Advanced Biological Statistics

Homework 1 F2017

Assignment: Work through the following problems, writing a **complete R markdown file** for the questions below that contains all of the appropriate components. Annotate your file to include such things as variable assignments, functions with arguments, as well as comments to explain each step. You can place all of your annotated R scripts into a single .Rmd file that I can just render.

Due: Submit your work **via Canvas** by the end of the day (midnight) on **October 19th**. I encourage you to work with other members of the class to help one another, but I expect each of you to eventually construct and run all of the scripts yourself. This is the best way to learn statistics and R, so don't cheat yourself!

Problem 1. Download and examine the data set 'HW_RNAseq.csv'. Notice that this is a RNA-Seq data set from two populations of threespine stickleback fish. There are numerous different variables, including measures of gene expression for ten different genes. Write an R script for **Exploratory Data Analysis**, including:

- List the variables including their characteristics (i.e. categorical, continuous, response, independent)
- Examine each continuous variable to determine its distribution using histograms, and put these into a single, multi-panel figure.
- Transform all original continuous variables to z-scores and plot the distribution for just two genes.
- Create boxplots of the original data and the z-scores for these two genes, split by population, treatment and sex. Use colors in your figures. Again, make a single, multi-panel figure.
- Calculate descriptive statistics (mean, variance, standard deviation) for each population for each variable, for each treatment for each variable, and for each sex for each variable. Compile these into a table.

Problem 2. Write an R script to calculate the **Standard Error (SE)** and **95% confidence interval (CI)** of the **mean** for each gene from above using both the parametric approach as well as the resampling approach that you learned in class. Assemble your results for all genes in a single table. Comment on whether the two approaches give you approximately the same result, and interpret your observation. In addition, modify the resampling approach to calculate the **SE of the variance** for all 10 genes, and add those to the table as well.

Problem 3. Pretend that you are sampling the number of Douglas Fir trees in each of 1000 plots around Mt. Pisgah. The data are as follows.

Number of trees observed in a plot	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Number of Plots with this many trees	74	149	228	181	169	84	49	24	19	12	9	4

Plot these raw data and determine how they are distributed. Using this sample, and the equation from your book for this particular sampling distribution, calculate the **mean number of trees per plot** and the **variance in the number of trees** per plot. How do these results compare, and why is this question significant?

Now, create a maximum likelihood function for the range of parameter values, estimate the MLE and confidence intervals.