Solutions to exercise #2 (assigned 26 September, due 19 October)

Overall I was very pleased with your performance on this exercise. It seems like most of you really understood the main points of the exercise. It also seems like most of you worked really hard on the exercise, and I can tell that what you get out of the exercise is proportional to what you put in. Can you tell that too? This was a pretty hard exercise, but no harder than the kinds of things you might really encounter in research science.

A few general comments: Some of you still need to think more in terms of plots and less in terms of data tables. Data tables do me no good – and they do you almost no good also, in terms of understanding your overall results. Plot up the data – it's the easiest, and frequently the only, way to see the global properties of the data. Is there still a linear trend in your residuals? The only way to tell is to look at the data. Watch out: Numbers close to one, or close to zero, are not the same as one (or zero). Small probabilities can really matter. And speaking of probability, a number of you did not say how you calculated the probabilities at 4, 5, or 6 sigma, and you should have. Finally, several of you did use canned routines for this assignment, which I explicitly asked you not to do.

A note about my grading scheme. This exercise had a number of parts, each of which I graded independently and assigned a value to. I tried to grade this exercise so that everyone's homework would be out of the same possible total, but somehow, mysteriously, I failed – sorry. The distribution of scores is on the web page. Please make sure that I counted your points correctly, and notify me if I made any mistakes. But, don't worry too much about your scores – instead, make sure that you are learning the material. I recommend reviewing the solution set with your homework carefully so that you understand what I did.

(1) We start with the data in `exercise2.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. When I plot the data I see right away that there is a sinusoidal signal (which is the "real" signal); an increasing linear trend (which is a systematic signal); and some scatter in the data (error in the measurements, presumably).

There are a couple of ways to approach this issue of cost versus fidelity of the signal that we determine. If cost is no object (ha!) then we should do fine sampling, and take 100 samples across this data range – no problem. However, if we have a cap on the amount of funds we have to spend to sample our data then we might only get, say, ten samples total. In this case, if we do fine sampling we might only get the first ten samples. If so, it would be clear that we have a sinusoid with some error on top, but not at all clear that we have a linear drift (the systematic error).

Similarly, if we do coarse sampling and take, say, ten samples across this data range, we might easily identify the linear trend but miss the sinusoid entirely. Instead, we'd just think we had a line with some scatter.

The advantage for logarithmic spacing is that you sample both the high frequency signal (sinusoid) and the low frequency signal (linear drift) because our sampling intervals change. Therefore, if you spread ten samples (or whatever number) out over this data range you would identify both the high frequency sinusoidal function and the low frequency linear drift.

How to demonstrate this? There are lots of different ways. Perhaps the easiest is just to sample your ten (or whatever) samples under the three cases I just described and plot those results. Then it will be easy to see that only the logarithmic sampling allows you to determine the existence of both the high and low frequency signals.

Most answers were generally fine for the fine/medium/coarse/logarithmic question. One thing to look out for: if your logarithmic sampling doesn't give you a very clear signal, how could you increase the number

of points? You could reduce the (logarithmic) sampling frequency from, say, every 2^n (where n is a set of integers) to every $2^{0.8n}$, or whatever.

Next I asked you what the magnitudes of the three signals are. By this I simply wanted you to compare the relative amplitudes of the signals. The line goes from around zero to six across the 100 samples, so the amplitude is (roughly) around 3, say. The sinusoid has an amplitude of around 3 as well. And the scatter on the measurements (by eye) is maybe 0.5. In other words, the magnitude of the systematic error and the true signal are about the same, and the magnitude of the errors is smaller by maybe a factor of 10 or so. It's good news that the random error magnitude is smaller than the detected signals! That means that we can do some data processing and improve the fidelity of our result. Many people did not even answer this question, but it only counted for 2 points total, since most of you could have easily answered it from the work you did.

