*CRD ANOVA*

ESTIMATING WHICH SCREEN SIZES HAVE SIGNIFICANTLY DIFFERENT AVERAGE PRICES

Pranav Gopalkrishna, 1940223

2021-09-01

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Including necessary libraries and setting work directory**

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.2

##   
## 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

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

# Data set

This dataset is for basic data analysis, and contains basic information about various different computer models, including RAM size, screen width and price.

I want to find out how a computer’s screen size can affect the computer’s price on average.

Since RAM size is generally a significant factor in the price of the computer, we will keep the data relatively homogenous by focusing on only one RAM size, 16GB. This is done because CRD is most effective for homogenous data, and not very accurate otherwise.

Hence, **x = screen, y = price**, given that the RAM size is 16GB.

myData = read.csv("computersBasics.csv")[c(1, 2, 5, 6)]  
# Only including relevant columns.  
myData$screen = as.factor(myData$screen)  
# Converted screen size to factor type for future purposes.

## Brief look at data

head(myData)

## X price ram screen  
## 1 1 1499 4 14  
## 2 2 1795 2 14  
## 3 3 1595 4 15  
## 4 4 1849 8 14  
## 5 5 3295 16 14  
## 6 6 3695 16 14

## **Data summary**

summary(myData)

## X price ram screen   
## Min. : 1 Min. : 949 Min. : 2.000 14:3661   
## 1st Qu.:1566 1st Qu.:1794 1st Qu.: 4.000 15:1992   
## Median :3130 Median :2144 Median : 8.000 17: 606   
## Mean :3130 Mean :2220 Mean : 8.287   
## 3rd Qu.:4694 3rd Qu.:2595 3rd Qu.: 8.000   
## Max. :6259 Max. :5399 Max. :32.000

## Filtering out

myData = filter(myData, myData$ram == 16)[c(4, 2)]

# (Only including the required columns)

myData$screen = as.factor(myData$screen)

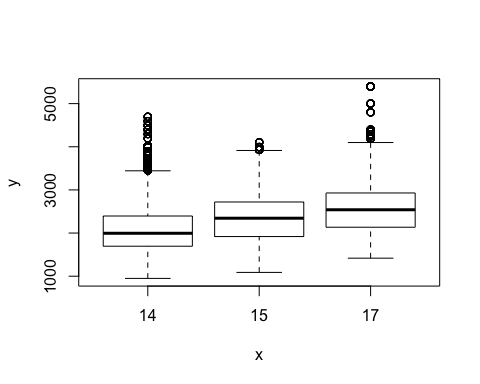
# Converting x values to factor values, as they should be.

head(myData)

## X price ram screen  
## 1 5 3295 16 14  
## 2 6 3695 16 14  
## 3 23 2895 16 14  
## 4 29 3095 16 14  
## 5 45 3495 16 14  
## 6 95 2995 16 14

## Visualising the distribution of data based on screen size.

boxplot(myData$price~myData$screen)



From the plot above, we can see that the means for each screen size distinct, but also somewhat close, at least visually. The same goes for the distribution of their values. Also, judging by the ranges, they follow a somewhat normal distribution each, with respect to prices.

# Hypotheses for ANOVA

**H0**: Mean prices are equal for all screen sizes.

**H1**: Mean prices are unequal for at least two screen sizes.

*Our level of significance will be 0.05.*

# Creating the ANOVA model

aovModel = aov(price ~ screen, myData)  
summary(aovModel)

## Df Sum Sq Mean Sq F value Pr(>F)   
## screen 2 2.341e+09 1.171e+09 3824 <2e-16 \*\*\*  
## Residuals 75105 2.299e+10 3.061e+05   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Pr(>F) is the p-value to consider here. It is the probability of the null hypothesis being true. It is lower than 0.05, the significance level.

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

Hence, at least two screen sizes have singificantly different mean prices.

# Post hoc analysis

Since the null hypothesis is rejected, we perform a post hoc analysis to find out exactly which pairs of screen sizes have significantly different mean prices.

#install.packages("multcompView")  
library(multcompView)   
tukey = TukeyHSD(x = aovModel, conf.level = 0.95)  
tukey = as.data.frame(tukey$x[,'p adj'])  
names(tukey)[1] = paste("p adj")  
tukey

## p adj  
## 15-14 0  
## 17-14 0  
## 17-15 0

From these adjusted p-values, we can say that, for 16GB RAM computers, every pair of screen sizes have significantly different mean prices, for a 0.05 significance level. This means that we may conclude (in general) that

1. Screen size is a significant factor in computer price, for 16GB RAM size.
2. All screen sizes differ from each other significantly in terms of price.