The Statistical Treatment of Experimental Data[1]

Introduction

The subject of statistical data analysis is regarded as crucial by most scientists, since error-free measurement is impossible in virtually all experimental sciences, natural or social. Experimentalists gather data with the aim of formulating a physically reasonable model to describe a particular phenomenon. But, given an experimental data set, how does one know whether it agrees with a theoretical prediction? There is no guarantee that a theory will describe the real situation — the basis of science is the demand that any idea be validated experimentally. To know whether your measurement is consistent with theory or not, you need to know whether the uncertainty in the measurement is smaller or larger than the discrepancy between the measurement and the theoretical prediction. In symbols we would say that if ν_{th} denotes the theoretically predicted value of a given quantity, while ν_{best} denotes the best experimental estimate, and $\delta\nu$ denotes the uncertainty expected in the measurement, then the theory is consistent with experiment only if

$$\left(\nu_{\text{best}} - \delta\nu\right) \lessapprox \nu_{\text{th}} \lessapprox \left(\nu_{\text{best}} + \delta\nu\right) \qquad \text{(S-1)}$$

The next question is how to determine $\delta\nu$ for a particular experiment and measurement apparatus. The candid truth is that $\delta\nu$ is often estimated from one's knowledge of the precision of the measuring device (sometimes this is called an "external" determination). For example in measuring the length of an object using a meter stick whose finest marks are 1 mm apart, one should *attempt* to make measurements to a fraction of 1 mm. Depending upon the object being measured, and the ability to clearly align the ruler with the object, one might estimate, for example, that $\delta\nu$ would be about 0.5 mm (but you might estimate something different from that!). You should be aware that when scientists refer to experimental error, they usually mean the uncertainty in their measurement, which is a measure of **precision** and **not accuracy**. With this in mind, we realize that the spread observed in a series of repeated measurements allows us to determine uncertainties based upon statistics (sometimes called an "internal" determination).

It is important to keep in mind the distinction between *systematic* and *statistical* uncertainty. Imagine that because of the humidity in Central Illinois at this time of year, the meter stick you are using has swelled in length by a small amount. In that case, it will underestimate by a corresponding amount the length of any object it is used to measure. We call

---

[1] Some of this advice originated within the Physics Department at Haverford College, where Gabe first taught as a Visiting Assistant Professor. The title and some of the material is based upon the treatment found in H. D. Young, *Statistical Treatment of Experimental Data* (McGraw-Hill, New York, 1962), which is a fine reference to the subject.

that sort of uncertainty *systematic* and try to design experiments to avoid it, because if it is overlooked, it will lead to an incorrect (inaccurate) result, despite a full statistical analysis which might give us a measure of the precision of our technique. You may end up finding some example of insidious systematic errors in this lab. *Statistical uncertainty,* on the other hand, is the uncertainty that is due to small random events beyond our control, that do not lead to bias in either direction. Statistical uncertainty can be reduced by repeating measurements, and it can be accurately determined from the spread of values that occur. Therefore we prefer it to systematic error. ;)

The following material is devoted mostly to determination of statistical uncertainty and something called the *propagation of uncertainty.* The theoretical results you will learn below are mathematically exactly true for a certain, commonly encountered type of statistical uncertainty (*i.e.*, based on a Gaussian distribution). [See note at end of write-up]

Dealing with Systematic Error

Generally speaking, a series of independent *measurements* allow us to reduce random errors, but independent *experiments* are required to gauge systematic errors. Still there are often simple checks that allow you to subtract offsets or drift which might otherwise be included in your dataset. Imagine you have been assigned to determine whether a particular radiation source represents a health hazard. To determine the health hazard, you might use a radiation detector called a Geiger counter to measure incoming high-energy particles given off by atomic nuclei as they undergo radioactive decay. However, if you turned on your detector, even with no source of radioactivity nearby, you would detect background radiation. This is because high-energy particles, mostly electrons and gamma rays, arrive at the earth from space every moment. These "cosmic rays" set off your detector, and they cannot be eliminated. In addition, there are some radioactive nuclei in building materials, so that the bricks and mortar of our lab building are emitting radiation, too! Even our bodies contain unstable ^{40}K nuclei that contribute slightly to the background radiation level. This background intensity (*number of counts per unit time*) represents a source of *systematic* uncertainty, since it adds to any measurement of radiation we make in the room. We would need to subtract off this background intensity from our measurement of any extra intensity caused by our radioactive source. The quantity of interest, the source intensity, is

$$I_s = I_t - I_b \tag{S-2}$$

where I_s indicates the source intensity, I_t the total intensity, and I_b the background intensity. The uncertainty in our measurement of I_s still needs to be determined.

Determining Statistical Uncertainty

It stands to reason that repeating measurements is a good way to increase the precision of the overall process. (You would never poll only three people to establish an approval rating for the President.) It turns out that from a reasonably large sample of measured values we can also accurately estimate the uncertainty!

