BIOL343 – Assignment #8

Fake It Till You Make It

Assigned Monday 23 March
Due Saturday 28 March 1159pm

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." Sir Ronald A Fisher

It seems to me that a lot of biologists start thinking about statistical analysis after they've collected their data. The goal of this assignment is to get you thinking about how an experiment might be analyzed and what the results might be while you're designing the experiment – before any data are collected.

In class, we have learned how to simulate data from a factorial experiment.

- Set the number of factors and the number of levels of those factors.
- Set a total sample size.
- Distribute that total sample size (total_n) among the various treatment combinations ("cells") of the experiment.
- Make the predictor variables using nested loops.
- Set the grand mean for the response variable.
- Set the within-cell SD of the response variable (residual variation, or "sres" in the simulation we made in class).
- Think about the effects of each factor on the response variable and set those "effects".
- Think about how those factors might interact and set the details of the interaction (the extent of nonadditivity).
- Using these effect sizes, draw individual data points from a random normal distribution with a mean equal to the grand mean plus the main effects and their interactions and a standard deviation equal to the residual SD (sres).
- Analyze your new simulated data to see if you reliably recover the effects and interactions that you set.

The simulation we worked on together can be set within a broader framework (i.e. a larger loop) to explore how various aspects of experimental design or the statistical properties of the biological system you're studying can affect your ability to recover the effects that you know are "in the data". The frequency with which you recover these effects is "statistical power", also known as $1-\beta$, where β is the type II error rate. We set the type I error rate (α, usually at 0.05) but not β.

Two parameters have a major influence on β, the sample size (total_n) and the residual variation (sres), so for this assignment you are tasked with exploring the effects of ONE of these on statistical power.

I'd like you to set up an experiment with two factors. You can decide how many levels of each you want. Make the effect sizes realistic, that is a 5–20% change in the response between the most

extreme factor levels. Your simulations should also include a reasonable interaction between the two factors. The interaction effects should be about the same general magnitude as the sex and pH effects. Describe the interaction in the text of your R notebook. Remember the rule with respect to the magnitude and sign of these effects.

Then embed the data simulation and analysis within two nested loops. The inner loop runs through the experiment "nreps" times to yield the distribution of *P* values under one set of parameters. You should set nreps at 1000. The outer loops runs through several prospective values of total_n or sres (pick one). Whether you pick sres or total_n to vary, you must have at least 10 reasonable values in your vector of possibilities. Of course, you're welcome to have many more; might make for more interesting results and your computer doesn't mind.

To add some biological realism to the simulation I want you to pick a trait in a particular species and find data suggesting the likely grand mean and sres.

For example, the simulation we played with in lecture imagined body size in zebra fish (*Danio rerio*) as a response variable. I set gmean = 10 and sres = 4 but I just made up those values. Based on data from the literature, however, the body mass of a 15-20 month old zebra fish is 0.62 g with SD = 0.09 g (Gilbert et al. 2013. Zebrafish (*Danio rerio*) as a model for the study of aging and exercise: Physical ability and trainability decrease with age. *Experimental Gerontology* (doi: 10.1016/j.exger.2013.11.013). Please provide the full reference and a doi for the source that provides the mean and base SD of your response variable. And no, you can't use zebra fish.

So, here's what Hana and I are expecting from you:

1. A text section at the beginning of your R notebook describing the experiment and its goals.
2. Code setting up the design of the experiment.
3. Code specifying the statistical properties of the biological system you are simulating plus the effect sizes for both factors and their interaction. Describe the interaction in words in the text below this code.
4. A full reference plus doi for a published study on which you based your values of gmean and sres.
5. Code for a loop making the part of the results data.frame containing the predictor variables.
6. Code for a loop drawing the data for the response variable from the appropriate normal distribution (the response variable should be added to the results data.frame).
7. Code for an inner loop running through a large number of replicates for each parameter set (nreps).
8. Code for an outer loop running through at least 10 values of the parameter (total_n or sres) that you are examining.
9. For each batch of simulated data, you will run the analysis of your experimental factors using lm() and collect the *P* values from applying the Anova() function to the resulting lm() object.
10. With these *P* values, you should calculate, for each value of total_n or sres, the statistical power (frequency with which you reject the null hypothesis when it is false) for each term in the model.
11. Make a publication-quality figure (with appropriate caption) illustrating your results. You should plot the lines for all three effects (sex, pH, interaction) on the same graph and annotate the graph so we know which line is which.

12. In a text section in your R notebook, interpret your results and describe how varying sample size (total_n) or trait variation (sres) affects frequency of statistical errors.
13. Your R notebook should be nicely organized (you can follow this list of expectations as an organizational structure), so that we can easily find everything asked for above and effectively evaluate your work.

You will submit an **PDF version** of your .html R notebook document called "StudentNumber_A8.pdf", where the file name starts with your student number.

Please make sure that your PDF file is **complete** and as nicely formatted and organized as possible, paying close attention to all the formatting guidelines and tips provided for previous assignments.

You should upload your file to the Assignment #8 OnQ dropbox by Saturday 28 March at 1159pm.

If you run into difficulties you can ask questions during my online office hours (Wednesdays 130-230pm) and online tutorials (Thursdays 930–1020am and 1030–1120am). Good luck!!