0: Systematic Error

> Physical scientists. . . know that measurements are never perfect and thus want to know *how* true a given measurement is. This is a good practice, for it keeps everyone honest and prevents research reports from degenerating into fish stories.
>
> Robert Laughlin (1998 Physics Nobel Laureate) p.10 *A Different Universe*

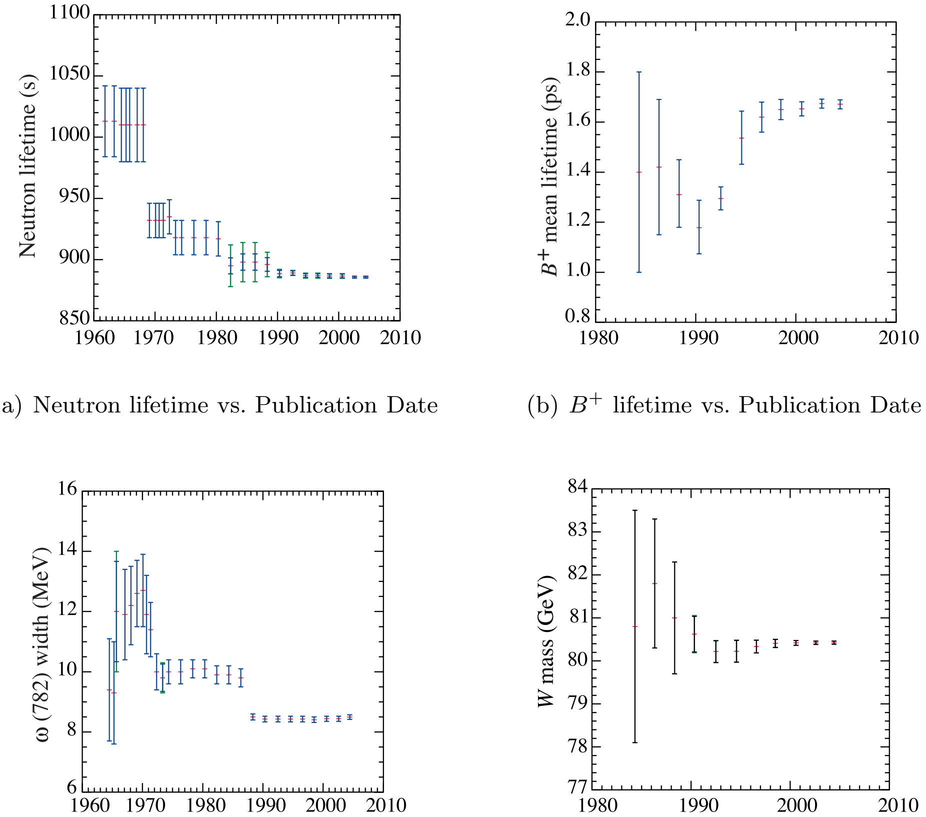
> A hypothesis or theory is clear, decisive, and positive, but it is believed by no one but the man who created it. Experimental findings, on the other hand, are messy, inexact things, which are believed by everyone except the man who did the work.
>
> Harlow Shapley, *Through Rugged Ways to the Stars.* 1969

Perhaps the dullest possible presentation of progress[1] in physics is displayed in Figure 1: the march of improved experimental precision with time. The expected behavior is displayed in Figure 1(d): improved apparatus and better statistics (more measurements to average) results in steady uncertainty reduction with apparent convergence to a value consistent with any earlier measurement. However frequently (Figs. 1(a)–1(c)) the behavior shows a 'final' value inconsistent with the early measurements. Setting aside the possibility of experimental blunders, *systematic error* is almost certainly behind this 'odd' behavior. Uncertainties that produce different results on repeated measurement (sometimes called *random errors*) are easy to detect (just repeat the measurement) and can perhaps be eliminated (the standard deviation of the mean $\propto 1/N^{1/2}$ which as $N \to \infty$, gets arbitrarily small). But systematic errors do not telegraph their existence by producing varying results. Without any telltale signs, systematic errors can go undetected, much to the future embarrassment of the experimenter. This semester you will be completing labs which display many of the problems of non-random errors.

