rightarrow \quad \text{Minimize } R \text{ with respect to } a.$$

Minimisation for the General Linear Model

The general linear model (with two parameters) takes the form $\mathbf{y} = a_1\mathbf{f}_1 + a_2\mathbf{f}_2$. We shall use as an example fitting a straight line $y = mx + c$ to the dataset:

$$\mathbf{x} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} 2.55 \\ 2.69 \\ 3.95 \\ 4.77 \\ 5.36 \end{pmatrix} \quad \text{so in this case} \quad a_1 = c,\ \mathbf{f}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } a_2 = m,\ f_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

The y-values are subject to a random error with standard deviation 0.2.
If we make a rectangular matrix out of the two column vectors we can write this as

$$\boxed{\mathbf{y}_{\text{fit}} = \mathbf{F}\mathbf{a}} \qquad \boxed{\mathbf{F} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \quad \mathbf{a} = \begin{pmatrix} c \\ m \end{pmatrix}}$$

The function to be minimised is thus

$$R(\mathbf{a}) = (\mathbf{y} - \mathbf{F}\mathbf{a})^T(\mathbf{y} - \mathbf{F}\mathbf{a}).$$

At the minimum we define $\mathbf{a} = \mathbf{a}_{\min}$, $R(\mathbf{a}_{\min}) = R_{\min}$, and the residuals are $\mathbf{r} = \mathbf{y} - \mathbf{F}\mathbf{a}_{\min}$.

At the minimum, any small change in parameters leaves R unchanged, to first order in the small change. So when $\mathbf{a} = \mathbf{a}_{\min} + \delta\mathbf{a}$

$$R(\mathbf{a}) = R_{\min} - \delta\mathbf{a}^T\mathbf{F}^T\mathbf{r} - \mathbf{r}^T\mathbf{F}\delta\mathbf{a} + \text{2nd-order term}$$

The two first-order terms are identical.
The minimization condition $\delta R = 0$ for any $\delta\mathbf{a}$ is equivalent to $\mathbf{F}^T\mathbf{r} = 0$: the residuals are orthogonal to all the $\mathbf{f}$.

$$\boxed{(\mathbf{F}^T\mathbf{F})\mathbf{a}_{\min} = \mathbf{F}^T\mathbf{y}.}$$

This is a system of linear equations for the parameters which we can solve as long as the matrix $\mathbf{F}^T\mathbf{F}$ is non-singular. (This can only happen if we have done something stupid like put the same functional form into two different $\mathbf{f}$'s. A much more common problem is that it can become nearly-singular, if we use $\mathbf{f}$'s that are too similar.) Thus the general linear model has a very simple solution involving just a matrix inverse:

$$\mathbf{a}_{\min} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y} \quad \text{which gives} \quad \mathbf{y}_{\text{fit}} = \mathbf{P}\mathbf{y} \quad \text{where} \quad \mathbf{P} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T.$$

The minimised residuals are then given by $\mathbf{r} = \mathbf{y} - \mathbf{y}_{\text{fit}} = \mathbf{Q}\mathbf{y}$ where $\mathbf{Q} = \mathbf{I}_N - \mathbf{P}$.

Applying this to our example:

$$(\mathbf{F}^T\mathbf{F}) = \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix} \qquad (\mathbf{F}^T\mathbf{F})^{-1} = \begin{pmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{pmatrix} \qquad \mathbf{F}^T\mathbf{y} = \begin{pmatrix} 19.32 \\ 46.34 \end{pmatrix}$$

These give

$$\mathbf{a}_{\min} = \begin{pmatrix} 2.324 \\ 0.770 \end{pmatrix} \quad \mathbf{P} = \begin{pmatrix} 0.6 & 0.4 & 0.2 & 0.0 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0.0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.0 & 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0.0 & 0.2 & 0.4 & 0.6 \end{pmatrix} \quad \mathbf{r} = \begin{pmatrix} +0.226 \\ -0.404 \\ +0.086 \\ +0.136 \\ -0.044 \end{pmatrix} \quad R_{\min} = 0.24212$$

