Advanced Biological Statistics

Independent Analysis 1 F2017

Assignment: Your task is to analyze the associated data set (zfish_diet.tsv) as if you generated it and are presenting the results for publication. You should use R for your work, and produce the appropriate analyses and figures to test distributional assumptions and to support your conclusions. You should encapsulate these findings in a results section that is modeled after one you would find in a journal such as Science or Nature, or PLoS Biology. You should not present every intermediate figure and analysis! Remember, the goal is to concisely and accurately present your results in figures, tables and statistics, each of which is correctly labeled and annotated. You may have additional figures that you can present as supplementary material if necessary, but be sure that they are referenced as such. In addition, include your final .Rmd as supplementary material.

Due: Submit your work to me **via Canvas** by the end of the day (midnight) on **Friday October 27th**. I encourage you to work with other members of class to help one another, but I expect each of you to construct and run all of the scripts yourself, and to write the results section on your own.

The Study:

Say that you performed a study to determine whether or not feeding zebrafish with enriched brine shrimp nauplii (treated with a particular supplement) affects mean zebrafish size. You are also wondering whether size might be affected differentially in terms of overall length of the fish, as determined by it's standard length in millimeters (mm), as compared to body mass, as measured by wet weight in milligrams (mg).

You sample 200 newly hatched zebrafish from the same clutch and randomly assign each to either an enriched shrimp diet or an unenriched shrimp ("control") diet. Assume that you have carried out your experiment carefully and not confounded diet with any other variables that may affect size. After two months you measure the standard length and weight of all 200 fish. The tab-delimited file "zfish_diet_IA.tsv" contains your data in four columns: "Individual," "Diet," "SL," and "Weight".

Here are some guidelines to help you think about what you should include in your results section:

1. State formally the null and alternative hypotheses relevant to your question for each of the continuous variables with respect to the categorical variable. Keep in mind you have no *a priori* expectations for how (or even whether) the altered diet will affect your fish.
2. Read the file into R and perform a thorough exploration (visualization and summarization) of the data. Produce relevant summary statistics and figures, including a single figure with dependent variable distributions for 1. the entire data set, 2. the control group, and 3. the enriched group. Remember, only a subset of these figures should be included in the main results section, others in the supplementary, and a few may be just for you.
3. Evaluate whether assumptions for a hypothesis test using the student's *t* distribution are met by the data, and actually perform a t-test using R. Make a clear, formal statement of the t-test results. Remember, you have two continuous variables. **Note** - *there is one mistake that will pop out pretty quickly during exploratory data analysis (EDA) that was a data entry error that you'll need to fix before you finish your analyses. You can just change the text file.*
4. Perform the same hypothesis test, but use a non-parametric approach (resampling in this case). Think carefully about what parameter you need to estimate, and make sure you re-sample it adequately (at least 1000 times). Plot the distribution of the resampled parameter, and report its mean and standard deviation. Most importantly, compute a 95% confidence interval for this parameter, and interpret your original point estimate in light of it and the hypothesis test. Also, produce a figure showing the confidence interval and your original point estimate on top of the distribution of the resampled parameter.
5. Evaluate the strength of the relationship between the two continuous variables using a simple linear regression assuming normality. Be sure to formally state your null and alternative hypotheses, and to report the relevant hypothesis test results. In addition, perform an examination of the residuals to determine whether the parametric assumptions of the linear model are met.
6. Draw conclusions from your results that answer the questions that you raised in your study!