Experiment: Measuring Resistance I

Consider the case of the digital multimeter (DMM). Typically repeated measurement with a DMM produces exactly the same value—its random error is quite small. Of course, the absence of random error does not imply a perfect measurement; Calibration errors are

[1] Great advancements is physics (Newton, Maxwell, Einstein) were not much influenced by the quest for more sigfigs. Nevertheless, the ability to precisely control experiments is a measure of science's reach and history clearly shows that discrepant experiments are a goad for improved theory.



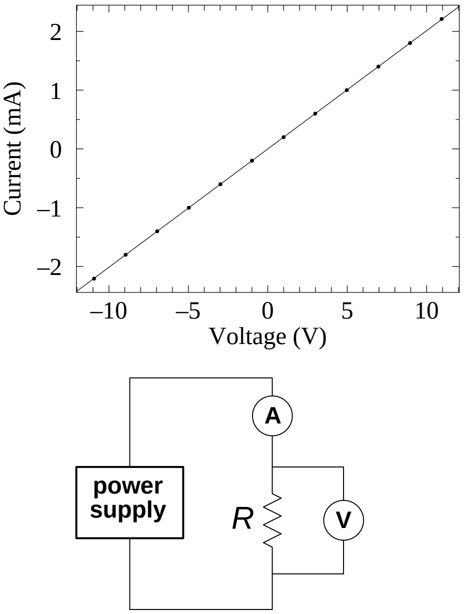
(a) Neutron lifetime vs. Publication Date

(b) B^+ lifetime vs. Publication Date

(c) ω width vs. Publication Date

(d) W mass vs. Publication Date

Figure 1: Measured values of particle properties 'improve' with time, but 'progress' is often irregular. The error bars (δx) are intended to be '$\pm 1\sigma$': the actual value should be in the range $x \pm \delta x$ 68.3% of the time (if the distribution were normal) and in the range $x \pm 2\delta x$ 95.4% of the time. These figures are from the Particle Data Group, `pdg.lbl.gov`.



V	δV	I	δI
(V)	(V)	(mA)	(mA)
0.9964	.005	0.2002	.001
2.984	.007	0.6005	.003
4.973	.009	1.0007	.005
6.963	.011	1.4010	.007
8.953	.013	1.8009	.009
10.942	.015	2.211	.012
−0.9962	.005	−0.1996	.001
−2.980	.007	−0.6000	.003
−4.969	.009	−1.0002	.005
−6.959	.011	−1.4004	.007
−8.948	.013	−1.8001	.009
−10.938	.015	−2.206	.012

Figure 2: A pair DM-441B DMMs were used to measure the voltage across (V) and the current through (I) a 4.99 kΩ resistor

expected and reported in the device's specifications. Using a pair of DM-441B multimeters, I measured the current through and the voltage across a resistor. (The circuit and results are displayed in Figure 2.) Fitting the expected linear relationship ($I = V/R$), Linfit reported $R = 4.9696 \pm .0016$ kΩ (i.e., a relative error of 0.03%) with a reduced χ^2 of .11. (A graphical display showing all the following resistance measurements appears in Figure 3. It looks quite similar to the results reported in Figs. 1.)

This result is wrong and/or misleading. The small reduced χ^2 correctly flags the fact that the observed deviation of the data from the fit is much less than what should have resulted from the supplied uncertainties in V and I (which were calculated from the manufacturer's specifications). Apparently the deviation between the actual voltage and the measured voltage does not fluctuate irregularly, rather there is a high degree of consistency of the form:

$$V_{\text{actual}} = a + bV_{\text{measured}} \tag{0.1}$$

where a is small and $b \approx 1$. This is exactly the sort of behavior expected with calibration errors. Using the manufacturer's specifications (essentially $\delta V/V \approx .001$ and $\delta I/I \approx .005$) we would expect any resistance calculated by V/I to have a relative error of $\sqrt{.1^2 + .5^2} =$ .51% (i.e., an absolute error of ±.025 kΩ for this resistor) whereas Linfit reported an error 17 times smaller. (If the errors were unbiased and random, Linfit could properly report some error reduction due to "averaging:" using all $N = 12$ data points—perhaps an error reduction by a factor of $N^{1/2} \approx 3.5$—but not by a factor of 17.) Linfit has ignored the systematic error that was entered and is basing its error estimate just on the deviation between data and fit. (Do notice that Linfit warned of this problem when it noted the small reduced χ^2.)