Suppose we had a second dataset with the same x-values: all the $\mathbf{F}$-matrices are the same.

$$\mathbf{y} = \begin{pmatrix} 2.17 \\ 2.62 \\ 3.94 \\ 4.39 \\ 5.46 \end{pmatrix} \quad \mathbf{F}^T\mathbf{y} = \begin{pmatrix} 18.58 \\ 45.51 \end{pmatrix} \quad \mathbf{a}_{\min} = \begin{pmatrix} 2.046 \\ 0.835 \end{pmatrix} \quad \mathbf{r} = \begin{pmatrix} +0.124 \\ -0.261 \\ +0.224 \\ -0.161 \\ +0.074 \end{pmatrix} \quad R_{\min} = 0.16507$$

(2) Is the data consistent with the best fit?

In neither case do the residuals show any obvious systematic deviation, and they are comparable with the standard deviation of 0.2 — the largest is 2 standard deviations, which is not large enough cause surprise in a dataset of 10. So I would say yes. However the residuals are somewhat smaller in the second case: we shall look at the implications of this in the next lecture.

(3) Accuracy of the coefficients.

The y-data are subject to statistical error:

$$\mathbf{y} = \mathsf{y} + \mathbf{e} \quad \text{where} \quad \langle \mathbf{e}\mathbf{e}^T \rangle = V_y \mathbf{I}_N \quad (cf \text{ p. 15: the } \epsilon_i \text{ are independent with equal variance.})$$

Then the true mean of the coefficient vector is given by $\mathsf{a} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathsf{y}$, and the true error in $\mathbf{a}_{\min}$ is $\mathbf{a}_{\min} - \mathsf{a} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{e}$. The variance-covariance matrix of the coefficients is thus

$$\langle (\mathbf{a}_{\min} - \mathsf{a})(\mathbf{a}_{\min} - \mathsf{a})^T \rangle = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \langle \mathbf{e}\mathbf{e}^T \rangle \mathbf{F} (\mathbf{F}^T\mathbf{F})^{-1}.$$

When we substitue for $\langle \mathbf{e}\mathbf{e}^T \rangle$ this simplifies rather beautifully:

$$\boxed{\langle (\mathbf{a}_{\min} - \mathsf{a})(\mathbf{a}_{\min} - \mathsf{a})^T \rangle = V_y (\mathbf{F}^T\mathbf{F})^{-1}.}$$

For the current example this matrix is

$$V_y(\mathbf{F}^T\mathbf{F})^{-1} = \begin{pmatrix} 0.024 & -0.008 \\ -0.008 & 0.004 \end{pmatrix} \quad \text{giving} \quad \begin{pmatrix} c \\ m \end{pmatrix} = \begin{pmatrix} 2.324 \pm 0.155 \\ 0.770 \pm 0.063 \end{pmatrix} \quad \begin{pmatrix} 2.046 \pm 0.155 \\ 0.835 \pm 0.063 \end{pmatrix}$$

These are consistent, which is not surprising since the underlying y-datasets are consistent.

If we don't have a prior number for V_y, can we estimate it from the residuals?

Yes! We re-write $R_{\min}$ in the equivalent form $R_{\min} = \mathrm{Tr}(\mathbf{r}\mathbf{r}^T) = \mathrm{Tr}(\mathbf{Q}\mathbf{y}\mathbf{y}^T\mathbf{Q}^T)$:

$$\langle R_{\min} \rangle = \mathrm{Tr}\Big(\mathbf{Q}\langle \mathbf{y}\mathbf{y}^T \rangle \mathbf{Q}^T\Big).$$

The only random variable here is $\mathbf{y} = \mathsf{y} + \mathbf{e}$ and $\langle \mathbf{y}\mathbf{y}^T \rangle = \mathsf{y}\mathsf{y}^T + V_y\mathbf{I}_N$.
Thus $\langle R_{\min} \rangle$ splits into 2 terms, a random one and a systematic one:

$$\langle R_{\min} \rangle = V_y\,\mathrm{Tr}\Big(\mathbf{Q}\mathbf{Q}^T\Big) + \mathrm{Tr}\Big((\mathbf{Q}\mathsf{y})(\mathbf{Q}\mathsf{y})^T\Big)$$

