Solutions to exercise #3 (assigned 5 October, due 26 October)

Overall I was very pleased with your performance on this exercise, again. I hope that you feel that you have learned a lot in this class this semester.

A few comments: (1) You guys have all learned to plot the data – great. But you need to also plot the *error bars*. How else can you tell whether your best-fit line makes sense? (2) A probability of 99.95% is *not* approximately equal to 100%. (3) r measures how well the two data sets are correlated, not how well the two data sets' errors are corrlated. (4) What does a `t test` test? Whether the *means* are the same (or not the same) – not whether the *data sets* are the same. And a `t test` and a K-S test test two different things. You might find that the means of two sets of data are similar but that the distributions of the data within the individual sets are different.

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. Most of the math and equations I used for this homework are presented in the IDL code on the exercise3 web page, so I mostly didn't repeat those expressions in this solution set. Please look at that IDL code to make sure you understand what I did.

(1) We start with the data in `exercise3.dat`. 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. Remember to plot the error bars too! Those error bars will be significant when you do the weighted least squares fitting.

The first thing I asked you to do was to calculate the unweighted and weighted least squares fits to these data (columns 1–4 only). The equations are in your notes, so this is not all that hard — you just have to do it. The only complication is what to do with w in the weighting formulae. I chose to make $1/w = \sigma_x^2$ for the x-only terms, $1/w = \sigma_y^2$ for the y-only terms, and $1/w = \sigma_x \sigma_y$ for the cross terms. I also made $1/w = \sigma_x \sigma_y$ for the terms that look like $\sum w$ (that is, with no x or y), though you could have chosen another approach. I found $b = [-0.75, 0.44]$ and $m = [5.1, 0.7]$ for [unweighted, weighted].

I found that my m and b for the unweighted fit were indistinguishable from those produced by IDL's canned routine. I plotted that line on the data (solid line in the first figure). My m and b for the weighted fit were quite different from the results from IDL's canned routine (IDL found $b = 3.6$ and $m = 3.7$). This is because IDL's canned routine only accepts errors on y, not on x – yet another reason to write your own equations and not rely on canned routines. I plotted those results over the data as well. My weighted fit is the dotted line on the first figure; the IDL weighted fit is the dashed line on the first figure. Why does my weighted fit look so "bad"? Well, my weighted least squares fit does pass through all the error bars, while putting the most weight (by far) on the first data point, which has the smallest (by far) error bars. In other words, the fit did exactly what it was supposed to!

I found a reduced chi-squared of 9.5 for the unweighted fit and 0.03 for the weighted fit. This means that we don't really have enough (good) data to derive a good weighted fit, but that the unweighted fit is not really all that good either.

I find a covariance for these data of 41. As a sanity check, I compared this value to $\sigma_x \sigma_y$, which is 47. I conclude that the data strongly covary (as you can tell by looking at the plot), but not as unity: the data do not perfectly covary.

I find $r = 0.98$. The probability of obtaining $r = 0.98$ with $N = 10$ for two uncorrelated variables is $<$0.05% (from Appendix C in Taylor), which means that the probability that these two data sets are correlated is $>$99.95%. This is a 3.5σ result (from Appendix A in Taylor).

(2) Now I plot all the data: x, y in black and x, z in yellow on the second figure.

I need to use the `t-test` to determine the significance of the difference between the means of the two data sets. I use the `t-test` equations to find that $t = 2.27$ and $df = 18$. Then I use a lookup table (for example, Appendix C of Richard Lowry's book) to find that the the level of significance of this result is between 0.05 and 0.02, so call it 0.04. Make sure you know how I did this! I chose a non-directional test – do you know why? This is about a 2.05σ result, but I would probably just call it 2σ.

Finally, I do the K-S test. I sort the data and then make the cumulative plots. (Third figure: x, y is diamonds, x, z is triangles.) I find (by eye) the greatest distance between the lines: it looks like $D \sim 0.45$ at $x = 18$ or so. Now I use a lookup table. There are many available; I gave you the one from Hawai'i earlier in the semester: when $n_1 = n_2 = 10$ the critical distance D_α for 2σ (which is $\alpha = 0.05$) is 0.7 and for 3σ ($\alpha = 0.01$) is 0.8. Therefore, my D is less than the critical D, so I cannot conclude that my samples are significantly different at the 2σ or 3σ levels. They are different at the $<2\sigma$ level.

How big of a sample size would I need for my $D = 0.45$ to imply samples that are significantly different? Let's assume that n_1 is always equal to n_2 (makes the math easier). I find that D_α, the critical value D at significance α, is $c(\alpha)\sqrt{2/n}$. The table tells me that $c(\alpha) = 1.36$ for 2σ, so $D_\alpha = 1.36\sqrt{2/n}$. Now if I set D_α to be equal to my D of 0.45 then I find $n = 18$. If my sample size were 18 measurements and the D were the same then those data sets would be significantly different at the 2σ level. At 3σ ($\alpha = 0.01$), $c(\alpha) = 1.63$ and $n = 26$. This makes sense – we'd need a bigger sample with the same D to have a more significant difference.

I have again used a canned IDL routine to check my results. That routine tells me that $D = 0.5$ (I was pretty close) and that the probability the two data sets are the same is 0.11. This means that the two data sets differ at around 1.6σ (Appendix A from Taylor again). This agrees with what we found above.