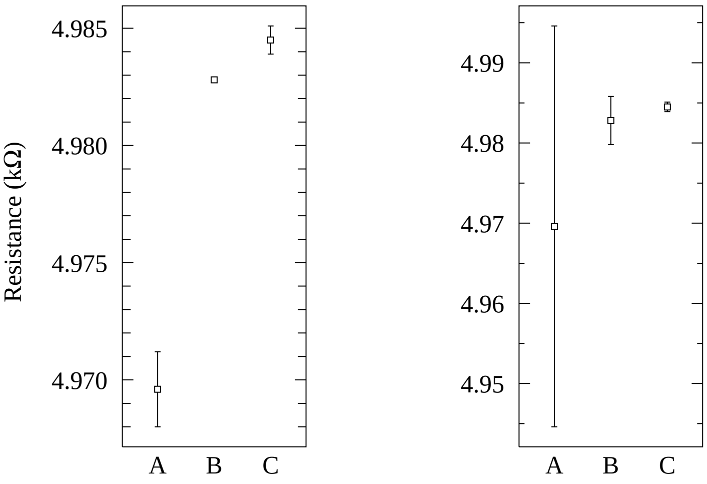


Figure 3: Three different experiments are used to determine resistance: (A) a pair of DM-441B: V/I, (B) a pair of Keithley 6-digit DMM: V/I, (C) a Keithley 6-digit DMM direct R. The left plot displays the results with error bars determined from Linfit; the right plot displays errors calculated using each device's specifications. Note that according to Linfit errors the measurements are inconsistent whereas they are consistent using the error directly calculated using each device's specifications.

When the experiment was repeated with 6-digit meters, the result was $R = 4.9828 \pm .0001$ kΩ with a reduced χ^2 of .03. (So calibration errors were again a problem and the two measurements of R are inconsistent.) Direct application of the manufacturer's specifications to a V/I calculation produced a 30× larger error: ±.003 kΩ

A direct measurement of R with a third 6-digit DMM, resulted in $R = 4.9845 \pm .0006$ kΩ.

Notice that if Linfit errors are reported as accurate I will be embarrassed by future measurements which will point out the inconsistency. On the other hand direct use of calibration errors produces no inconsistency. (The graphical display in Figure 3 of these numerical results is clearly the best way to appreciate the problem.) How can we know in advance which errors to report? Reduced χ^2 much greater or much less than one is always a signal that there is a problem with the fit (and particularly with any reported error).

Lesson: Fitting programs are designed with random error in mind and hence do not properly include systematic errors. When systematic errors dominate random errors, computer reported 'errors' are some sort of nonsense.

Comment: If a high precision resistance measurement is required there is no substitute for making sure that when the DMM reads 1.00000 V the actual voltage is also 1.00000 V. Calibration services exist to periodically (typically annually) check that the meters read true. (However, our SJU DMMs are not calibrated periodically.)

Warning: Statistics seems to suggest that arbitrarily small uncertainties can be obtained simply by taking more data. (Parameter uncertainties, like the standard deviation of the mean, will approach zero in proportion to the inverse square-root of the number of data points.) This promise of asymptotic perfection is based on the assumption that errors are exactly unbiased — so that with a large number of data points the errors will cancel and the underlying actual mean behavior will be revealed. However, in real experiments the errors are almost never unbiased; systematic errors cannot generally be removed by averaging. Care is always required in interpreting computer reported uncertainties. You must always use your judgment to decide if your equipment really has the ability to determine the parameters to accuracy suggested by computer analysis. You should particularly be on your guard when large datasets have resulted in errors much smaller than those reported for the individual data points.