The systematic term is just the sum of squares of the systematic deviations from the straight line: $\mathrm{Tr}\Big((\mathbf{Q}\mathsf{y})(\mathbf{Q}\mathsf{y})^T\Big) = (\mathbf{Q}\mathsf{y})^T(\mathbf{Q}\mathsf{y})$. If the y actually lie on a straight line then $\mathbf{Q}\mathsf{y} = 0$.

Thus *provided we know that y is truly given by the functions in* $\mathbf{F}$ we can indeed use the random deviations from the fit to find an estimator for V_y:

$$\boxed{\langle R_{\min} \rangle = V_y(N - M)}$$

[Details:
We note that $\mathbf{P}$, and hence $\mathbf{Q}$, are symmetric $\mathbf{Q} = \mathbf{Q}^T$.
We then note that $\mathbf{P}\mathbf{P} = \mathbf{P}$ (obvious when you consider it projects an arbitrary data vector onto one fitting the functions in $\mathbf{F}$), and $\mathbf{Q}\mathbf{Q} = \mathbf{Q}$.
Thus the random term in $R_{\min}$ is just $\mathrm{Tr}(\mathbf{I}_N - \mathbf{P}) = N - \mathrm{Tr}(\mathbf{P})$.
But $\mathrm{Tr}(\mathbf{P}) = \mathrm{Tr}\Big(\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\Big) = \mathrm{Tr}\Big((\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{F}\Big) = \mathrm{Tr}(\mathbf{I}_M)$. Thus $\mathrm{Tr}(\mathbf{Q}) = (N - M)$ where N is number of data values and M is the number of functions.]

Of course all the fitting routines simply assume this condition is satisfied and use $R_{\min}$ to assign errors to the coefficient without comment. Here as always it is up to the user to determine whether there are relevant systematic errors that could affect the result.

So in our example, if we did *not* have the y-standard deviation of 0.2 we could estimate it from the values of $R_{\min}$ for the two datasets:

	$R_{\min}$	V_y	σ_y	m	c
Set 1	0.24212	0.081	0.28	2.324(220)	0.770(90)
Set 2	0.16507	0.055	0.23	2.046(182)	0.835(74)

Taken at face value this implies that the parameter values from the second experiment are more accurate than the first *because* they fit the straight line better. So we *could* make a weighted mean of these values, giving more weight to the second one.

BUT: given the two y-values for each x we would find a 'best estimate' for it by taking a simple mean, giving them *equal* weight.

So we have two different approaches for combining the datasets: which is right?

Consider repeated runs of an experiment with *no changes* in the equipment or procedures, just random variation in the results. Is a result derived from a run with low scatter better or not? The assigned error suggests that it is, but how accurate is the assigned error?

We shall look at two important properties of the error in a one-parameter fit to N y-values (such as a proportionality $y = ax$).

(1) The variance of the y-variance estimator

(2) The covariance of the variance estimator with the actual squared error.

Notation

We have the data $\mathbf{y} = \mathsf{y} + \mathbf{e}$ and the single function vector $\mathbf{F} = \mathbf{f}$, from which we calculate the following:

- The variance estimator $V_y = \dfrac{R_{\min}}{(N-1)} = \dfrac{\mathbf{y}^T\mathbf{Q}^T\mathbf{Q}\mathbf{y}}{(N-1)}$ (*cf* pp 23-24).
- The parameter $a_{\min} = \dfrac{\mathbf{f}^T\mathbf{y}}{\mathbf{f}^T\mathbf{f}}$ (*cf* p. 20). Its true error is $a_{\min} - \mathsf{a} = \dfrac{\mathbf{f}^T\mathbf{e}}{\mathbf{f}^T\mathbf{f}}$.
- The estimator for its variance depends V_y: $V_a = \dfrac{V_y}{\mathbf{f}^T\mathbf{f}}$ (*cf* p. 22).

