Homework 3: Sample Variation and Visualization

Biol 607

September 16, 2014

1) Problems from Whitlock and Schluter (5 points each)

Complete problems 10 and 17-18 on pg. 109-111. Use R where possible. Data sets (so you don't have to type things in) are available at http://www.zoology.ubc.ca/~whitlock/ABD/teaching/datasets.html.

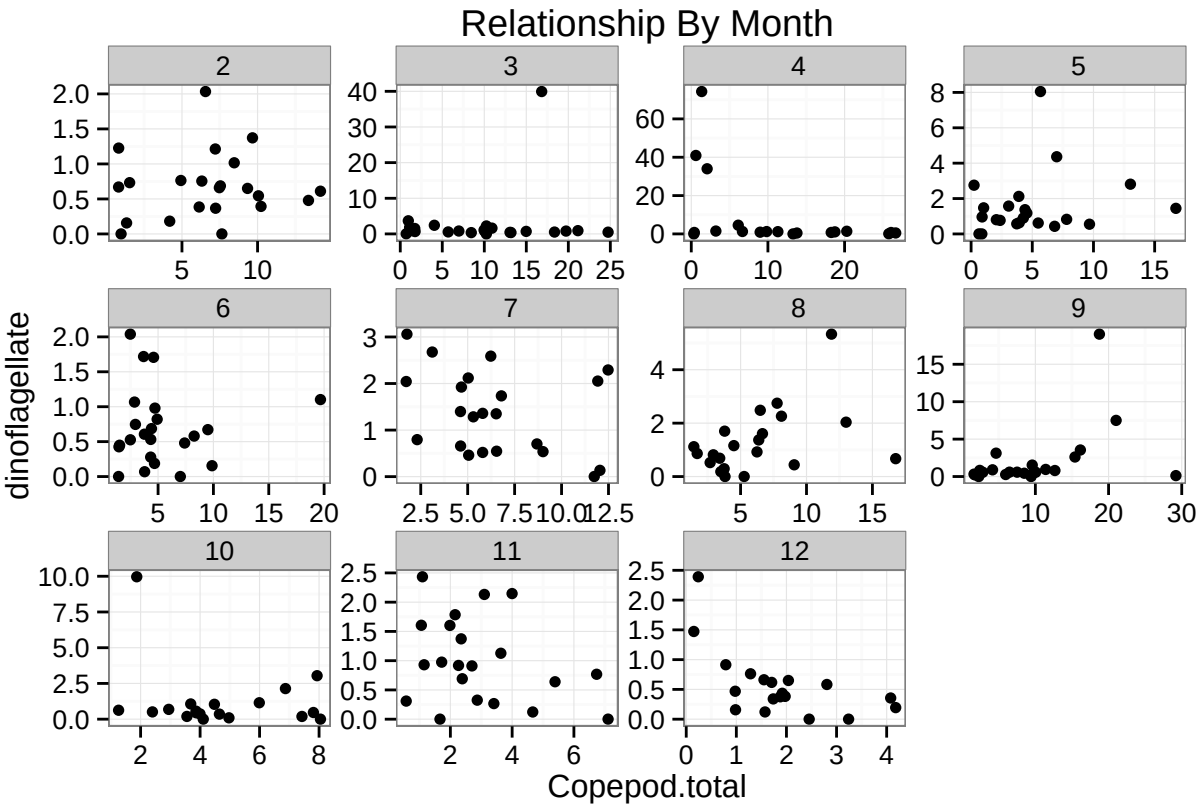
2) 95% Confidence Intervals (10 points) In class, we discussed how a 95% confidence interval around an estimate of a parameter tells you that, there is a 95% chance of the true mean falling within the 95% CI that you have calculated. This takes some mental gymnastics to absorb the true implication. One implication of that definition is that, were you to go out and sample a population 100 times, the 95% CI will overlap the true value of the estimated parameter. It **does not** mean that you have a probability distribution for potential values of that parameter. Those are Bayesian Credible Intervals, and we'll talk about those later in the semester. \ The 95 times out of 100 bit is an interesting concept, though, and once you actually see it in practice, it's a bit easier to understand what a 95% CI (or 99%CI or 50%CI, etc) is actually telling you. \ So, I'd like you to generate a population of 1000 individuals, drawn from a normal distribution with mean 30 and SD 5. Take 100 samples from this population of sample size 10. For each sample, calculate the 95% CI of the mean. How many times does your 95% CI overlap the *true* mean? Try re-running the simulation a few times to feel more comfortable, as the answer will often not be exactly what you expect (this is a simulation, after all!)

Also show this graphically. You may need to either use functions like `lines` and `segments` or explore a few additional geoms, depending on what you use to plot.