Measure Twice: Systematic Error's Bane

In the thermionic emission lab you will measure how various properties of a hot tungsten wire are affected by its temperature. The presence of some problem with the thermionic lab measurements is revealed by the odd reduced χ^2 in fits, but how can we determine which measurements are the source of the problem? Systematic errors are most commonly found by measuring the same quantity using two different methods and not getting the same result. (And this will be the approach in this course: you will often be asked to measure a quantity (e.g., path length, temperature, plasma number density) using two different methods, and find different answers.) Under these circumstances we can use the deviation between the two different measurements as an estimate for the systematic error. (Of course, the error could also be even larger than this estimate!)

Problem of Definition

Often experiments require judgment. The required judgments often seem insignificant: Is this the peak of the resonance curve? Is A now lined up with B? Is the image now best in focus? Is this the start and end of one fringe? While it may seem that anyone would make the same judgments, history has shown that often such judgments contain small observer biases. "Problem of definition errors" are errors associated with such judgments.

Historical Aside: The "personal equation" and the standard deviation of the mean.

Historically the first attempts at precision measurement were in astrometry (accurate measurement of positions in the sky) and geodesy (accurate measurement of positions on Earth). In both cases the simplest possible measurement was required: lining up an object of interest with a crosshair and recording the data point. By repeatedly making these measurements, the mean position was very accurately determined. (The standard deviation of the mean is the standard deviation of the measurements divided by the square root of the number of measurements. So averaging 100 measurements allowed the error to be reduced by a factor of 10.) It was slowly (and painfully: people were fired for being 'poor' observers) determined that even as simple an observation as lining up A and B was seen differently by different people. Astronomers call this the "personal equation": an extra adjustment to be made to an observer's measurements to be consistent with other observers' measurements. This small bias would never have been noticed without the error-reduction produced by averaging. Do notice that in this case the mean value was not the ‘correct’ value: the personal equation was needed to remove unconscious biases. Any time you use the standard deviation of the mean to substantially reduce error, you must be sure that the random component you seek to remove is exactly unbiased, that is the mean answer is the correct answer.

In the bubble chamber lab, you will make path-length measurements from which you will determine a particle’s mass. Length measurements (like any measurement) are subject to error, say 0.1 mm. A computer will actually calculate the distance, but you have to judge (and mark) the beginning and end of the paths. The resulting error is a combination of instrument errors and judgment errors (problem of definition errors). Both of these errors have a random component and a systematic component (calibration errors for the machine, unconscious bias in your judgments). A relatively unsophisticated statistical treatment of these length measurements produces a rather large uncertainty in the average path length (and hence in the particle’s mass calculated from this length). However, a more sophisticated treatment of the same length data produces an incredibly small estimated length error much less than 0.1 mm. Of course it’s the aim of fancy methods to give ‘more bang for the buck’ (i.e., smaller errors for the same inputs), however no amount of statistical manipulation can remove built in biases, which act just like systematic (non-fluctuating) calibration errors. Personal choices about the exact location of path-beginning and path-end will bias length measurements, so while random length errors can be reduced by averaging (or fancy statistical methods), the silent systematic errors will remain.

Experiment: Measuring Resistance II

If the maximum applied voltage in the resistance experiment is increased from ±10 V to ±40 V a new problem arises. The reduced χ^2 for a linear fit balloons by a factor of about 50. The problem here is that our simple model for the resistor $I = V/R$ (where R is a constant) ignores the dependence of resistance on temperature. At the extremes of voltage (±40 V) about $\frac{1}{3}$ W of heat is being dumped into the resistor: it will not remain at room temperature. If we modify the model of a resistor to include power’s influence on temperature and hence on resistance, say:

$$I = \frac{V}{k_1(1 + k_2 V^2)} \tag{0.2}$$

(where fitting constant k_1 represents the room temperature resistance and k_2 is a factor allowing the electrical power dissipated in the resistor to influence that resistance), we return to the (too small) value of reduced χ^2 seen with linear fits to lower voltage data. However even with this fix it is found that the fit parameters depend on the order the data is taken. Because of ‘thermal inertia’ the temperature (and hence the resistance) of the resistor will lag the $t \to \infty$ temperature: T will be a bit low if the resistor heating up during data collection or a bit high if the resistor is cooling down. The amount of this lag will depend on the amount of time the resistor is allowed to equilibrate to a new applied voltage. Dependence of data on history (order of data collection) is called *hysteresis*.

