Solutions to exercise #1 (assigned 19 September, due 28 September)

Overall I was reasonably pleased with your performance on this exercise. A few general comments: I do not need you to turn in pages and pages of data tables. For some people I do need to see more clearly what equations you used. Mostly, the amount of description was not too bad. Several homeworks were so good that they helped me find mistakes I had made in my solution set! I'm looking forward to seeing your next exercises.

A note about my grading scheme. This exercise had a number of parts, each of which I graded independently and assigned a value to: excellent, good, okay, or needs work. I then went back and assigned numerical scores to these values: 4, 3, 2, 1. Depending how you answered the questions, you might have had about four parts to your homework, so your total score will be out of 16. Some people have a total possible score of 12, or of 20. The final score I gave you is a fraction. The distribution of scores can be found on the course web page.

We start with the data in `exercise1.dat`. A technical note: I did all my analysis for this exercise with IDL. You might have used Matlab, or Excel — doesn't matter. I didn't use any canned (preprogrammed) routines except to check my results.

The first thing I did is plot the data, just to see what's there and to help me visualize how to approach the questions. I suggest that you do the same for this and for future assignments.

First we want to find the error on the mean and the mean of the errors for the first three measurements (rows) for $(y, \Delta y)$ (columns 1 and 2) only.

The error on the mean is $\sigma_y/\sqrt{N}$, where σ_y is

$$\sqrt{\frac{1}{N}\sum(\bar{y} - y_i)^2} \tag{1}$$

and where

$$\bar{y} = \frac{1}{N}\sum y_i. \tag{2}$$

So we find that $\bar{y} = 10.2263$ and $\sigma_y = 0.168062$. Note that if you compared the answer derived from our formalism to one you found from a canned routine (say, to check your results), you may have gotten a different answer! This is because formally we should use $1/(N-1)$ here, and not $1/N$, under the radical.

Finally, we find that the error on the mean is 0.0970308.

The mean of the errors is

$$\frac{1}{3}\sum_{i=1}^{3}\Delta y_i \tag{3}$$

which is 0.583072. Note that I used $|\Delta y_i|$, although you could have used Δy (signed). That's a bit ambiguous, because this isn't really a real data set, just something I made up, so it's not clear whether you want to use the absolute value or not. I think it made sense to do so, because that captures the magnitude of the error, and because I didn't want a significantly positive and significantly negative error to cancel each other out. The real test, though, is that if you use signed values then the mean of the error goes to zero, which you know can't be right — should be an indication that something funny is going on, and that perhaps you should use unsigned values.

How many significant figures did you use for these two steps, and why?

Now we just proceed. I obtained the following results:

Number of rows	E.O.M.	M.O.E.
3	0.0970308	0.583072
10	0.311331	0.478259
50	0.157199	0.416409
100	0.113999	0.483937

Do the E.O.M. and M.O.E. behave as I expected? Mostly. The E.O.M. should go down as N increases and, except for the first case (just 3 rows), that is what we see. M.O.E. should be roughly independent of N, and that's mostly what we observe, again with the exception of $N = 3$.

Now we want to calculate the mean, median, and mode for these same (sub)samples. The mean is simply $(1/N)\sum y_i$ – easy to find. The median is the data point at which half the values are larger and half are smaller, so just sort the data and find the middle value. For an even-number-sized data set, you could report the mean of the two middle values as the median, or choose one, or give both. I chose to take the mean value of the two middle points, which is a little bit cheating, since there is no actual data point with the value of the median that I report!

Finding the mode is tricky. I didn't even bother trying to find the mode for $N = 3$, since that's kind of meaningless. For $N = 10$ and larger I binned the data. I tried different binsizes (from small to large) until I found one that seemed to produce a histogram where there was a single peak that was taller than the rest. In the following table, I list the bin size I used for the various N, and I give as the mode the middle value of the bin.

