MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Physics Department

Physics 8.01T Fall Term 2004

Introductory Error Analysis

These guidelines will be augmented as the term progresses. For each topic, two sets of guidelines will be given:

• A brief description of the terminology and method of calculuation.

• An outline of the theory behind the formulas. An attempt has been made to have these presentations be algebra-based, while the calculus-based formulas are presented for those so inclined.

Sampling a Population: Averages and Standard Deviation

Definitions and Notation

(This presentation will use the notation $\langle x \rangle$ for the "average value of x," as opposed to other notation, such as $\bar{x}$.)

The *average*, more precisely the *arithmetic mean* of n samples of a measurement of the quantity x, denoted x_1, x_2, $\ldots x_n$, or more compactly, $\{x_1, x_2, \ldots, x_n\}$, or even $\{x_i\}_{i=1}^{n}$, is denoted

$$\langle x \rangle = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The *variance*, σ^2, is the average of the squares of the distances from the average, or

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \langle x \rangle\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \langle x \rangle^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2.\end{aligned}$$

The last expression, derived later in these notes, is "the average of the square minus the square of the average," and is included here to demonstrate the relative ease

of calculation; the measurememnts and their squares are stored separately, and the average need not be recalculated if data are added or altered.

The *standard deviation* σ is then

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^2} = \sqrt{\langle x^2\rangle - \langle x\rangle^2}$$

is sometimes known as the "root-mean-square" of the sample.

Another commonly-used set of measures is the *sample variance*, and the *sample standard deviation* defined in a manner similar to σ^2,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^2},$$

so that $s = \sqrt{\frac{n}{n-1}}\,\sigma$. Of course, for n large, $s \sim \sigma$. For small n, however, the distinction should be made, and users should be aware of which measurement is being used. For instance, the nearest handheld calculator I can find has both s and σ; once one is calculated, the other is very simple, and so it's easy to include both features.

Numerical Methods

For a small number of data measurements, it might well be more convenient to use a handheld calculator with statistics capabilities. It should not be too much of a generalization to say that any scientific calculator sold today has such capabilities. The operators of these handhelds will be left to interpret the uses thereof on their own - the notation varies too much from make to make and model to model.

What follows are instructions on how to use Xess, the default spreadsheet on server at MIT. Other spreadsheet programs (such as Excel) will use similar notation or commands, but not all can be given here.

Let's assusme we have ten data points,

$$\{x_i\} = \{3.23,\ 3.15,\ 3.25,\ 3.21,\ 3.00,\ 3.05,\ 3.17,\ 3.18,\ 3.15,\ 3.13\}\,.$$

These were generated by the "Random Number" feature on a handheld, and are not meant to correspond to anything physical. The overall closeness to π is, we hope, a fluke. Maybe we'll see how "random" they are.

To start Xess, use the Dash/Menubar, whichever is convenient.

When the Xess window appears, merely enter the data in Column A (of course, it doesn't matter which column you choose). Data are typed in the line at the top of the sheet (the "Edit line") and entered with a return, checking the green checkmark or hitting the down-arrow key. Note the the 3.00 data point is recorded as 3. This can be fixed by highlighting the column, going to "Format" on the menu bar, selecting "Cell Format", choosing "Fixed" and setting the number of decimal places at 2.

The needed statistical functions are evaluated as follows: Click on the cell where you want the result to appear, and type the given command into the Edit line.

Average:	`=@AVG(A1..A10)`
Standard Deviation:	`=@STD(A1..A10)`
Sample Standard Deviation:	`=@STDS(A1..A10)`

In each of the above, the "`=`" merely indicates that a numerical result is given; the "`@`" is used to call a specific function; and a range must be given, in each of the above cases the cells A1 through A10. Entries for the function and the range are not case-sensitive. The precision with with the results are displayed may be changed as described above using the "Cell Format" feature.

I get $\sigma \sim 0.0776$; If I took many more samples, and the sampling were truly random, I would expect $1/(8\sqrt{3}) \sim 0.0722$.

In one of the earlier experiments, you are told that it's okay to use an "average deviation," the average of the sum of the absolute values of the deviations, specifically

$$\text{Average Deviation} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \langle x \rangle| = \overline{|x - \langle x \rangle|}.$$

It turns out that if there are more than a few points, this is difficult to calculate by hand, as the absolute value does not reduce algebraically as the square of a binomial does. However, the machines don't care. You do, however, have to use an extra step, in that the average has to be calculated first.

