MAR536: Biological Statistics II

R Laboratory Exercise 5

February 15, 2023

Open a new R markdown file. Save it. (name it lastname_lab5.Rmd or something similar)
At the top of the script, add comments with your name and lab 5. Work in pairs or individually. Submit your Rmd and any other files via myCourses before lab next week.

Lab exercise 1 - permutation test

Use the Laengelmavesi data in `../data/Laengelmavesi2.csv`

a. Obtain the data for just the lengths of perch and bream.

b. Plot the distribution of lengths for both species, and calculate the mean lengths for both species.

c. Conduct a permutation test to assess whether the difference in mean length between bream and perch is statistically clear.

d. Plot the distribution for the test statistic under the null hypothesis of no difference in length, and indicate the true value for the test statistic relative to the two-tailed 95th percent of the null hypothesis distribution.

e. What are your conclusions?

Lab exercise 2 - Bootstrapping

`hake.csv` contains abundance data for silver hake from tows in the 2015 NMFS spring bottom trawl survey.
a. Produce 5,000 bootstrapped estimates for the mean abundance per tow based on case resampling (5000 samples).
b. Compare the standard deviation of the bootstrapped estimates of the mean to the standard error of the mean from the original sample.
c. Compute an approximate 95 percent confidence interval for the mean based on the bootstrap, assuming normality. Compare this to the interval based on percentiles of the bootstrap sampling distribution.
d. *BONUS* Plot how the bootstrap confidence interval for the mean changes with the number of bootstrap samples. (100, 500, 1000, 2000, 5000, 10000)

Lab exercise 3, k-fold cross validation

Using the `ISLR::Wage` data set, evaluate the predictive ability of models for wages.
a. Define a unique random number seed. Use 10-fold cross validation to estimate the test error rate for models fitting a polynomial of age of order 2, 3, 4, 5, and 6.
b. Conduct the validation 20 times for each polynomial. Plot the distribution (use boxplots) for the validation test error rate as a function of the degree of polynomial. Based on the results, what order polynomial would you use?
c. Use 5-fold cross-validation to compare the performance of models that include combinations of:
- a polynomial of age
- education level
- race
- industry

e. What model would you choose based on the test error rates?
f. *BONUS* How does the model chosen by 5-fold CV compare to that from using AIC as a model selection tool?