**R coding tutorial – how to do a one-way analysis of variance in R**

**Introduction**

An analysis of variance (ANOVA) is a useful and widely used statistical tool to evaluate statistically significant differences between different levels of a single factor (one-way ANOVA) or different levels of multiple factors (two-way ANOVA). An ANOVA determines if the means between each group for whatever variable being measured are significantly different. This improves upon the t-test, which compares means of two groups, but can only determine if there is a difference between two groups; ANOVAs allow comparison between multiple groups at once. This tutorial will focus on one-way ANOVAs, due to the character of the example dataset.

ANOVAs should be done if there are multiple sampling levels within an overall factor, such as varying habitat types that are being compared for significant differences in a biodiversity metric like species richness or individual abundances. To run an ANOVA, R requires a numeric vector (the values that the ANOVA will calculate the mean on; these are the measured data) and a factor vector (the different levels that are being compared to determine if there is a difference). This means that the categorical data will have to be in factor format; if it is not already, R can transform numeric data into factor data using the function “factor(numeric data)”. When R runs the ANOVA, it will return several statistics, including values used to calculate different steps in the ANOVA, and the p-value, which is the most important statistic when it comes to interpreting the ANOVA. The automatic α-value that R uses is 0.05, or a confidence interval of 0.95, meaning that the results will be considered statistically significant if they have a less than 5% chance of occurring if the factor levels have no effect on the mean. Therefore, if the p-value is less than 0.05, the results will be considered significant, as they have less than a 5% chance of occurring randomly. R can use another α-value if it is specified in the code.

While the ANOVA itself does not tell the user which levels are significantly different from each other, just that they are, there are posthoc tests that tell the user which levels differ. One of the most commonly used ones is the Tukey’s posthoc test. When a Tukey’s test is run in R, it returns the p-value for each pairwise comparison; p-values lower than 0.05 indicate that there is a significant difference between the means of that pair. Both ANOVA and Tukey’s tests are built into R base package.

ANOVAs have several assumptions that have to be met, as the data for an ANOVA has to be parametric, meaning that it is sample data from a population that follows a probability distribution. There are two assumptions that can be tested for using R: the assumption of normality, and the assumption of homoscedasticity, meaning that the variances of the populations tested are roughly equal. There are several options to test normality using R, including a Shapiro-Wilk test, boxplots, qqnorm plots, and histograms. A Shapiro-Wilk test tests against the assumption of normality; it returns a p-value that, if below 0.05, rejects the assumption of normality. Boxplots, qqnorm plots, and histograms are all graphical representations of the normality of the data. To test for homoscedasticity, R can run a Bartlett’s test, which returns a p-value like the Shapiro-Wilk test; a p-value lower than 0.05 indicates that the assumption of homoscedasticity should be rejected. Normality tests in R are run on the numeric vector, while the homoscedasticity test is run using the numeric vector and the factor vector, illustrated below in the example code. Note that an ANOVA is generally robust to deviations from the normality assumption, especially if the sample size is large, and fairly robust to deviations from the assumption of homoscedasticity, but this is contingent on the samples being the same size.

The example dataset used for this tutorial is suitable for a one-way ANOVA. There are nine groups, the unifying variable of urban park type, which compose the factor vector. Each group has ten replicates that are random samples of the population. There are two measured variable columns, species richness and total individual abundances per sample. For the example, the possible significant difference in species richness between each group is being considered. Note that in the results of the test, the data is shown to be nonnormally distributed and homoscedastic: the ANOVA is still proceeded with as it is an example. The ANOVA shows that there is a significant difference and the Tukey’s test shows which groups differ significantly from each other: fourteen pairwise comparisons in all significantly differ.

**Annotated ANOVA code**

#This line of code clears the global environment before you start working

rm(list=ls(all=TRUE))

#setwd() sets the directory in which the files you want are located

#getwd() tells you what directory you are working in at the moment

#dir() tells you all the files that are in your directory

setwd("C:/Users/Christine/Documents/Biology 607")

getwd()

dir()

#First step in doing an ANOVA is to input your dataset

#read.csv will read in a csv file, header=TRUE tells R to expect that

#the first lines of the dataset are column names

#The head() function shows the first few lines of every column,

#so it's useful to make sure that your dataset was imported entirely

ants.anova<-read.csv("tutorial example dataset (csv).csv",header=TRUE)

head(ants.anova)

#To make sure the data meets the assumptions for an ANOVA, we do a normality

#test, which is the shapiro.test() function, and a homoscedasticity test,

#which is the bartlett.test() function

#Both of these are conducted on the numeric vector

#To tell R to conduct the test on one specific column in the dataset, use the $ to

#separate the dataset name and the column name

shapiro.test(ants.anova$sp\_richness)

bartlett.test(ants.anova$sp\_richness~ants.anova$park)

#To run the actual ANOVA, we use the aov() function

#The structure of the aov() function is (numeric vector~factor vector)

#The results of the ANOVA are stored in the object named ant.aov

#To get the statistical summary output for the ANOVA, the summary() function is used

ant.aov<-aov(sp\_richness~park,ants.anova)

ant.aov

summary(ant.aov)

#To run the posthoc test, Tukey's test, TukeyHSD() function is used

#The structure of TukeyHSD() is (ANOVA object, factor being compared, confidence level)

#The results of the Tukey's test are stored in the object named ant.posthoc

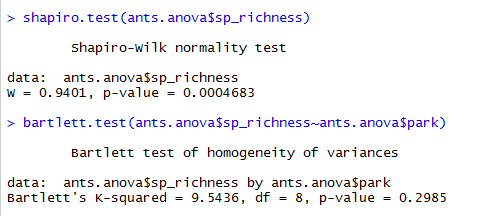
#To get the statistical summary output for the Tukey's test, run that object alone

ant.posthoc<-TukeyHSD(ant.aov, 'park', conf.level=0.95)

ant.posthoc

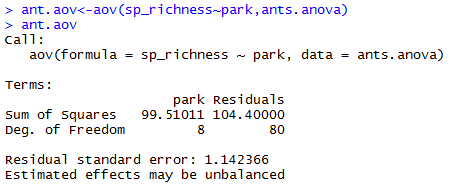
**Example images of the final statistical outputs for each test**

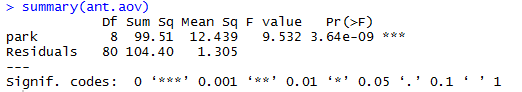
Results of Shapiro-Wilk test and Bartlett’s test:



The p-value for this can be found under “p-value” in the summary table.

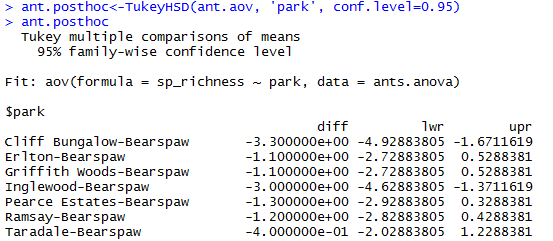
Results of ANOVA:

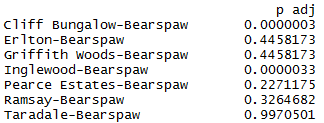




The p-value for the ANOVA is found under “Pr(>F)” where F refers to the ratio of variances.

Results of Tukey’s test





The p-value for each pairwise comparison can be found in the column “p adj”.