You might guess that the solution to this ‘problem’ is to always use the most accurate model of the system under study. However it is known that that resistance of resistors depends on pressure, magnetic field, ambient radiation, and its history of exposure to these quantities. Very commonly we simply don’t care about things at this level of detail and seek the fewest possible parameters to 'adequately' describe the system. A resistor subjected to extremes of voltage does not actually have a resistance. Nevertheless that single number does go a long way in describing the resistor. *With luck*, the fit parameters of a too-simple model have some resemblance to reality. In the case of our Ohm's law resistance experiment, the resulting value is something of an average of the high and low temperature resistances. However, it is unlikely that the computer-reported error in a fit parameter has any significant connection to reality (like the difference between the high and low temperature resistances) since the error will depend on the number of data points used.

The quote often attributed[2] to Einstein: "things should be made as simple as possible, but not simpler" I hope makes clear that part of art of physics is to recognize the fruitful simplifications.

Lesson: We are always fitting less-than-perfect theories to less-than-perfect data. The meaning of of the resulting parameters (and certainly the error in those parameters) is never immediately clear: judgment is almost always required.

The Spherical Cow

> I conceive that the chief aim of the physicist in discussing a theoretical problem is to obtain 'insight' — to see which of the numerous factors are particularly concerned in any effect and how they work together to give it. For this purpose a legitimate approximation is not just an unavoidable evil; it is a discernment that certain factors — certain complications of the problem — do not contribute appreciably to the result. We satisfy ourselves that they may be left aside; and the mechanism stands out more clearly freed from these irrelevancies. This discernment is only a continuation of a task begun by the physicist before the mathematical premises of the problem could even be stated; for in any natural problem the actual conditions are of extreme complexity and the first step is to select those which have an essential influence on the result — in short, to get hold of the right end of the stick.
>
> A. S. Eddington, *The Internal Constitution of the Stars*, 1926, pp 101–2

As Eddington states above, the real world is filled with an infinity of details which a priori might affect an experimental outcome (e.g., the phase of the Moon). If the infinity of details are all equally important, science cannot proceed. Science's hope is that a beginning may be made by striping out as much of that detail as possible ('simple as possible'). If the resulting model behaves —at least a little bit— like the real world, we *may* have a hold on the right end of the stick.

The short hand name for a too-simple model is a "spherical cow" (yes there is even a book with that title: Clemens QH541.15.M34 1985). The name comes from a joke that every physicist is required to learn:

> Ever lower milk prices force a Wisconsin dairy farmer to try desperate—even

[2] "The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience" p.9 *On the Method of Theoretical Physics* is an actual Einstein quote, if not as pithy—or simple.

> crazy—methods to improve milk production. At he end of his rope, he drives to Madison to consult with the greatest seer available: a theoretical physicist. The physicist listens to him, asks a few questions, and then says he'll take the assignment, and that it will take only a few hours to solve the problem. A few weeks later, the physicist phones the farmer, and says "I've got the answer. The solution turned out to be a bit more complicated than I thought and I'm presenting it at this afternoon's Theory Seminar". At the seminar the farmer finds a handful of people drinking tea and munching on cookies—none of whom looks like a farmer. As the talk begins the physicist approaches the blackboard and draws a big circle. "First, we assume a spherical cow..." (Yes that is the punch line)

One hopes (as in the spherical cow story) that approximations are clearly reported in derivations. Indeed, many of the 'problems' you'll face this semester stem from using high accuracy test equipment to test an approximate theory. (It may be helpful to recall the 191 lab on measuring the kinetic coefficient of friction in which you found that accurate measurement invalidated $F = \mu_k N$ where μ_k was a constant. Nevertheless 'coefficient of friction' is a useful approximation.)