Number of rows	mean	median	mode	bin size
3	10.2263	10.2500	N/A	N/A
10	11.1300	10.8331	10.35	0.3
50	11.0639	11.1640	10.35	0.3
100	11.2265	11.2153	11.3	0.2

You can see that the mode jumped up from $N = 50$ to $N = 100$ – probably something unusual about the data values in the second fifty. It is possible that the mode is not the most useful way to describe this data set ....

Now we look at all four columns, and two functions: $x = f_1(y, z) = y + 2z$ and $x = f_2(y, z) = yz^2$.

The first thing we want to do is calculate x and Δx for each row. Calculating x is easy. For $f_1(y, z)$, $\sigma_x^2 = \sigma_y^2 + 4\sigma_z^2$, taking the covariance to be negligible (see Bevington, page 43), and the E.O.M. is $\sigma_x/\sqrt{N}$. Note that when $N = 1$, the E.O.M. is zero (undefined) because σ_x is zero when $N = 1$. In other cases ($N > 1$), the E.O.M. is the errors on the mean of the values x_i. The M.O.E. values are $(1/N)\sum \Delta x_i$, which is the same as $(1/N)\sum \sigma_{x_i}$ (sorry, I switched nomenclature there). So what I do is to calculate σ_x at each i and then calculate E.O.M. and M.O.E. in the usual ways for the samples of interest. Here are my answers for $f_1(y, z)$:

Number of rows	$\overline{x}$	E.O.M.	M.O.E.	$\overline{\Delta x}$
1	84.0068	N/A	1.44542	1.44542
3	88.8740	2.46363	2.85569	3.77153
10	69.1231	5.94786	6.98137	9.17151
50	63.5037	2.84819	44.4942	44.5853
100	59.6722	2.00742	79.7955	79.8207

You can see that the M.O.E. is growing with N here! Why? Well, σ_z gets larger as you go down the list (i.e., with bigger N), and that drives σ_x up and therefore the M.O.E. That it turns makes $\overline{\Delta x}$ grow with increasing

N. You can see that for N somewhere in the range between 10 and 50, the M.O.E. starts to dominate the total error.

Now let's take the next function, $f_2(y, z)$. This is not really a very complicated function, but it's plenty hairy to deal with, as you'll see. Following the arguments on Bevington (page 43–44), and taking the covariance to be small, we can write that

$$\sigma_x^2 = \left(\frac{\partial x}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial x}{\partial z}\right)^2 \sigma_z^2 \tag{4}$$

$$= z^4\sigma_y^2 + 4y^2z^2\sigma_z^2 \tag{5}$$

Divide both sides by $x = yz^2$ and we can rewrite this relationship as the following:

$$\frac{\sigma_x^2}{x^2} = \frac{z^4\sigma_y^2}{y^2z^4} + \frac{4y^2z^2\sigma_z^2}{y^2z^4} \tag{6}$$

$$= \frac{\sigma_y^2}{y^2} + \frac{4\sigma_z^2}{z^2} \tag{7}$$

Now we can solve for all x_i and Δx_i and, as before, find the E.O.M, M.O.E., and total error. I get the following:

Number of rows	$\overline{x}$	E.O.M.	M.O.E.	$\overline{\Delta x}$
1	13940.2	N/A	1464.08	1464.08
3	15872.8	1075.86	1600.62	1928.59
10	10159.8	1652.24	2310.21	2840.24
50	8652.79	804.879	12756.9	12782.3
100	7703.01	569.421	20470.6	20478.5

As before, the total error is dominated by the M.O.E. for N larger than 10 or so largely because of σ_z.

Overall, this was a reasonably grueling problem set, but nowhere near as nasty as it could have been, or as real scientific analysis often is. The lesson is that real data analysis is messy and never elegant, and there are always a lot of things to keep track off – even when you think you are finished, you might not be, because you thought of one more thing. The other lesson is that there is not necessarily just one single correct way to carry out the analysis. You may have done something differently than I did on your homeworks. As long as it makes sense and is self-consistent and is argued convincingly, I'm reasonably happy.