ANOVA TESTS

NGAIRA PETER

2024-05-11

##ONE WAY ANOVA  
#ANOVA IS USED TO TEST 2 OR MORE STATISTICAL MEANS TO PROVE WHETHER THEY STATISTICALLY SIGNIFICANT  
#Analysis of variance compares the between variance and the within different group variance  
#If the between variance if larger than the within variance, the group variance is declared to be different   
#It is tested using the ratio Variance between / variance within.  
  
# Importing online data from penguins  
# Loading data  
require(palmerpenguins)

## Loading required package: palmerpenguins

## Warning: package 'palmerpenguins' was built under R version 4.3.3

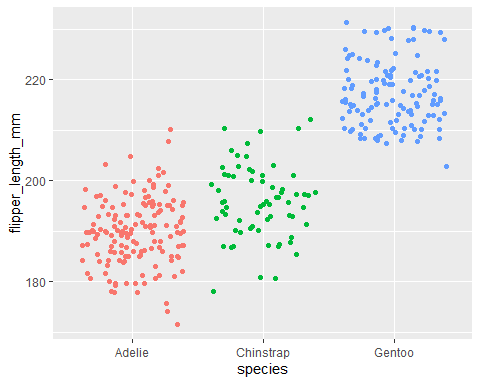
require(tidyverse)  
  
#focusing only on the flipper length and the species using pipes  
anova<- penguins %>%   
 select(species, flipper\_length\_mm)  
  
# Saving the selected data set into the directory  
write\_excel\_csv(anova, "anova.xlsx")  
  
# Summary of the selected data set  
summary(anova)

## species flipper\_length\_mm  
## Adelie :152 Min. :172.0   
## Chinstrap: 68 1st Qu.:190.0   
## Gentoo :124 Median :197.0   
## Mean :200.9   
## 3rd Qu.:213.0   
## Max. :231.0   
## NA's :2

#Flipper length varies from 172 to 231 mm, with a mean of 200.9 mm. There are respectively 152, 68 and 124 penguins of the species Adelie, Chinstrap and Gentoo.

# A ggplot of Flipper\_length\_mm dispersion over Species  
library(ggplot2)  
ggplot(anova)+  
aes(x=species, y= flipper\_length\_mm, color=species)+  
 geom\_jitter()+  
 theme(legend.position = "none")

## Warning: Removed 2 rows containing missing values or values outside the scale range  
## (`geom\_point()`).



# Here, the factor is the species variable which contains 3 modalities or groups (Adelie, Chinstrap and Gentoo).

#ANOVA CALCULATIONS IN PREPARATION TO TEST NORMALITY ASSUMPTION  
res\_aov <- aov(flipper\_length\_mm ~ species, data= anova)

# Visual check of Normality  
#Combine plots  
par(mfrow= c(1,2))  
  
#Histogram of the residuals  
hist(res\_aov$residuals)  
  
# QQ plot of the residuals  
library(car)

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

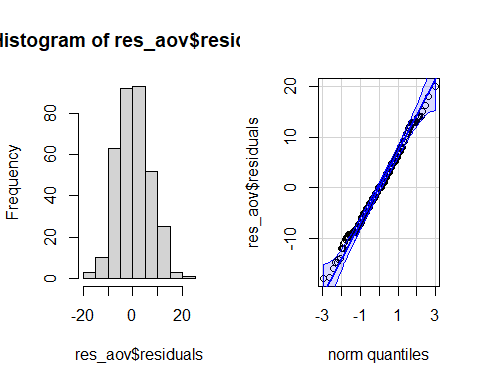
## Warning: package 'carData' was built under R version 4.3.3

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

qqPlot(res\_aov$residuals,id= FALSE)



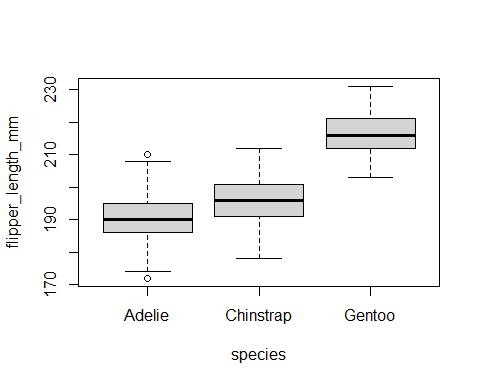
# From the histogram and QQ-plot above, we can already see that the normality assumption seems to be met. Indeed, the histogram roughly form a bell curve, indicating that the residuals follow a normal distribution. Furthermore, points in the QQ-plots roughly follow the straight line and most of them are within the confidence bands, also indicating that residuals follow approximately a normal distribution.