For example, in the Langmuir's probe lab we assume that the plasma is in thermal equilibrium, i.e., that the electrons follow the Maxwell-Boltzmann speed distribution and make a host of additional approximations that, when tested, turn out to be not exactly true. In that lab, you will find an explicit discussion of the error (20% !) in the *theoretical* equation Eq. 7.53.

$$J_i \approx \frac{1}{2}\, e n_\infty \sqrt{\frac{kT_e}{M_i}} \tag{0.3}$$

Again this 'error' is not a result of a measurement, but simply a report that if the theory is done with slightly different simplifications, different equations result. Only rarely are errors reported in theoretical results, but they almost always have them! (Use of flawed or approximate parameters is actually quite common, particularly in engineering and process control—where consistent conditions rather than fundamental parameters are the main concern.)

What can be done when the model seems to produce a useful, but statistically invalid fit to the data?

0. **Use it!** Perhaps the deviations are insignificant for the engineering problem at hand, in which case you may not care to expore the reasons for the 'small' (compared to what matters) deviations, and instead use the model as a 'good enough' approximation to reality.

1. **Find a model that works.** This obvious solution is always the best solution, but often (as in these labs) not a practical solution, given the constraints.

2. **Monte Carlo simulation of the experiment.** If you fully understand the processes going on in the experiment, you can perhaps simulate the entire process on a computer: the computer simulates the experimental apparatus, producing simulated data sets which can be analyzed using the flawed model. One can detect differences (biases and/or random fluctuation) between the fit parameters and the 'actual' values (which are known because they are set inside the computer program).

3. **Repeat the experiment and report the fluctuation of the fit parameters.** In some sense the reporting of parameter errors is damage control: you can only be labeled a fraud and a cheat if, when reproducing your work, folks find results outside of the ballpark you specify. You can play it safe by redoing the experiment yourself and finding the likely range (standard deviation) of variation in fit parameters. In this case one wants to be careful to state that *parameter values* are being reported not physical parameters (e.g., 'indicated temperature' rather than actual temperature). Again, since systematic errors do not result in fluctuation, the likely deviation between the physical parameters and the fit parameters is not known. This was the approach used in the 191 μ_k experiment.

4. **Use bootstrapping[3] to simulate multiple actual experiments.** Bootstrapping 'resamples' (i.e., takes subsets) from the one in-hand data set, and subjects these subsets to the same fitting procedure. Variation in the fit parameters can then be reported as bootstrap estimates of parameter variation. The program `fit` can bootstrap. (Again: report that an unknown amount of systematic error is likely to be present.)

5. **Fudge the data.**

   > In dire circumstances, you might try scaling all your x and y error bars by a constant factor until the probability is acceptable (0.5, say), to get plausible values for σ_A and σ_B.
   >
   > *Numerical Recipes* by Press, et al., 3rd ed. p. 787

   Increase the size of your error bars so you get reduced $\chi^2 = 1$, and then calculate errors as in the usual approach. Clearly this is the least legitimate procedure (but it is what *LINFIT* does). One must warn readers of the dicey nature of the resulting error estimates. The program `fit` can fudge.

Special Problem: Temperature

Measuring temperature is a particular problem. (Here, the first two labs involve measuring temperatures above 1000 K in situations a bit removed from the experimenter.) You may remember from 211 that while temperature is a common part of human experience, it has a strikingly abstruse definition:

$$\frac{1}{kT} \equiv \frac{\partial \ln \Omega}{\partial E} \tag{0.4}$$

While the usual properties of Newtonian physics (mass, position, velocity, etc.) exist at any time, temperature is a property that exists contingent on a situation: 'thermal equilibrium'. And thermal equilibrium is an idealization only approximately achieved—never exactly achieved—in real life. Furthermore in these experiments, thermal equilibrium is not even closely approximated, so the resulting temperatures have somewhat restricted meanings.

