ANOVA.R

monca016

Mon Oct 08 14:33:29 2018

```
## ANOVA Section
# Are the mean number of daily visitors to a ski resort the same for three types of snow conditions?

SnowType <- c(rep("Powder", times = 4), rep("Machine Made", times = 6), rep("Packed", times = 5))
SnowType
```

```
##  [1] "Powder"       "Powder"       "Powder"       "Powder"
##  [5] "Machine Made" "Machine Made" "Machine Made" "Machine Made"
##  [9] "Machine Made" "Machine Made" "Packed"       "Packed"
## [13] "Packed"       "Packed"       "Packed"
```

```
NumbVisitors <- c(1210, 1080, 1537, 941, 2107, 1149, 862, 1870, 1528, 1382, 2846, 1638, 2019, 1178, 2233)

# Powder_NumbVisitors(1210, 1080, 1537, 941)
# MachineMade_NumbVisitors(2107, 1149, 862, 1870, 1528, 1382)
# Packed_NumbVisitors(2846, 1638, 2019, 1178, 2233)

# ANOVA
fit <-aov(NumbVisitors ~ SnowType)

# Total DF (n - 1)
total_df = length(SnowType) - 1
total_df
```

```
## [1] 14
```

```
# Treatment (k - 1)
treatment = length(unique(SnowType)) - 1
treatment
```

```
## [1] 2
```

```
# Error (n - k)
error = total_df - treatment
error
```

```
## [1] 12
```

```
# Sum of Squares Treatment
sst = 1468909
sst
```

```
## [1] 1468909
```

```
# Sum of Squares Error
sse = 2819077
sse
```

```
## [1] 2819077
```

```
# Sum of Squares Total
sstotal = sst + sse
sstotal
```

```
## [1] 4287986
```

```
# Mean of Squares Treatment
mst = 1468909 / 2
mst
```

```
## [1] 734454.5
```

```
# Mean of Squares Error
mse = 2819077 / 12
mse
```

```
## [1] 234923.1
```

```
# F-Value
f_value = round(mst / mse, 3)
f_value
```

```
## [1] 3.126
```

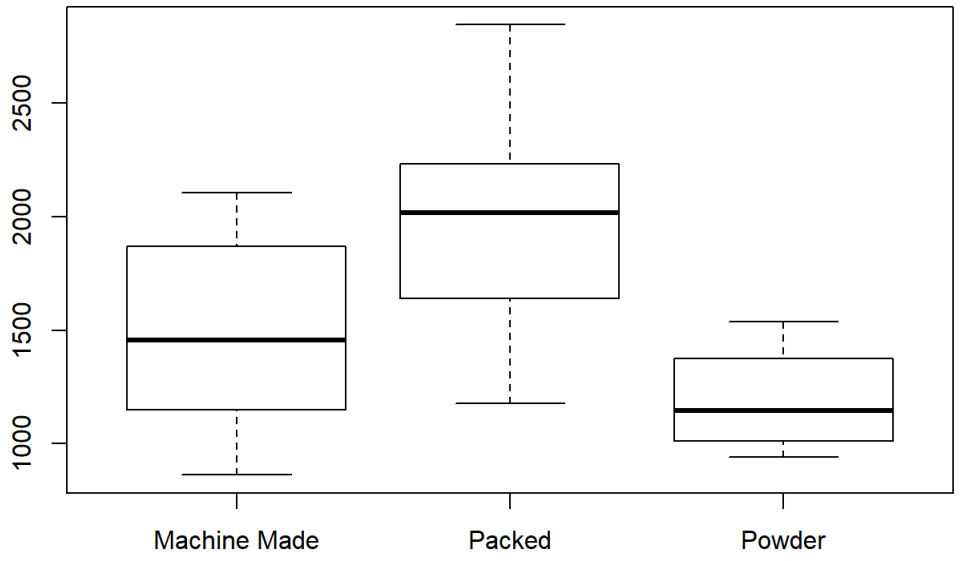
```
summary(fit)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## SnowType     2 1468909  734455   3.126 0.0807 .
## Residuals   12 2819077  234923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(model.tables(fit, "means"))
```

