Second Question: ANOVA

Arnau Abella

10/03/2020

Anova Test

a) Generate three population using your own algorithm.

In order to generate the three normal populations I used the following Haskell script

```haskell
stdDev :: Double
stdDev = 1.0

main = do
  IO.withFile "normal.csv" IO.WriteMode $
    \handle -> do
      vss <- traverse (normalV 10000) [0.0, 0.0, 10.0]
      forM_ vss $ \vs ->
        let bs = Csv.encode [GV.toList vs]
        in LBS.hPut handle bs
  where
    normalV n mean =
      withSystemRandom $
        \(gen::GenST s) -> normalVector mean stdDev gen n :: ST s (UV.Vector Double)

normalVector :: (PrimMonad m, Vector v Double)
             => Double             -- ^ Mean
             -> Double             -- ^ Standard deviation
             -> Gen (PrimState m)
             -> Int                -- ^ vector length
             -> m (v Double)
normalVector mean std gen n =
  GV.replicateM n (MWCD.normal mean std gen)

standardVector :: (PrimMonad m, Vector v Double)
             => Gen (PrimState m)
             -> Int                -- ^ vector length
             -> m (v Double)
standardVector = normalVector 0.0 1.0
```

b) Analyze using an ANOVA if these three populations are different (or not) depending on the parameter selected.

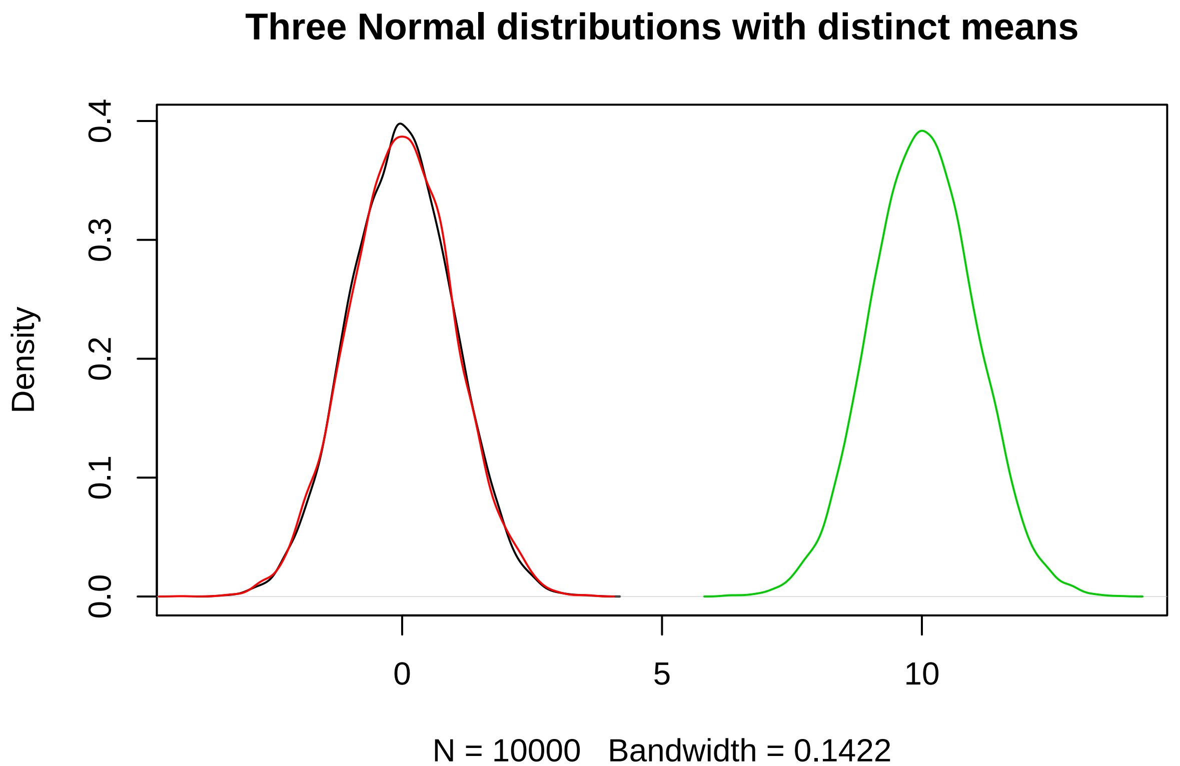
```r
# 30,000 values in total.
v <- sapply(read.csv( paste(root, 'normal.csv', sep='/'), header = FALSE, sep = ","), as.numeric)
v1 <- v[1, ] # 10,000 values
v2 <- v[2, ] # 10,000 values
v3 <- v[3, ] # 10,000 values

plot(density(v1),xlim=c(-4,14),main="Three Normal distributions with distinct means")
lines(density(v2),col=2)
lines(density(v3),col=3)
```

```r
v1n=data.frame(x1=v1, x2="v1")
v2n=data.frame(x1=v2, x2="v2")
v3n=data.frame(x1=v3, x2="v3")

library(RcmdrMisc)
```

```
## Loading required package: car
## Loading required package: carData
## Loading required package: sandwich
```