So, if you have found the average in, for instance, cell B1, and enter into an empty cell (is C1 still empty?) `=@ABS(A1-$B$1)`. The dollar signs mean that the cell reference is "absolute" (nothing to do with `ABS`, the absolute value function), as opposed to `A1`, which is a "relative" cell reference. This means that if you now enter the function, click on cell C1 and either "Control+f" or select from the Edit feature

on the menu bar to copy the formula into cells C2..C10, the respective absolute values of the differences will be entered. Then, in any other available cell, entering the function `=@AVG(C1..C10` will give the average deviation.

There's much that can be done, but little that's useful for our immediate purposes. The data can be graphed by selecting the desired column and using the "Graph" feature from the Menu bar and selecting "New Graph." Either "Scatter" or "Bar Graph" might tell you something.

The real advantage to using the spreadsheet is being able to display all of the data and calculations simultaneously, making editing easier.

Where the Formulas Come From

Of course, everyone knows what the "average" is, but it turns out that there's a slightly subtle reason why we use the arithmetic mean, a reason that will serve us well in other applications.

Suppose we call the "best guess" of our sample c, and consider the variance with respect to c, in the form

$$V_c = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 .$$

Then, V_c clearly is positive ($V_c = 0$ will not be considered) and depends on c. We want to adjust c so that V_c is minimized.

To minimize V_c, note that

$$\begin{aligned} V_c &= \frac{1}{n}\sum_i \left(x_i^2 - 2c\,x_i + c^2\right) \\ &= \frac{1}{n}\sum_i x_i^2 - 2c\,\langle x\rangle + c^2 \\ &= \frac{1}{n}\sum_i x_i^2 - \langle x\rangle^2 + \left(c^2 - 2c\,\langle x\rangle + \langle x\rangle^2\right) \\ &= \frac{1}{n}\sum_i x_i^2 - \langle x\rangle^2 + (c - \langle x\rangle)^2. \end{aligned}$$

The value of c that minimizes this expression is that which makes the term in parantheses in the last expression above zero. If one chooses to use calculus,

$$\begin{aligned} \frac{dV_c}{dc} &= -\frac{1}{n}\sum_i 2(x_i - c) = -\frac{2}{n}\left(\sum_i x_i - \sum_i c\right) \\ &= -\frac{2}{n}(n\langle x\rangle - n\,c) = 0. \end{aligned}$$

Either way, V_c is minimized at $c = \langle x \rangle$. As we have seen, this least variance will be denoted as σ^2. Explicitly,

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_i (x_i - \langle x \rangle)^2 = \frac{1}{n}\sum_i \left(x_i^2 - 2\langle x \rangle x_i + \langle x \rangle^2\right) \\ &= \frac{1}{n}\left(\sum_i x_i^2 - 2\langle x \rangle \sum_i x_i + n\langle x \rangle^2\right) \\ &= \frac{1}{n}\sum_i x_i^2 - 2\langle x \rangle\langle x \rangle + \langle x \rangle^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2,\end{aligned}$$

where $\left\langle x^2 \right\rangle \equiv \frac{1}{n}\sum_i x_i^2$, the average of the squares, has been used.

This convenient algebraic expression allows us to calculate only two averages instead of having to calculate $\langle x \rangle$, then recalculate $(x_i - \langle x \rangle)$ a total of n times, then squaring and averaging. So, as a result,

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}.$$

σ is known as the "root mean square" of the sample, that is, the square root of the average square of the distance from the mean.

There's one slight catch. How do we know the $\langle x \rangle$ measured from our n trials is the true mean? Actually, sometimes we know it can't be. For example, if we toss a fair coin n times, assigning $x_i = 0$ if toss i is heads and $x_i = 1$ if toss i is tails, we know that the true mean is $1/2$. If n is odd, however, we can't possibly get $\langle x \rangle = 1/2$.