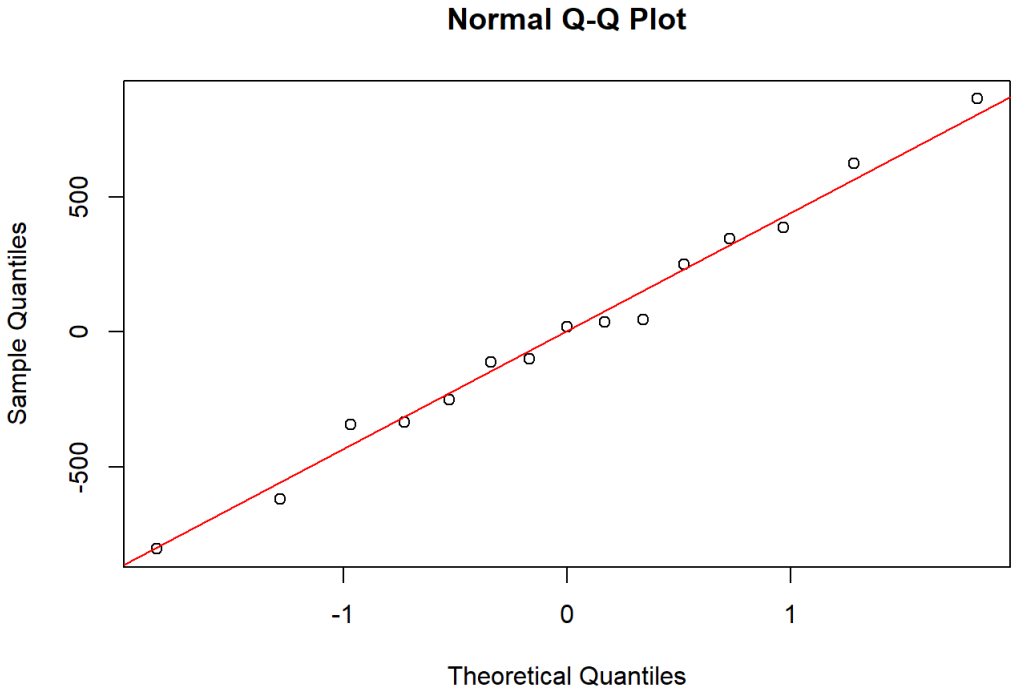
```
## Tables of means
## Grand mean
##
## 1572
##
##  SnowType
##     Machine Made Packed Powder
##             1483   1983   1192
## rep            6      5      4
```

```
boxplot(NumbVisitors ~ SnowType)
```




```
# Assumption checks
  # Residual = Actual - Category mean
  # check the Normal probability plot as a check on the normality assumption

qqnorm(fit$residuals)
qqline(fit$residuals, col = 'red')
```




```
  # goodness of fit test of H0: normal
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.98745, p-value = 0.9974
```

```
  # to check equal variances
# install car package first
#install.packages('car')
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.4.4
```

```
leveneTest(NumbVisitors, SnowType)
```

```
## Warning in leveneTest.default(NumbVisitors, SnowType): SnowType coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.0519 0.3794
##       12
```

```
# to check all the pairwise contrasts
TukeyHSD(fit, conf.level = .90)
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = NumbVisitors ~ SnowType)
##
## $SnowType
##                         diff        lwr        upr     p adj
## Packed-Machine Made  499.8  -165.1191 1164.71913 0.2440083
## Powder-Machine Made -291.0  -999.8062  417.80617 0.6324909
## Powder-Packed       -790.8 -1527.4130  -54.18702 0.0753350
```

```
# R 5.3 Cholesterol
# Drug company compares three different drugs (A, B, C) being developed to reduced cholesterol levels
# Each drug is administered to six patients for 6 months
# After 6 months, reduction in cholesterol level is recorded for each patient
# measures of cholesterol reduction are in Cholesterol.xlsx

library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.4
```

```
Cholesterol <- read_excel("Cholesterol.xlsx", col_names = TRUE)

# set columns to variables to make it faster
chol_drug = Cholesterol$Drug
chol_reduction = Cholesterol$CholReduction

# print variables to make sure values are right
print(chol_drug)
```

```
##  [1] "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
## [18] "C"
```

```
print(chol_reduction)
```

```
##  [1] 22 31 19 27 25 18 40 35 47 41 39 33 15  9 14 11 21  5
```

```
chol_fit <-aov(chol_reduction ~ chol_drug)

# A. Do the three drugs differ?
# Based on the the F value of 40.79 (LARGE), and the p-value of the f distribution being very small
# we reject the null hypothesis that there is no relationship between the drugs and the reduction of
# cholesterol levels, and conclude that the three drugs differ in effectiveness of reducing
# cholesterol levels

summary(chol_fit)
```