Now we can do some math. I asked you what is the standard deviation of the residuals (once you've corrected for the linear drift and the sinusoid). Through five minutes' worth of guess-and-check I found that the function is close to something like this:

$$\text{flux} = A\sin(t/13.5) + 0.02t \tag{1}$$

where A is the amplitude of the sine curve. (Your mileage may vary, of course – somewhat, but probably not a lot.) This is not meant to be an exact solution – just enough to get us started, i.e., to be close to the real solution.

I wrote some IDL code (see the web page) to find the best fit to the data. The way it works is that I found the above best-fit (by eye) function. Then I carried out a grid search over each of the four variables of interest (amplitude, phase, period, and slope), with 11 trials for each variable, centered on my best guess, for each of the dimensions. Why eleven? That gives me a reasonably good density of trials but the code runs in only five minutes on my underpowered laptop – it's a compromise. This gives me 11^4 (which is 14641) solutions. Note: this code produces some pretty neat-looking movies, so let me know if you want to watch it with me.

I calculate the residuals for each of the 14641 solutions. You can see in the code that I defined my best-fit criterion by looking for the minimum of the mean of the absolute value of the residuals. If the residuals were gaussian, which they frequently are not, then this value would indicate the narrowest gaussian that is achieved.

Finally, I *minimize* over all the variables. That is, I plot the minimum mean absolute residual for the eleven amplitude trials, summed over the other three variables. That is, for the first amplitude value, I ask what is the minimum mean absolute residual for all combinations of phase, period, and slope. Then I do it again for the second amplitude value, and the third, and so on. Then I do this same game by asking, for each value of phase, what is the minimum mean absolute residual, minimizing over the other three variables. I do the same for period and slope. In this way, I can derive what are the best-fit amplitude, phase, period, and slope.

There are lots of other ways to do this. One way would be to calculate the χ^2 value for each solution and find the best solution that way – similar, but slightly different.

Now, after I find the best solution according to my grid search ($A = 2.4$, phase of zero, $1/\omega$ of 13.5, and slope of 0.02) I calculate the residuals for this best-fit solution. This too is done in the code, and I plot a histogram of the residuals. It looks sort of gaussian, and we can measure the HWHM to be around 1.2, which means that σ is close to 1. (Make sure you know how I did that last step.) Why doesn't it look more gaussian? Maybe we don't have enough samples. Maybe our best-fit solution is not really that good, so our residuals are asymmetric, or broadened. Maybe the errors aren't really random!

Most people did fairly well on this previous part (residuals, best fit solutions). An important thing to remember is to show your results in a plot. I am not a machine and cannot interpret or understand several pages of data that are printed out. Even worse, if you don't show any data then I cannot judge that you have found a good (best) solution. Remember to show your results in plots.

Now we do some easier stuff. What is the probability of having a measurement that is 1σ from the true signal? That's 32%, as you well know. Right? You know how I got that, right? (I wasn't very clear here in what I was asking. The answers I am giving you are the probabilities of having values at that sigma or *more* from the mean. You might have answered what is the probability of having that sigma or less from the mean, in which case your answers are simply one minus my answers.) For 2σ, it's 5%. What about 2.5σ?

Remember that you solve this by looking up the values for the error function. So the probability of obtaining a value that is 2.5σ from the mean is

$$\frac{1}{\sqrt{2\pi}} \int_{-2.5}^{+2.5} e^{-z^2/2} dz. \tag{2}$$

Now you go find an online error function calculator or a lookup table, or similar. **Be careful!** Different people use slightly different definitions of the error function. Make sure you "calibrate" your error function calculator before you trust it. You know what the answer should be at 1σ and 2σ – test it! I found several online error function calculators that did *not* give the answers I expected. Appendix A of Taylor works correctly, as does Appendix A from Richard Lowry's online statistics book[1] that I have cited in class.

I find that the probability of having values beyond 2σ is 0.0455; for 2.5σ it's 0.0124; and for 3.2σ it's around 0.0013.

In 100 measurements, you should have around 32 measurements at 1σ (or greater); 5 measurements at 2σ (or greater); perhaps 1–2 at 2.5σ, and less than one (that is, most likely zero) at 3.2σ.

