One-way ANOVA

ESTIMATING WHICH LOCATIONS HAVE SIGNIFICANTLY DIFFERENT AVERAGE RAINFALL

Pranav Gopalkrishna, 1940223

Semester 5, Practical 2

2021-08-11

Table of Contents

[Setting work directory 1](#_Toc79610320)

[Data set 1](#_Toc79610321)

[Formatting and simplifying the data set 2](#_Toc79610322)

[Creating the ANOVA model 3](#_Toc79610323)

[Checking the linear regression model 3](#_Toc79610324)

[Performing ANOVA 3](#_Toc79610325)

[Hypotheses 3](#_Toc79610326)

[Test 4](#_Toc79610327)

[Result interpretation 4](#_Toc79610328)

[Post-hoc analysis (pairwise comparison of means) 4](#_Toc79610329)

[Significantly different means 5](#_Toc79610330)

[Final notes on potential inaccuracies 6](#_Toc79610331)

# Setting work directory

setwd("~/Documents/Study/computerScience/programming/r/data/")

# Data set

This data set contains several fields of meteorological data from five different locations in Australia, measured across a span of around 8 years and 6 months, starting from 2008-12-01 and ending at 2017-06-25. For this assignment, we are concerned about only two fields: location and daily rainfall.

Our aim here is to figure out whether the average daily rainfall in these five locations is statistically significantly different from each other or not, as measured from from 2008-12-01 to 2017-06-25, and if yes, which locations have significantly different daily rainfall levels. Our significance level will be 0.05 or 5%, meaning that if the test statistic derived from the difference between the means of two locations falls in the range of the 95% most probable values, we do not consider the differences to be significantly different.

# We need only the location and rainfall fields, and date fields to establish the time frame.  
myData = read.csv("weatherAustralia.csv")[c(1, 2, 5)]  
# Converting dates from strings to dates.  
myData$Date = as.Date(myData$Date, "%d/%m/%y")  
head(myData)

## Date Location Rainfall  
## 1 2008-12-01 Albury 0.6  
## 2 2008-12-02 Albury 0.0  
## 3 2008-12-03 Albury 0.0  
## 4 2008-12-04 Albury 0.0  
## 5 2008-12-05 Albury 1.0  
## 6 2008-12-06 Albury 0.2

summary(myData)

## Date Location Rainfall   
## Min. :2008-12-01 Albury :3040 Min. : 0.000   
## 1st Qu.:2011-01-18 BadgerysCreek:3009 1st Qu.: 0.000   
## Median :2013-05-10 Cobar :3009 Median : 0.000   
## Mean :2013-04-03 CoffsHarbour :3009 Mean : 2.386   
## 3rd Qu.:2015-06-03 Moree :3009 3rd Qu.: 0.200   
## Max. :2017-06-25 Max. :371.000   
## NA's :342

Before moving ahead, it is important to note that the location is the factor, and the rainfall level depends on it (to some degree). The different locations are the different levels of this factor.

# Different levels of the factor "location"  
levels = unique(myData$Location)  
i = 1  
for(level in levels)  
{  
 print(paste(i, ":", level))  
 i = i + 1  
}

## [1] "1 : Albury"  
## [1] "2 : BadgerysCreek"  
## [1] "3 : Cobar"  
## [1] "4 : CoffsHarbour"  
## [1] "5 : Moree"

## Formatting and simplifying the data set

# Removing NA values  
max = length(myData$Location)  
sum(is.na(myData$Location))

## [1] 0

sum(is.na(myData$Rainfall))

## [1] 342

x = y = c()  
for(i in c(1:max))  
{  
 if(!is.na(myData$Location[i]) & !is.na(myData$Rainfall[i]))  
 {  
 x = c(x, myData$Location[i]);  
 y = c(y, myData$Rainfall[i]);  
 }  
}  
  
# Converting x to factor form

**Why independent variable is to be changed to factor type**

Firstly, locations are supposed to be factor type, as they are distinct nominal values.

Secondly, for the levels of the independent variable (i.e. treatment or variety) to be treated as levels, you must convert the values to factor values i.e. non-numeric. This is vital, since aov (analysis of variance) function calls the lm (linear model) function, and if lm encounters the independent variables values as numeric, it will create a single linear regression model between the independent variable and the response.

Making the independent variable values as factors makes each lm call make a separate linear regression model for each level i.e. each value of the independent variable i.e. each class.

x = as.factor(x)  
  
# Creating the new data frame  
myData = data.frame(x, y)

# Creating the ANOVA model

## Checking the linear regression model

To test how the linear regression model will be made for each level of the factor...  
lm(y ~ x)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Coefficients:  
## (Intercept) x2 x3 x4 x5   
## 1.9141 0.2790 -0.7868 3.1474 -0.2839

## Performing ANOVA

### Hypotheses

**H0:** *The mean daily rainfall for each each location is equal.*

**H1:** *The mean daily rainfall for at least two locations are unequal to each other.*

Note that neither hypothesis will be proved exactly. Rather, equality is concluded by statistically insignificant differences in means, and inequality is concluded by statistically significant differences in at least two means.