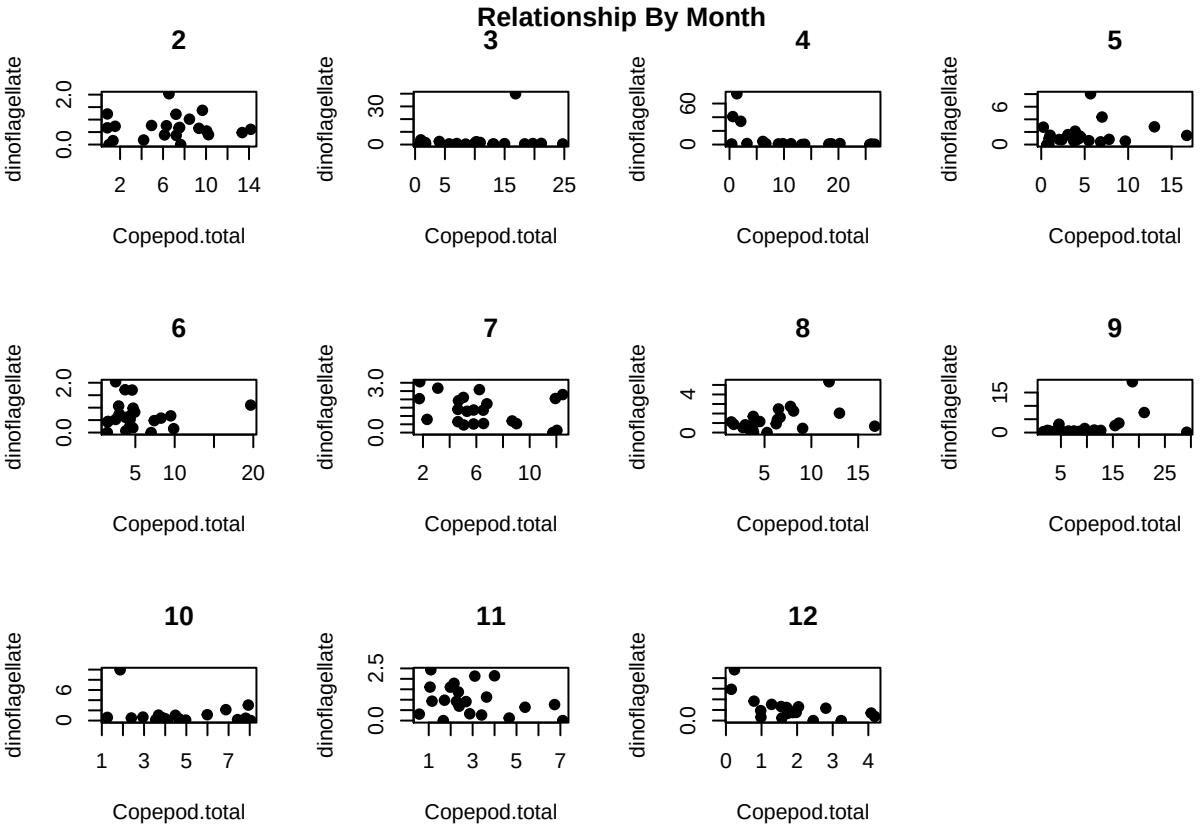
3) Loops and Plotting

3.1) Load the Data (5 points) We'll be working with the Lake Baikal Plankton Data from the data visualization lecture. To learn more about it, the instrumentation, etc., see http://knb.ecoinformatics.org/knb/metacat/nceas.290/nceas. Load it in. Screen it for any bad data or obvious outliers. Should they be eliminated? Why or why not?

3.2 (10 points) One of the really interesting ways to look at the relationships in this data is to split them by month. This lets us see trends within months so that we can directly compare processes between years. For example, we can look at the Dinoflagellate-Copepod relationship as follows.



Reproduce this plot using both ggplot2 and the basic R graphing package. The former is straightforward. The latter should look something like:



3.3 Bootstrapped Sample Estimates (10 points) As we discussed in class, the re-sampling based approach to assessing error in parameter estimates can be incredibly simple and powerful. In particular, it can be quite powerful in the case of variables that have asymmetric confidence intervals. To estimate asymmetric confidence intervals, one re-samples their data as usual to calculate a test statistic, but then looks at the quantiles or percentiles of the test statistic to determine the range of values in which 95% of their sample estimates fall.

Let's look at how this works for medians.

(a) Calculate the naive bootstrapped standard error and 95% confidence intervals for the median of the values of diatom in the data. Use 5000 bootstrapped replicates.

(b) Compare this naive estimate to the percentile confidence intervals. Take a look at the arguments for the function quantile. Are they different? Why or why not?

(c) Look at the bootstrap function in the bootstrap package. Can you use it to get the 95% CIs in two lines of code?