# Shapiro test of normality  
# Assumptions   
#1. H0: data come from a normal distribution  
#2. H1: data do not come from a normal distribution  
  
shapiro.test(res\_aov$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: res\_aov$residuals  
## W = 0.99452, p-value = 0.2609

#P-value of the Shapiro-Wilk test on the residuals is larger than the usual significance level of α=5%, so we do not reject the hypothesis that residuals follow a normal distribution (p-value = 0.261).  
# residuals are the distance between the actual value of Y and the mean value of Y for a specific value of X

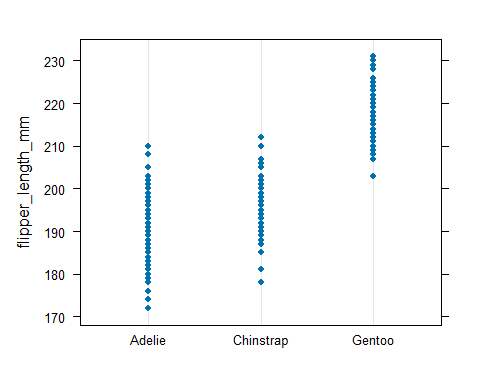
## Equality of variances  
#Assuming residuals follow a normal distribution, Equality is tetsted using either boxplot or dotplot  
#BOXPLOT  
boxplot(flipper\_length\_mm ~ species, data=anova)



#Dotplot  
require(lattice)

## Loading required package: lattice

dotplot(flipper\_length\_mm ~ species, data= anova)



#Both the boxplot and the dotplot show a similar variance for the different species. In the boxplot, this can be seen by the fact that the boxes and the whiskers have a comparable size for all species.  
  
#There are a couple of outliers as shown by the points outside the whiskers, but this does not change the fact that the dispersion is more or less the same between the different species.

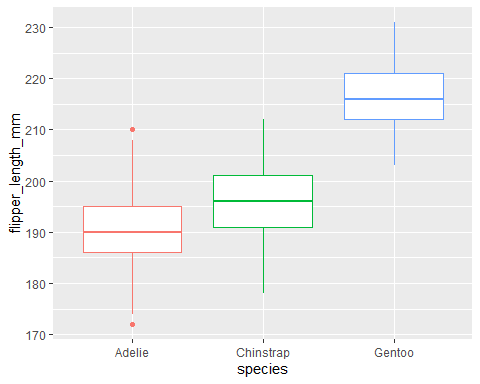
#Like the normality assumption, if you feel that the visual approach is not sufficient, you can formally test for equality of the variances with a Levene’s or Bartlett’s test. Notice that the Levene’s test is less sensitive to departures from normal distribution than the Bartlett’s test.  
  
#The null and alternative hypothesis for both tests are:  
  
#1. H\_0: variances are equal  
#2. H\_1: at least one variance is different.  
  
#Levennes test  
leveneTest(flipper\_length\_mm ~ species, data=anova)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.3306 0.7188  
## 339

#The p-value being larger than the significance level of 0.05, we do not reject the null hypothesis, so we cannot reject the hypothesis that variances are equal between species (p-value = 0.719).  
  
#This result is also in line with the visual approach, so the homogeneity of variances is met both visually and formally.

