Homework 8 – Replication, Statistical Power, and Type M Error

Imagine that you want to test how a warming climate affects primary production in grasslands. You will perform an *in situ* warming experiment, in which you use ceramic heaters to warm 2 x 2 meter experimental plots by 4ºC, continuously for two years. Control plots are unwarmed, and the treatment and control plots will be randomly assigned. Biomass production is measured at the plot scale using appropriate methods.

(1) You have limited time and money, which means that the number of replicate treatment and control plots is limited. Before designing and performing the study, let's consider how the number of replicates will affect our *statistical power*: our ability to detect an effect of warming.

Based on previous studies in similar environments, we can get a sense for what the *effect size* might be: in other words, the true magnitude of the warming effect. We can also get a sense for how variable productivity is between plots in the same treatment, for reasons outside of our control (i.e., the residual noise). Assume that the true treatment mean is 460, and the true control mean is 415 (the numbers are aboveground net primary production (ANPP) - grams per square meter per year). Assume that the standard deviation of the variation between plots is 110. These numbers are based on real data synthesized in the attached paper by Lemoine et al.

Start by assuming three replicates each of the treatment and control plots. Perform 1000 simulations, where each time you draw three treatment ANPP values and three control ANPP values, using the appropriate means and standard deviation, from a normal distribution. For each random draw, fit a linear model testing for a treatment effect. Save the p-value from an F-test for the treatment effect [you can extract this with Anova(model)$P[1]]. Also save the model coefficient that quantifies the difference between the treatment and control groups. [You can extract this with coef(model)]. At the end you should have 1000 p-values and 1000 coefficient values.

What proportion of the p-values are less than 0.05? This is your statistical power, under this design, effect size, and residual noise. Now repeat this whole process using different numbers of replicates: 5, 10, 20, 50, 100. Plot the statistical power vs. the number of replicates. How much replication do you need to achieve a power of 0.25? This means that when there is a real treatment effect, you will detect it (only) 25% of the time.

(2) Now let's use the simulation results to answer a different question. For situations where statistical power is low, a treatment effect will only be 'significant' if it is quite large, and *this may cause the treatment effect to be exaggerated by chance*. This has been termed Type M error, where the M stands for *magnitude* (see the attached paper).

For each of the simulations you performed above, you saved the model coefficients. So you should have 1000 coefficients for each level of replication (3, 5, 10, 20, 50, 100).

Take those coefficients, and for each level of replication calculate *the mean of the coefficients using only models where $p < 0.05$*. This is simulating the following process: if you perform the experiment, and $p > 0.05$, you report 'no effect', but if $p < 0.05$ you report 'significant effect'. We want to know if the reported significant effects are biased. How does the mean of the significant coefficients change as the number of replicates increases? Recall that because this is a simulation, we know the true value of the treatment difference: it's 461 - 415 = 46. How much larger is the simulated value from the significant experiments, compared to the true value? This is the type M error. What are the potential implications for our understanding of climate change, if most warming experiments have low power?