Conducting Analysis of Variance (ANOVA) in R

Dr. Kevin Geyer
Assistant Professor of Biology
Young Harris College
11/29/2019

Focus:

This lesson offers an introduction to ANOVA, including 1) how this statistical test can be differentiated from others and 2) a step-by-step guide to conducting and interpreting ANOVA results in R, including assumption testing and post-hoc analysis.

Overview:

This lesson focuses on recognizing situations in which ANOVA can be used and the necessary steps for conducting such a test. Students will review the different types of variables commonly collected by biologists (continuous vs. categorical; dependent vs. independent) to discern scenarios is which ANOVA is appropriate. Next, students follow a step-by-step guide to conducting and interpreting an ANOVA test using the 'iris' dataset made available in base R. A variety of multiple choice and command-response questions are used to check student comprehension throughout. The skills learned within should serve as a foundation for additional tests utilizing other variable combinations and may require testing of other assumptions and/or data transformations.

Learning Objectives:

Learning objectives for this lesson include:

a) Gaining familiarity with basic R coding, including the syntax for functions (e.g., function.name()) and accessing R help files both in R Studio and online.
b) Discerning the differences in data variable types, and how experimental context can influence these.
c) Conducting an ANOVA test with particular attention paid to:
    - Accessing and interpreting ANOVA tables and p-values
    - Checking assumptions of residual normality and homoscedasticity
    - Transformation of data to meet assumptions
    - Post-hoc analysis of results

Lesson Sequence:

This lesson includes the following sequence of activities:

1) A discussion of different variable types, specifically dependent vs. independent and continuous vs. categorical. Students are asked to view the 'iris' dataframe made available with the course as an example dataset. Multiple choice questions presented to check comprehension about variable types.

2) Upon recognizing a combination of variables suitable for ANOVA (sepal length vs. iris species), students will plot these candidate variables in order to make predictions as to whether sepal length is significantly different between the three species in the dataset.
3) Next, students will conduct the ANOVA (aov()), then save this result in order to explore it more thoroughly (summary()), specifically isolating the p-value.
4) A definition of p-values is given for students to critically think about the results they've just generated ("Will you accept or reject the null hypothesis?")
5) Two major assumptions of the ANOVA test are now introduced.
    a. Homoscedasticity
        i. A definition is given and students are led to conduct a test of this assumption (bartlett.test()).
        ii. The p-value for this test is <0.05, indicating violation of the assumption (a heteroscedastic dataset). We will return to resolve this problem later.
    b. Normality of residuals
        i. A definition is given and students are led to conduct a test of this assumption (shapiro.test()).
        ii. The p-value for this test is >0.05, indicating the assumption is met.
    c. Data transformation
        i. In order to resolve the heteroscedasticity, students are instructed to log transform sepal length and repeat the Bartlett test (which should now pass).
        ii. Given this, the ANOVA test is repeated with a log transformed sepal length. The new test results are assessed and any earlier conclusions revised (this won't be necessary, as the p-value remains <<< 0.05).
6) A post-hoc analysis (TukeyHSD()) is introduced in order to test how many pairwise significant differences exist between iris species.
7) Finally, students are asked to debrief from this lesson by writing an outline of the necessary steps for conducting an ANOVA test. This can be turned in for assessment and used by students in future class activities where ANOVA is to be used.

Pre-lesson Activities:

I recommend beginning this activity with the entire class working collectively. This allows the instructor to verify that all students have been able to open swirl and access the course. It also provides an opportunity for the instructor to review some basic coding terminology, how to navigate swirl (particularly how to leave the course environment using ESC or bye()), and expectations for the debrief activity. This activity would also be most effective when prefaced with a broader discussion about scientific data and the objectives of statistical analysis.

Post-lesson Activities:

Students are assigned at the end of the lesson to reflect on the steps of conducting an ANOVA test by writing their own outline. This can be collected for assessment and used by students in future activities where they are expected to perform an ANOVA test more independently.