Summarising the Iris dataset using stratification

Jeevan

7 March 2019

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

attach(iris)  
names(iris)

## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"   
## [5] "Species"

View(iris)  
iris %>%  
 filter(Species == 'versicolor') %>%  
 summarise(mean(Petal.Length,na.rm = TRUE))

## Warning: package 'bindrcpp' was built under R version 3.5.1

## mean(Petal.Length, na.rm = TRUE)  
## 1 4.26

iris %>%  
 filter(Species == 'versicolor') %>%  
 summarise(mean(Sepal.Length,na.rm = TRUE))

## mean(Sepal.Length, na.rm = TRUE)  
## 1 5.936

iris %>%  
 filter(Petal.Width>1) %>%  
 group\_by(Species) %>%  
 summarise(mean(Sepal.Width,na.rm = TRUE))

## # A tibble: 2 x 2  
## Species `mean(Sepal.Width, na.rm = TRUE)`  
## <fct> <dbl>  
## 1 versicolor 2.83  
## 2 virginica 2.97

strata1 = filter(iris,Species == 'setosa')  
strata2 = filter(iris,Species == 'versicolor')  
strata3 = filter(iris,Species == 'virginica')  
s1 = round(nrow(strata1)/nrow(iris)\*49)  
s2 = round(nrow(strata2)/nrow(iris)\*50)  
s3 = round(nrow(strata3)/nrow(iris)\*50)  
s1+s2+s3

## [1] 50

cor.test(strata1$Petal.Length,strata1$Sepal.Length)

##   
## Pearson's product-moment correlation  
##   
## data: strata1$Petal.Length and strata1$Sepal.Length  
## t = 1.9209, df = 48, p-value = 0.0607  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01206954 0.50776233  
## sample estimates:  
## cor   
## 0.2671758

cor.test(strata2$Petal.Length,strata2$Sepal.Length)

##   
## Pearson's product-moment correlation  
##   
## data: strata2$Petal.Length and strata2$Sepal.Length  
## t = 7.9538, df = 48, p-value = 2.586e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6020680 0.8532995  
## sample estimates:  
## cor   
## 0.754049

cor.test(strata3$Petal.Length,strata3$Sepal.Length)

##   
## Pearson's product-moment correlation  
##   
## data: strata3$Petal.Length and strata3$Sepal.Length  
## t = 11.901, df = 48, p-value = 6.298e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7714542 0.9210172  
## sample estimates:  
## cor   
## 0.8642247

CONCLUSION From the above code, we see that the dplyr package has been attached to perform the method of stratification. Using the names function - we see the different columns that are there in the dataset. The view function allows to see the full dataset. We find that the mean petal length and sepal length for the species veriscolor is 4.26 and 5.936 respectively. The mean sepal width whose petal width is more than 1 for each species is displayed separately. The setosa species does not have sepal width greater than 1 and hence it does not display in the table with versicolor and virginica. The dataset has been startified on the basis of species i.e. into 3 species which are setosa, versicolor & virginica. Samples have been taken using the nrow command for each species and the total number of samples have been added to 50. Using Pearson’s product-correlation coefficient, the correlation between Petal Length & Sepal length for the species setosa = 0.267158. The correlation for the species versicolor = 0.754049 The correlation for the species virginica = 0.8642247