### Test

aovModel = aov(y ~ x, myData)  
summary(aovModel)

## Df Sum Sq Mean Sq F value Pr(>F)   
## x 4 28282 7071 76.83 <2e-16 \*\*\*  
## Residuals 14729 1355443 92   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

An alternate and equivalent way to perform ANOVA through aov, given factor x and response y, is as follows...

linearRegressionModel = lm(y ~ x)

aovModel = aov(linearRegressionModel)

### Result interpretation

Pr(>F) is the p-value associated with the calculated F statistic. This is well below 0.05, the significance level we have chosen.

**Hence, we reject the null hypothesis.**

This means location is significant enough of a factor that the average daily rainfall of at least two locations is statistically significantly different, as measured from 2008-12-01 to 2017-06-25. To find out which locations display this difference, we will perform the post-hoc analysis.

I assume that Pr(>F) refers to the probability that the calculated F-ratio estimated for the population (which is not known in this case, but is estimated for based on the sample values), is lesser than the table F-ratio for the degrees of freedom (4, 1472). Note that calculated value being lesser than table value means the null hypothesis will be accepted. Hence, Pr(>F) refers to the probability of the null hypothesis being accepted, or the alternate hypothesis being rejected. Here, this probability is extremely small, and definitely below 0.05, the significance level, hence we do not accept the null hypothesis.

# Post-hoc analysis (pairwise comparison of means)

Tukey test allows to find pairs of factor levels whose means are significantly different from each other, comparing all possible pairs of means with a t-test like method. Tukey test is a single-step multiple comparison procedure and statistical test. It is a post-hoc analysis method, what means that it is used in conjunction with an ANOVA.

In R, the multcompView allows to run the Tukey test thanks to the TukeyHSD() function. We are using the following arguments...

TukeyHSD(x = aovModel, conf.level = 0.95)

The variable aovModel contains the ANOVA model. Note that the independent variable in the ANOVA model must be of factor type i.e. non-numeric. We have already handled that aspect. The argument conf.level refers to the confidence interval.

#install.packages("multcompView")  
library(multcompView)  
tukey = TukeyHSD(x = aovModel, conf.level = 0.95)  
tukey

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = y ~ x, data = myData)  
##   
## $x  
## diff lwr upr p adj  
## 2-1 0.2789862 -0.4002676 0.9582400 0.7957830  
## 3-1 -0.7868057 -1.4625931 -0.1110183 0.0129881  
## 4-1 3.1473819 2.4695874 3.8251764 0.0000000  
## 5-1 -0.2839117 -0.9676154 0.3997921 0.7891229  
## 3-2 -1.0657919 -1.7463333 -0.3852504 0.0001889  
## 4-2 2.8683957 2.1858611 3.5509303 0.0000000  
## 5-2 -0.5628979 -1.2513010 0.1255053 0.1683718  
## 4-3 3.9341875 3.2551026 4.6132724 0.0000000  
## 5-3 0.5028940 -0.1820890 1.1878770 0.2645331  
## 5-4 -3.4312936 -4.1182568 -2.7443303 0.0000000

**QUICK NOTE ON P-VALUE**

The adjusted p (p adj) value is the smallest significance level at which a particular comparison will be declared statistically significant as part of the multiple comparison testing. In general, p-value for a test statistic refers to the probability of it occurring for a sample. To calculate the probability, we assume that the population distribution from which the sample is drawn is transformable to (i.e. follows the same kind of distribution as) the standard distribution that the test statistic follows. In this case, that distribution is the t-distribution.

For example, for the 1st pair 2-1, p adj = 0.7957830, meaning if the difference between the means of 2 and 1 is to be considered statistically significant, then you must choose a significance level of at least 0.7957830.

However, our chosen significance level is 0.05 or 5%, way below the required minimum. Hence, with 0.05 or 5% significance, the difference is not considered statistically significant i.e. the difference is within the range from which the values (differences between means here) are expected to occur 95% of the time (confidence interval is 95%, hence the range of ‘non-outlier’ values, as determined by us, covers 95% of the most probably occurring values).

## Significantly different means

First note the index-to-location name mapping...

## [1] "1 : Albury"  
## [1] "2 : BadgerysCreek"  
## [1] "3 : Cobar"  
## [1] "4 : CoffsHarbour"  
## [1] "5 : Moree"

From the results of the Tukey test, noting the adjusted p-values for each pair of locations, we see that the following location pairs have significantly different means from each other, as measured from 2008-12-01 to 2017-06-25...

3-1 i.e. Cobar - Albury  
4-1 i.e. CoffsHarbour - Albury  
3-2 i.e. Cobar - BadgerysCreek  
4-2 i.e. CoffsHarbour - BadgerysCreek  
4-3 i.e. CoffsHarbour - Cobar  
5-4 i.e. Moree - CoffsHarbour

The rest of the locations have statistically insignificant differences in average daily rainfall level, as measured from 2008-12-01 to 2017-06-25.

# Final notes on potential inaccuracies

The data is historic. Hence our estimates reflect the climate than recent weather patterns, assuming the climate has not changed decisively over the period of time that the data was collected. The climate may have undergone changes during the 8 and a half years for which this data has been measured, possibly leading to misleading conclusions.