(1) The Error on the Error:

The error in our variance estimator is $V_y - \mathsf{V}_y$, and so its variance is

$$\mathsf{V}_V = \langle (V_y - \mathsf{V}_y)^2 \rangle = \left\langle \left(\frac{\mathbf{y}^T \mathbf{Q}^T \mathbf{Q} \mathbf{y}}{(N-1)} - \mathsf{V}_y \right)^2 \right\rangle = \left\langle \left(\frac{\mathbf{y}^T \mathbf{Q}^T \mathbf{Q} \mathbf{y}}{(N-1)} \right)^2 \right\rangle - \mathsf{V}_y{}^2$$

In this one-parameter case $\mathbf{Q} = \mathbf{I}_N - (\mathbf{f}^T\mathbf{f})^{-1}\mathbf{f}\mathbf{f}^T$ and hence $\mathbf{Q}^T = \mathbf{Q}$ and $\mathbf{Q}\mathbf{Q} = \mathbf{Q}$.
Setting $\mathbf{y} = \mathsf{y} + \mathbf{e}$, we also note that $\mathbf{Q}\mathsf{y} = 0$ so we can replace $\mathbf{y}$ with $\mathbf{e}$:

$$\mathsf{V}_V = \frac{\langle \mathbf{e}^T \mathbf{Q} \mathbf{e} \mathbf{e}^T \mathbf{Q} \mathbf{e} \rangle}{(N-1)^2} - \mathsf{V}_y{}^2$$

We can bring all the $\mathbf{e}$'s together if we go into index notation:

$$\mathsf{V}_V = \frac{\sum_{i,j,k,l} \langle \epsilon_i \epsilon_j \epsilon_k \epsilon_l \rangle \mathbf{Q}_{ij} \mathbf{Q}_{kl}}{(N-1)^2} - \mathsf{V}_y{}^2$$

Thus we need to consider how to average four errors.

Evaluating $\langle \epsilon_i \epsilon_j \epsilon_k \epsilon_l \rangle$:

We extend the concept of independent errors to propose $\langle f(\epsilon_i) g(\epsilon_j) \rangle = \langle f(\epsilon_i) \rangle \langle g(\epsilon_j) \rangle$.
We then consider the following cases:

1. $i,\ j,\ k,\ l$ all distinct: $\langle \epsilon_i \rangle \langle \epsilon_j \rangle \langle \epsilon_k \rangle \langle \epsilon_l \rangle = 0$.
2. $i,\ j,\ k$ distinct, $k = l$: $\langle \epsilon_i \rangle \langle \epsilon_j \rangle \langle \epsilon_k^2 \rangle = 0$.
3. $i,\ j$ distinct, $j = k = l$: $\langle \epsilon_i \rangle \langle \epsilon_j^3 \rangle = 0$.
4. $i,\ j$ distinct, $i = k$ and $j = l$: $\langle \epsilon_i^2 \rangle \langle \epsilon_j^2 \rangle = \mathsf{V}_y{}^2$.
5. $i = j = k = l$: $\langle \epsilon_i^4 \rangle$.

The fourth power depends on the fourth power of the width and on shape of the PDF:

$$\boxed{\langle \epsilon_i^4 \rangle = k \mathsf{V}_x{}^2}$$

where k depends on the Probability Distribution Function: for a normal distribution $k = 3$. However the fourth power makes it sensitive to the tail of the distribution: for a hard cut-off, with no tail, it is less (uniform distribution $k = 9/5$) and for a longer tail it is greater.

Cases 1.–4. are covered by $\langle \epsilon_i \epsilon_j \epsilon_j \epsilon_l \rangle = \mathsf{V}_y{}^2 \Big(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} \Big)$.
However for Case 5. this gives $\langle \epsilon_i^4 \rangle = 3\mathsf{V}_y{}^2$ which is wrong, so . . .

$$\boxed{\langle \epsilon_i \epsilon_j \epsilon_k \epsilon_l \rangle = V_y{}^2 \Big(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + (k-3)\Delta_{ijkl} \Big)}$$

where $\Delta_{ijkl} = 1$ if $i = j = k = l$ and 0 otherwise.