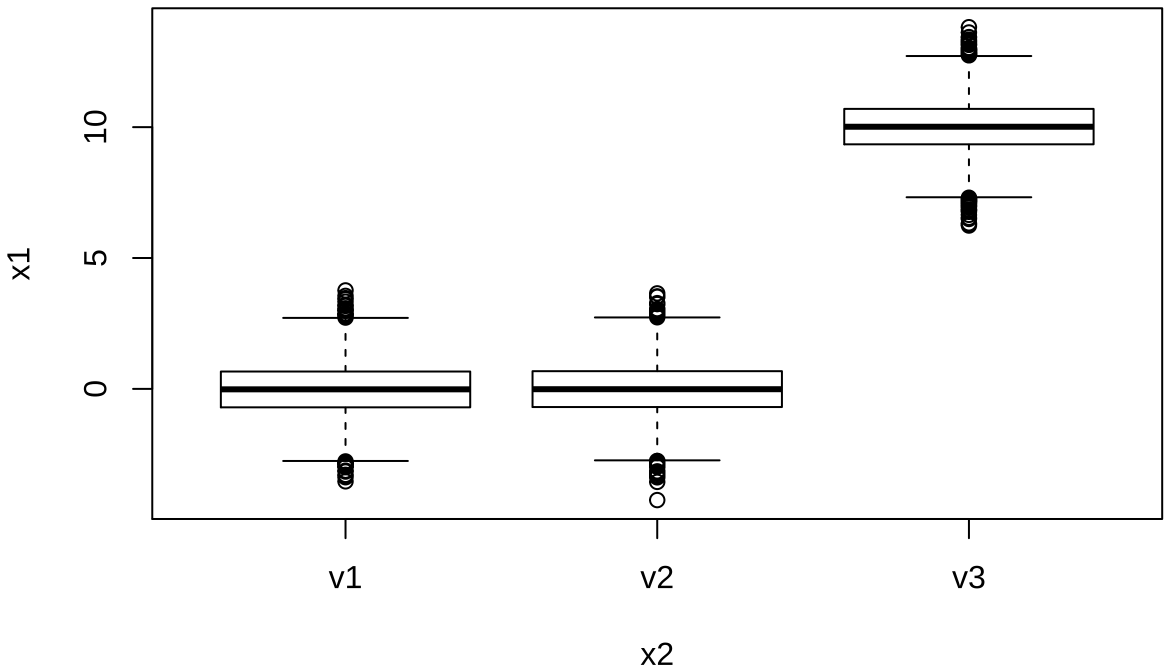



```r
# We create a single data frame
data=mergeRows(v1n, v2n, common.only=FALSE)
data=mergeRows(as.data.frame(data), v3n, common.only=FALSE)

AnovaModel.1 <- aov(x1 ~ x2, data=data)
summary(AnovaModel.1) # Pr(>F) = p-value
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## x2              2 671726  335863  334459 <2e-16 ***
## Residuals   29997  30123       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Boxplot(x1~x2,data=data,id=FALSE)
```



From the output of the ANOVA test, we can see that the PR(>F) is smaller than the p-value so we **can** refuse the null hypothesis that there is no significant difference between means of the different groups.