To account for this, we revise our determination of the standard deviation;

$$\begin{aligned}s^2 &= \frac{1}{n-1}\sum_i (x_i - \langle x \rangle)^2 = \frac{n}{n-1}\,\sigma^2 \\ s &= \sqrt{\frac{n}{n-1}}\,\sigma.\end{aligned}$$

The factor of $n-1$ instead of n is often taken to mean that "one measurement is needed to find the mean, so $n-1$ are left to find the standard deviation." Another way to see the necessity of this factor is to realize that one measurement tells us nothing about the standard deviation.

If $n \gg 1$, s is essentially the same as σ. If, however, we know the mean in advance, all we are measuring is the standard deviation, so we use σ. For example, with two honest dice, we know that the average of the sum of the numbers is seven (2 is as probable as 12, 3 as likely as 11, etc).

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Physics Department

Physics 8.01T Fall Term 2004

Propagation of Error - Experiment 2

A more detailed explanation of the source of the formulas cited here will be provided later in the term. For now, we will state without proof an extension of the Pythagorean Theorem for several variables.

Specifially, if $f(x,\, y,\, z, \ldots)$, we can use calculus and geometry to find Δf. We consider how a small change in one variable affects f, square these changes and add to obtain Δf. Symbolically,

$$(\Delta f)^2 = \left(\frac{\partial f}{\partial x}\Delta x\right)^2 + \left(\frac{\partial f}{\partial y}\Delta y\right)^2 + \cdots .$$

The relation to the Pythagorean Theorem is that the terms in parentheses on the right side of the above expression represent "how far away from f" the uncertainties Δx, Δy, ... would allow the measurement to vary.

For Experiment 2, the expression for the gravitational constant g in terms of the ball diamter D, the time T measured for the ball to pass the photogate, the height y measured above the ground, the horizontal range x of the ball's path and the initial angle θ_0 is

$$\begin{aligned} g &= \frac{D^2}{T^2\, x^2}\left[2\, \cos\theta_0\, (x\, \tan\theta_0 - y)\right] \\ &= 2\, \frac{D^2}{T^2} \cos\theta_0 \left[\tan\theta_0 \frac{1}{x} - \frac{y}{x^2}\right]. \end{aligned}$$

For the purposes of estimating the uncertainty Δg, we will assume that imprecision in the diameter of the ball (given as $D = 12$ mm) and the uncertainty in the initial angle, (fixed at a multiple of $\pm 15^\circ$) contribute much less to Δg than the uncertainties in T, x and y. Be sure to realize that while the experiment writeup and your table of results use "ΔT" for the time it takes the ball to pass the photogate, that usage would be inappropriate here, so that time is denoted $T \pm \Delta T$.

The needed calculus is not hard, but not worth having everyone redo the calculations. To simplify things somehwhat, denote the term in square brackets in the last line above as $u = u\,(x,\, y) = \left[\tan\theta_0 \frac{1}{x} - \frac{y}{x^2}\right]$. The needed results (cited but not derived) are

$$\frac{\partial g}{\partial T} = -2\, \frac{g}{\Delta T}, \qquad \frac{\partial g}{\partial x} = \frac{g}{u}\frac{\partial u}{\partial x}, \qquad \frac{\partial g}{\partial y} = \frac{g}{u}\frac{\partial u}{\partial y}.$$

The net result is

$$\left(\frac{\Delta g}{g}\right)^2 = 4\left(\frac{\Delta T}{T}\right)^2 + \frac{\left[\left(\left(\tan\theta_0/x^2\right) + 2\,\left(y/x^3\right)\right)^2 (\Delta x)^2 + \left(1/x^2\right)^2 (\Delta y)^2\right]}{\left[\tan\theta_0 \frac{1}{x} - \frac{y}{x^2}\right]^2}$$

$$= 4\left(\frac{\Delta T}{T}\right)^2 + \frac{\left[\left(\tan\theta_0 + 2(y/x)\right)^2 (\Delta x)^2 + \cos^2\theta_0\,(\Delta y)^2\right]}{[x\,\tan\theta_0 - y\,]}.$$

It would be nice if the second term on the right above simplifies, but in general that won't be the case. It certainly might be expected that the more complicated the expression for f, the more complicated the expression for Δf. Consider, however, that the expression for g was not that complicated. If we introduced uncertainties in D and θ_0 as well, we'd have quite a bit of number-crunching to do.