Putting this into the expression for V_V:

$$V_V = V_y{}^2 \Big[\frac{\sum_{i,j} \Big(\mathbf{Q}_{ii}\mathbf{Q}_{jj} + \mathbf{Q}_{ij}\mathbf{Q}_{ij} + \mathbf{Q}_{ij}\mathbf{Q}_{ji} \Big) + (k-3) \sum_i \mathbf{Q}_{ii}\mathbf{Q}_{ii}}{(N-1)^2} - 1 \Big]$$

The first term gives $\mathrm{Tr}(\mathbf{Q})^2$, where the trace is $N-1$. The second and third terms give $\mathrm{Tr}(\mathbf{Q}\mathbf{Q}^T) + \mathrm{Tr}(\mathbf{Q}^2)$, which are both equal to $N-1$. Thus in the important case of normal errors ($k = 3$):

$$\boxed{V_V = \frac{2V_y{}^2}{(N-1)} \quad \text{and hence } \sigma_V = V_y \sqrt{\frac{2}{N-1}}}$$

where σ_V is the error on V_y. The fractional error is thus $\sqrt{2/(N-1)}$ which propagates to :

$$\boxed{\text{The Fractional Error on } \sigma_y \text{ in a 1-parameter fit is } 1/\sqrt{2(N-1)}.}$$

This result generalises in the obvious way to to fitting M functions instead of 1.

In our earlier example $N = 5$ and $N - M = 3$ so the fractional error on V_y as found from $R_{\min}$ is $1/\sqrt{6} = 41\%$. So the difference between the σ_y from the two datasets (Table on p.25) is entirely to be expected.

However this leaves open the question: does the second dataset, which is more linear, and hence has a smaller σ_y derived from $R_{\min}$, also give a more accurate result for a?

(2) The Covariance of Linearity and Accuracy:

We consider the covariance of the deviation of the variance estimator from its mean with the squared error in a:

$$C = \left\langle \Big(V_y - V_y \Big) \Big(\frac{(\mathbf{f}^T\mathbf{e})^2}{(\mathbf{f}^T\mathbf{f})^2} \Big) \right\rangle = \left\langle \left(\frac{\mathbf{e}^T\mathbf{Q}^T\mathbf{Q}\mathbf{e}}{(N-1)} - V_y \right) \left(\frac{(\mathbf{f}^T\mathbf{e})^2}{(\mathbf{f}^T\mathbf{f})^2} \right) \right\rangle$$

Again we go into index notation:

$$C = \langle \epsilon_i \epsilon_j \epsilon_k \epsilon_l \rangle \frac{\mathbf{Q}_{ij} f_k f_l}{(N-1)(\mathbf{f}^T\mathbf{f})^2} - \frac{V_y{}^2}{\mathbf{f}^T\mathbf{f}}$$

Using the boxed equation for the quadruple product, the first term cancels the last, the next two terms vanish because $\mathbf{Q}\mathbf{f} = 0$ to leave

$$\boxed{C \propto (k-3).}$$

THUS FOR NORMAL ERRORS LINEARITY AND ACCURACY ARE UNCORRELATED!

Indeed for a uniform distribution they are *negatively* correlated: parameters obtained from data with a large scatter are (on average) more accurate than from data with a small scatter.

Closing Remarks:

- We have to assign errors to results.
- These come from the deviation of results from expected patterns:
  ·In a single value, from the fact that the results differ (*cf* pp 8–9).
  ·In a related value value, by propagation of errors (*cf* pp 10–16).
  ·In an (x, y) dataset, from the fact that they do not fit the expected functional form (*cf* pp 17–24).
- Unless you have an enormous dataset these error estimates are very uncertain (*cf* pp 25–29).
- You can't beat statistics: if you find a dataset with small scatter it isn't likely to be any more accurate, so you should always combine results with equal weight unless there is an objective reason to do otherwise (*cf* p. 30).