Red and White Wine Quality

We want to analyze if in both (type or quality) affects some properties of the wine. After combining the two datasets (one for red wines and one for white wines), you should create two variables: "type" that identifies if the wine is red or white, and wine quality categorized in three groups: <5 (low), 5-6(medium) and >6 (high). Once you complete preprocessing steps, please answers the following questions applying appropriate statistical techniques:

```r
red  <-read.csv2(paste(root, 'winequality-red.csv', sep='/'), dec=".") # 1599 x 12
white<-read.csv2(paste(root, 'winequality-white.csv', sep='/'), dec=".") # 4898 x 12

# Combine the rows
winequal<-rbind(red,white)

# Categorical Variable: type
winequal$type<-as.factor(rep(c(1,2),c(nrow(red),nrow(white))))
levels(winequal$type)<-c("red","white")
summary(winequal$type)
```

```
##   red white
##  1599  4898
```

```r
# Categorical Variable: category (low, medium, high)
winequal$category<- cut(winequal$quality,c(1,5,6,10))
summary(winequal$category)
```

```
##  (1,5]  (5,6] (6,10]
##   2384   2836   1277
```

Before answer the questions, we are going to check if the assumptions of ANOVA are fulfilled for each numerical variable.

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
anova1 <- aov(alcohol ~ quality, data=winequal)
```

Independent obs.

```r
# Durbin Watson, Ho = autocorrelation of the disturbances is 0.
dwtest(anova1, alternative ="two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  anova1
## DW = 1.488, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

Normality

```r
#Shapiro test (Normality)
# shapiro.test(residuals(anova1))
```

Homogeneity

```r
#Breusch Pagan test (Variance equality)
bptest(anova1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  anova1
## BP = 101.75, df = 1, p-value < 2.2e-16
```

```r
# leveneTest(alcohol~quality, data=winequal)
```

a) Which of the chemical properties influence the quality of the wines?

```r
for (i in 1:11){
  print(colnames(winequal)[i])
  print(summary(aov(winequal[,i]~category,data=winequal)))
}
```

```
## [1] "fixed.acidity"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## category       2     57  28.455   17.01 4.27e-08 ***
## Residuals   6494  10861   1.672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "volatile.acidity"
##               Df Sum Sq Mean Sq F value Pr(>F)
## category       2  13.09   6.547   260.9 <2e-16 ***
## Residuals   6494 162.98   0.025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "citric.acid"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## category       2   0.89  0.4472   21.31 5.98e-10 ***
## Residuals   6494 136.28  0.0210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "residual.sugar"
##               Df Sum Sq Mean Sq F value   Pr(>F)
## category       2    614  307.11   13.62 1.25e-06 ***
## Residuals   6494 146434   22.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "chlorides"
##               Df Sum Sq Mean Sq F value Pr(>F)
## category       2  0.345 0.17233   146.7 <2e-16 ***
## Residuals   6494  7.628 0.00117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "free.sulfur.dioxide"
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## category       2    4122  2060.8   6.553 0.00144 **
## Residuals   6494 2042386   314.5
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "total.sulfur.dioxide"
##                Df   Sum Sq Mean Sq F value   Pr(>F)
## category       2    73818   36909   11.59 9.44e-06 ***
## Residuals   6494 20679084    3184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "density"
##                Df  Sum Sq  Mean Sq F value Pr(>F)
## category       2 0.00629 0.003144   391.7 <2e-16 ***
## Residuals   6494 0.05212 0.000008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "pH"
##                Df Sum Sq Mean Sq F value Pr(>F)
## category       2   0.15 0.07318   2.832  0.059 .
## Residuals   6494 167.79 0.02584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "sulphates"
##                Df Sum Sq Mean Sq F value  Pr(>F)
## category       2   0.25 0.12739   5.761 0.00316 **
## Residuals   6494 143.59 0.02211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "alcohol"
##                Df Sum Sq Mean Sq F value Pr(>F)
## category       2   2069  1034.7   936.9 <2e-16 ***
## Residuals   6494   7172     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All properties except for the pH affect the quality of the wine.

b) Which of the chemical properties are related with type of the wines ?