Now imagine that we've got 10,000,000 data points. At what level do we believe that an event is not a statistical fluctuation, but a true real measurement? In other words, at how many sigma do we expect the number of randomly occuring events to be less than 1, so that if we make a measurement there then we believe it to be a real significant measurement, and not just a random fluctuation? Make sure you understand how this question works – several of you seemed to miss the point here.

You can't calculate this, of course, because you need to use the error function here too. I used the one at Wolfram[2] (and yes I checked it first with 1σ) and found that 5.327σ gives about 1 part in 10^7 for the probability. To be conservative, then, you might require events to be 5.4σ or something in order to be sure that the detected event is not due to random fluctuations.

The probability of obtaining a 3σ or more event is 0.00270, so there will be around 27,000 of these in your data set of 10,000,000 measurements! Of course, approximately half of these would be beyond $+3\sigma$ and half would be beyond -3σ. The probability of obtaining a 4σ or more event is 0.00006, so there will be around 600 of these in your data set of 10,000,000 measurements. Therefore, the number of measurements you'd expect between three and four sigma would be about 26,400.

(2) The Adam Burgasser question has a lot of different parts to it, but fundamentally it's about using equation A3. I wrote a computer program (see the web page) that solves this equation. One nuance is to understand where 0.84 and 0.16 come from; hopefully you figured out that the integral of probability from negative infinity to $+1\sigma$ is 0.84, and the integral of probability from negative infinity to -1σ is 0.16. You could use other values here if you were interested in $\pm 2\sigma$, for instance.

The first thing I did with my computer code is to make sure I got the same answer as Burgasser: when $N = 10$ and $n = 2$ I get $\epsilon_b = 0.20^{+0.17}_{-0.07}$, as Adam did. What if $N = 10$ and $n = 1$? Then $\epsilon_b = 0.10^{+0.17}_{-0.03}$. What if $N = 10$ and $n = 0$? Then $\epsilon_b = 0.00^{+0.15}_{-0.00}$ or maybe, depending on your experiment, you might write $\epsilon_b < 0.15$. A lot of people found that the "lower bound" here was 0.016, which doesn't make sense: ϵ_b is zero, and the lower bound cannot be greater than (or less than) zero! You should think about the answers that you get numerically to make sure they make sense ....

[1] http://faculty.vassar.edu/lowry/webtext.html

[2] http://tinyurl.com/3p8w7v5

What if $N = 20$ and $n = 4$? Then $\epsilon_b = 0.20^{+0.12}_{-0.06}$. You can see that the error bars here are closer to equal than they were when $N = 10$, as expected. When N gets bigger, the shape of the curve should get more and more gaussian, which means error bars closer and closer to symmetric. With $N = 40$ and $n = 8$ we get $\epsilon_b = 0.20^{+0.08}_{-0.05}$. Something to watch out for here is how to write these upper and lower limits. Please note my syntax and formatting carefully – this is the correct way.

At what sample size does binomial become within 1% of gaussian? You guys had a lot of different ways to work this out. I thought a good way to do this would be to find N such that ϵ_b^U and ϵ_b^L are equal to within 1%. Because of the discrete steps in my code, I think that a precision of 1% is a little bit beyond what is calculable, but I did find that at $N \approx 128$ my error bars are symmetric to the precision that I can calculate things. Adam says N bigger than around 100 you should get symmetric error bars, and I'm not too far off from that. Your answers may differ slightly.

Many of you solved this by, as a first step, setting $n = 0$. I'm a little worried about this (and you should be too!). Will you ever have a gaussian – which ought to be symmetric – when $n = 0$? Several of you found answers this way, but I suspect you ran into the precision limit of the machine, which happened to give you more-or-less the correct answer. However, I think that the way I did it above, which allows $\epsilon_b = 0.2$, might be a better approach (among several approaches that people chose). Don't forget: Adam wrote in his paper that you should get the right answer when $N \approx 100$. That should give you a hint that if you are getting answers that are very different then something is not right.