# Outlier detection using bxplots   
par(mfrow=c(1,2))  
library(ggplot2)  
ggplot(anova)+ aes(x= species, y= flipper\_length\_mm, color=species)+  
 geom\_boxplot()+ theme(legend.position = 'none')

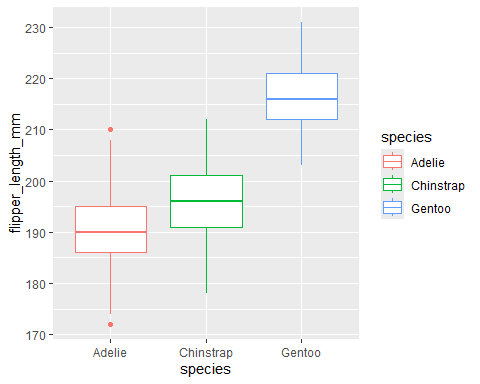
## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



#There is one outlier in the group Adelie, as defined by the interquartile range criterion. This point is, however, not seen as a significant outlier so we can assume that the assumption of no significant outliers is met.

#Preliminaries analysis  
#A good practice before actually performing the ANOVA in R is to visualize the data in relation to the research question. The best way to do so is to draw and compare boxplots of the quantitative variable flipper\_length\_mm for each species.  
  
ggplot(anova)+  
 aes(x= species, y = flipper\_length\_mm, color= species)+  
 geom\_boxplot()+  
 theme(legend.position = "right")

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



#The boxplots above show that, at least for our sample, penguins of the species Gentoo seem to have the biggest flipper, and Adelie species the smallest flipper.  
  
#Besides a boxplot for each species, it is also a good practice to compute some descriptive statistics such as the mean and standard deviation by species.  
  
library(dplyr)  
group\_by(anova, species) %>%   
 summarise(  
 mean= mean(flipper\_length\_mm, na.rm = TRUE),  
 sd= sd(flipper\_length\_mm, na.rm = TRUE)  
 )

## # A tibble: 3 × 3  
## species mean sd  
## <fct> <dbl> <dbl>  
## 1 Adelie 190. 6.54  
## 2 Chinstrap 196. 7.13  
## 3 Gentoo 217. 6.48

#Mean is also the lowest for Adelie and highest for Gentoo

## “Is the length of the flippers different between the 3 species of penguins?”.  
# This can be achieved through performance of several tests   
# 1. One way ANOVA test  
oneway.test(flipper\_length\_mm ~ species, data = anova, var.equal = TRUE)

##   
## One-way analysis of means  
##   
## data: flipper\_length\_mm and species  
## F = 594.8, num df = 2, denom df = 339, p-value < 2.2e-16

# 2. With the summary() and aov() functions:  
res\_aov <- aov(flipper\_length\_mm ~ species, data = anova)  
summary(res\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## species 2 52473 26237 594.8 <2e-16 \*\*\*  
## Residuals 339 14953 44   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 2 observations deleted due to missingness

#As you can see from the two outputs above, the test statistic (F = in the first method and F value in the second one) and the p-value (p-value in the first method and Pr(>F) in the second one) are exactly the same for both methods, which means that in case of equal variances, results and conclusions will be unchanged.  
  
#The advantage of the first method is that it is easy to switch from the ANOVA (used when variances are equal) to the Welch ANOVA (used when variances are unequal). This can be done by replacing var.equal = TRUE by var.equal = FALSE, as presented below:

#Welch ANOVA  
oneway.test(flipper\_length\_mm ~ species, data = anova, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: flipper\_length\_mm and species  
## F = 614.01, num df = 2.00, denom df = 172.76, p-value < 2.2e-16

##Interpretations of ANOVA results  
#Given that the p-value is smaller than 0.05, we reject the null hypothesis, so we reject the hypothesis that all means are equal. Therefore, we can conclude that at least one species is different than the others in terms of flippers length (p-value < 2.2e-16).

require(report)

## Loading required package: report

## Warning: package 'report' was built under R version 4.3.3

report(res\_aov)

## The ANOVA (formula: flipper\_length\_mm ~ species) suggests that:  
##   
## - The main effect of species is statistically significant and large (F(2, 339)  
## = 594.80, p < .001; Eta2 = 0.78, 95% CI [0.75, 1.00])  
##   
## Effect sizes were labelled following Field's (2013) recommendations.