```r
for (i in 1:11){
  print(colnames(winequal)[i])
  print(summary(aov(winequal[,i]~type,data=winequal)))
}
```

```
## [1] "fixed.acidity"
##                Df Sum Sq Mean Sq F value Pr(>F)
## type           1   2587  2586.7    2017 <2e-16 ***
## Residuals   6495   8331     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "volatile.acidity"
##                Df Sum Sq Mean Sq F value Pr(>F)
## type           1  75.09   75.09    4829 <2e-16 ***
## Residuals   6495 100.99    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "citric.acid"
##                Df Sum Sq Mean Sq F value Pr(>F)
## type           1   4.82   4.817   236.4 <2e-16 ***
```

```
## Residuals   6495 132.36   0.020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "residual.sugar"
##               Df Sum Sq Mean Sq F value Pr(>F)
## type           1  17892   17892   899.8 <2e-16 ***
## Residuals   6495 129156      20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "chlorides"
##               Df Sum Sq Mean Sq F value Pr(>F)
## type           1  2.096  2.0956    2316 <2e-16 ***
## Residuals   6495  5.877  0.0009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "free.sulfur.dioxide"
##               Df  Sum Sq Mean Sq F value Pr(>F)
## type           1  455241  455241    1858 <2e-16 ***
## Residuals   6495 1591267     245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "total.sulfur.dioxide"
##               Df   Sum Sq  Mean Sq F value Pr(>F)
## type           1 10179301 10179301    6253 <2e-16 ***
## Residuals   6495 10573600     1628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "density"
##               Df  Sum Sq  Mean Sq F value Pr(>F)
## type           1 0.00891 0.008914    1170 <2e-16 ***
## Residuals   6495 0.04950 0.000008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "pH"
##               Df Sum Sq Mean Sq F value Pr(>F)
## type           1  18.19  18.192     789 <2e-16 ***
## Residuals   6495 149.75   0.023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "sulphates"
##               Df Sum Sq Mean Sq F value Pr(>F)
## type           1  34.15   34.15    2022 <2e-16 ***
## Residuals   6495 109.70    0.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "alcohol"
##               Df Sum Sq Mean Sq F value  Pr(>F)
## type           1     10  10.045   7.068 0.00787 **
## Residuals   6495   9231   1.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test, all p-values of each property are below the acceptance area, hence all properties are directly correlated with the type of wine.

c) How does type and quality of wines affect (separately and together) the percentage of alcohol present in the wine ?

```
print(summary(aov(winequal$alcohol~category,data=winequal)))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## category       2   2069  1034.7   936.9 <2e-16 ***
## Residuals   6494   7172     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(summary(aov(winequal$alcohol~type,data=winequal)))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## type           1     10  10.045   7.068 0.00787 **
## Residuals   6495   9231   1.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the ANOVA test, the p-value for the category $< 2e - 16$ is smaller than the p-value for the type 0.00787 so the **category** has a bigger impact on the quantity of alcohol of the wine.

```
print(summary(aov(winequal$alcohol~category+type,data=winequal)))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## category       2   2069  1034.7 937.567 <2e-16 ***
## type           1      6     6.0   5.458 0.0195 *
## Residuals   6493   7166     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two-way ANOVA shows that the alcohol mean of the wine is still affected by both the category and the type, altough the category has a greater impact on the amount of alcohol of the wine.

d) Detail the results of a two-way ANOVA considering as dependent variable "fixed acidity", and independent variable "type" and "quality".

```
print(summary(aov(winequal$fixed.acidity~category+type,data=winequal)))
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## category       2     57    28.5   22.18 2.51e-10 ***
## type           1   2531  2531.1 1972.91  < 2e-16 ***
## Residuals   6493   8330     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the two-way ANOVA table we can conclude that both category and type are statistically significant. The category is the most significant factor variable. These results would lead us to believe that changing the category or the quality of the wine, will impact significantly the mean of the acidity level.

Not the above fitted model is called *additive model*. It makes an assumption that the two factor variables are independent.