Let's suppose that, here, N is the number of groups reporting measurements, supplying us with N independently made measurements, $\{I_1, I_2, \ldots I_N\}$, of some quantity I, such as the *background* radiation intensity. Then the best estimate of the actual value of I is the *mean* (or average) of the N values.

$$\bar{I} = \frac{1}{N}\sum_{i=1}^{N} I_i. \tag{S-3}$$

We also want to deduce δI, the uncertainty associated with this measured value. We can estimate this quantity from the spread of individual values of the I_i's that occurred. It should also depend on how many measurements were made, but we'll deal with that after we decide how to characterize the spread of values obtained. Statisticians and physicists have devised an accepted measure of the spread of values that occur in a distribution and call it the *standard deviation*, σ. It is defined in terms of an average of the deviation of the individual measurements from the average value $\bar{I}$. (If we just averaged the deviations from the mean, we would get roughly zero, since the measurements above and below the mean would almost cancel out. Therefore we average the square of the deviations, to get rid of the signs of those deviations, and take the square root at the end to get a measure with the right units.) The definition of σ will be

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(I_i - \bar{I}\right)^2}. \tag{S-4}$$

Eq. (S-4) is a bit of a pain even on a calculator and it can be shown that the following expression is equivalent for large N, so we usually use it:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} I_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} I_i\right)^2} \tag{S-5}$$

The second term under the square root sign is just $\bar{I}^2$, so that part is easy to obtain. The only real new punching of calculator buttons is the accumulation of the sum of the squares of the I_i's in the first term.

NOTE: If N is a relatively small number, the uncertainty estimated from (S-5) is slightly underestimated. (After all, if $N = 1$, you would get $\sigma = 0$ erroneously.) To approximately correct for this underestimate, use $(N - 1)$ instead of N in equation S-5 if N is less than 10 or so.

One might be tempted to assume that σ is also a good estimate of δI. That is, if you took many identical data sets and used each of them to compute I_{best}, one might suppose that the standard deviation of that *set* of I_{best} would be σ. That is *not* the case. Particularly for large *N* we will know $\bar{I}_{\text{best}}$ considerably more accurately than $\bar{I}_{\text{best}} \pm \sigma$. In fact you would find that if we all did another measurement, thereby doubling *N*, we would find that σ *itself* would not change appreciably, since it is just a measure of the natural spread in values resulting from our measuring procedure, and yet our knowledge of $\bar{I}_{\text{best}}$ would be improved. It can be shown that a better estimate for the error bar is the so-called *standard deviation of the mean*, σ_m, given by

$$\sigma_m = \frac{\sigma}{\sqrt{N}} \tag{S-6}$$

Thus we would typically say we have determined *I* to be $I = \bar{I} \pm \sigma_m$. The exact meaning of this statement is discussed in texts on Error Analysis, but the following interpretation is good enough for most purposes: there is about a 2/3 chance that the actual value of *I* lies in the interval $\left[\bar{I} - \sigma_m, \bar{I} + \sigma_m\right]$, and a 95% chance that it is in the interval $\left[\bar{I} - 2\sigma_m, \bar{I} + 2\sigma_m\right]$, and a 99.7% chance that it is in the interval $\left[\bar{I} - 3\sigma_m, \bar{I} + 3\sigma_m\right]$, *etc*. In many situations scientists will choose caution, and would report the uncertainty, δI, as being larger than σ_m, *e.g.* $2\sigma_m$, but ***context*** determines whether caution of this sort is appropriate or inappropriate. Our advice is that you let your audience know, explicitly, what choices you've made.

Propagation of Uncertainties

It frequently happens in experimental science that the desired quantity depends on *several* other quantities, each of which has an associated uncertainty. In that case we need to know how to compute the uncertainty in the final quantity from those of the measured quantities. This process is called "propagation of uncertainty."

This topic is treated in numerous texts, such as *Experimentation: An Introduction to Measurement Theory and Experiment Design*, by D.C. Baird or *Data Reduction and Error Analysis for the Physical Sciences*, by P.R. Bevington and D.K. Robinson. We will simply summarize several simple results that will be useful to you in this course. First, suppose that you have measured two quantities, *a* and *b*, that are added or subtracted to produce a final result. If we may presume that there are **no correlations** between the *fluctuations* in our measurements of *a* and the *fluctuations* in our measurement of *b*, then the best estimate of the uncertainty of the sum or difference is

$$\delta\left(a \pm b\right) = \sqrt{(\delta a)^2 + \left(\delta b\right)^2}. \tag{S-7}$$

We say that the uncertainties are added "in quadrature." Note that if the two uncertainties are equal, the resultant uncertainty is larger than either one alone by a factor of $\sqrt{2}$. Also, **if the uncertainty in one is less than about 1/4 of the other, its contribution to the overall uncertainty is negligible**.

In other experiments, you may be multiplying or dividing two quantities. In that case, the *relative* or *fractional* uncertainties that are added in quadrature:

$$\frac{\delta(ab)}{ab} = \sqrt{\left(\frac{\delta a}{a}\right)^2 + \left(\frac{\delta b}{b}\right)^2}. \tag{S-8}$$

This implies, for example, that if δa is 2.0% of a and δb is 4.0% of b, then $\delta a\, b$ is $[(0.020)^2 + (0.040)^2]^{1/2}$, or 4.5%, of ab.

There are more complicated rules for other manipulations. For example, if you raise a measured value to a power the relative uncertainty is scaled by that power

$$\frac{\delta(a^n)}{a^n} = n\frac{\delta a}{a} \tag{S-9}$$

where it has been assumed that there is no uncertainty in n, only in a.

Such rules can be ***summarized*** as follows: for function of two independent variables, $z = f(x, y)$:

$$\delta z = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 \delta x + \left(\frac{\partial z}{\partial y}\right)^2 \delta y^2} \tag{S-10}$$

If z is a function of more than two variables the equation is extended by adding similar terms.

> NOTE: These results are mathematically exact only for *Gaussian* distributions. Some sorts of data, radioactive **counting** measurements for example, actually follow *Poisson* distributions, which are slightly different. However, the error introduced by this fact is often relatively small.