In the photometry lab 'the temperature of stars' is measured. In fact stars do not have a temperature and are not in thermal equilibrium. Nevertheless, astronomers find it useful to define an 'effective temperature' which is really just a fit parameter that is adjusted for the best match between the light produced by the star and the light predicted by the model.

[3] wiki Bootstrapping (statistics)

Special Problem: Assuming Away Variation

In the 191 μ_k lab, you assumed the sliding motion was characterized by one value of μ_k, whereas a little experimentation finds usually slippery and sticky locations (handprints?). In the thermionic emission lab you will measure how various properties of a hot wire depend on temperature, however the hot wire does not actually have a temperature: near the supports the wire is cooled by those supports and hence is at a lower temperature. Our spherical cow models have simplified away actual variation. The hope is that the fit model will thread between the extrems and find something like the typical value. Of course, real variations will result in deviations-from-fit which will be detected if sufficiently accurate measurements are made.

Special Problem: Derive in Idealized Geometry, Measure in Real Geometry

Often results are derived in simplified geometry: perfect spheres, infinite cylinders, flat planes, whereas measurements are made in this imperfect world. In these labs (and often in real life) these complications are set aside; instead of waiting for perfect theory, experiment can test if we have "the right end of the stick". Thus a Spherical Cow is born. The theory should of course be re-done using the actual geometry, but often such calculations are extremely complex. Engineering can often proceed perfectly adequately with such a first approximation (with due allowance for a safety factor) and, practically speaking, we simply may not need accuracy beyond a certain number of sigfigs. Indeed it takes a special breed of physicist to push for the next sigfig; such folks are often found in national standards labs like `nist.gov`.

The Fit Elephant

I remember a public lecture in the late 1970s by the theoretical astrophysicist Don Cox, in which he said

> Give me one free paramter and I'll give you an elephant. Give me two and I'll make it wag its tail

Cox was certainly not the originator[4] of this sort of statment, for example, Freeman Dyson writes[5] that in 1953 Enrico Fermi quoted Johnny von Neumann as saying:

> with four parameters I can fit an elephant and with five I can make him wiggle his trunk

The fit elephant is the opposite of the spherical cow: totally unconstrained parameters are added willy-nilly to the model in order to chase the data. The parameter k_2 in the hot

[4] Brown & Sethna, Phys.Rev.E, **68** 021904 (2003), reports attributions to C.F. Gauss, Niels Bohr, Lord Kelvin, Enrico Fermi, and Richard Feynman; I would add Eugene Wigner. The first Google Books hit is in 1959.

[5] *Nature* **427**, 297 (2004)

resistor equation (Eq. 0.2) is potentially such a dangerously free parameter: I will accept any value the computer suggests if ony it improves the fit. While I have provided a story which suggests why such a term might be present, I have not actually checked that there is any truth in the story (for example, by measureing the actual temperature of the resistor at high voltage and by measureing the resistance of the resistor when placed in an oven). Skepticism about such inventions is expressed as Occam's razor[6] and the law of parsimony.

Purpose:

In all your physics labs we have stressed the importance of 'error analysis'. However, in this course you will have little use for that form of error analysis (because it was based on computer reports of random errors). Instead, my aim in this course is to introduce you to the problems of non-random error. In the bubble chamber lab you will see how increasingly sophisticated analysis can reveal systematic error not important or evident in more elementary analysis. In the other labs you will see how systematic error can be revealed by measuring the same quantity using different methods. In all of these labs you will use too simple theory to extract characterizing parameters, which are not exactly the same quantity as might occur in a perfect version of the problem.

Comment:

The lesson: "measure twice using different methods" is often impractical in real life. The real message is to be constantly aware that the numbers displayed on the meter may not be the truth. Be vigilant; check calibrations and assumptions whenever you can. But the opening Shapley quotation tells the truth: "Experimental findings... are messy, inexact things, which are believed by everyone except the man who did the work".

[6] "entia non sunt multiplicanda praeter necessitatem", roughly (Wiki) translated as "entities must not be multiplied beyond necessity".