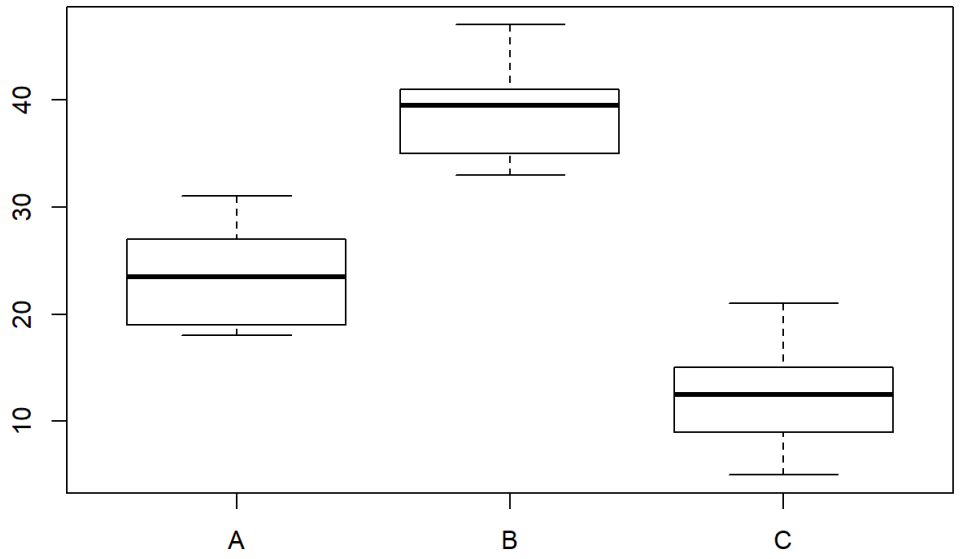
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## chol_drug    2 2152.1  1076.1   40.79 8.59e-07 ***
## Residuals   15  395.7    26.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The means of reduction in cholesterol levels vary between the different drugs
# A = 23.67, B = 39.17, C = 12.50

print(model.tables(chol_fit, "means"))
```

```
## Tables of means
## Grand mean
##
## 25.11111
##
##  chol_drug
## chol_drug
##     A     B     C
## 23.67 39.17 12.50
```

```
# Same with the boxplots for each drug
boxplot(chol_reduction ~ chol_drug)
```

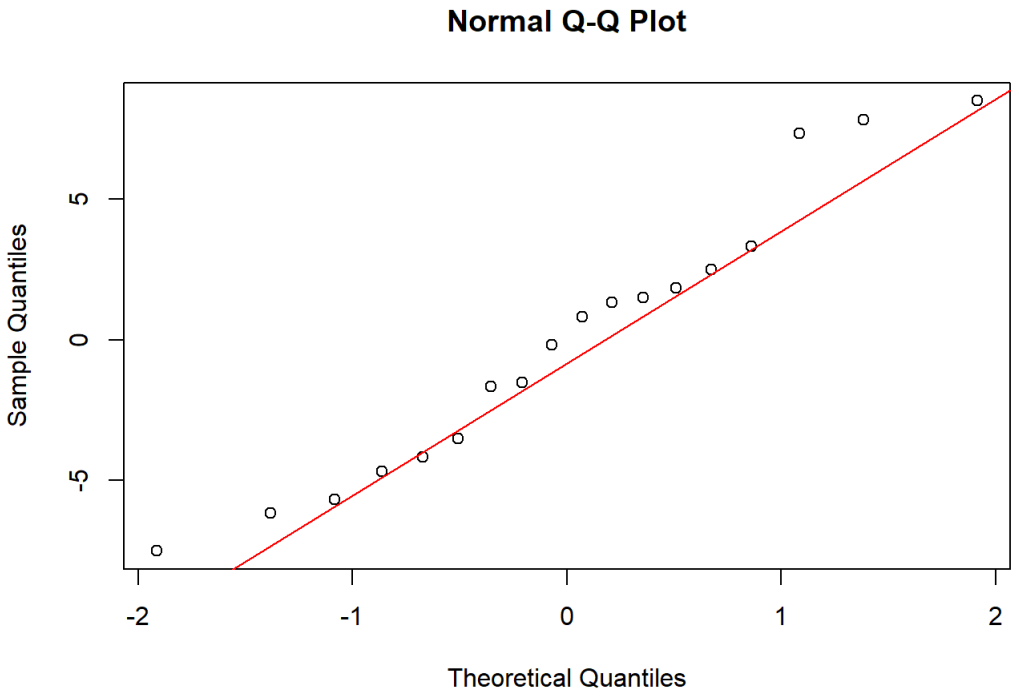



```
# B.  What assumptions are needed for the test?  Check these as possible.

  # (1) Random Sampling
  # (2) Stability of the process
  # (3) Since n < 30 for all three predictor variables, we have to check to see that
  #     the response variable is normally distributed within all the groups, which
  #     it is.
  # (4) We assume a common standard deviation across all of the groups

  # Residual = Actual - Category mean
  # check the Normal probability plot as a check on the normality assumption

qqnorm(chol_fit$residuals)
qqline(chol_fit$residuals, col = 'red')
```




```
  # goodness of fit test of H0: normal
shapiro.test(chol_fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chol_fit$residuals
## W = 0.95396, p-value = 0.4904
```

```
# C.  If a difference is observed, determine which drugs differ from each other

# Since our F - value for our Levene test is small, we accept
# the null hypothesis that our variances across the three groups
# are equal
leveneTest(chol_reduction, chol_drug)
```

```
## Warning in leveneTest.default(chol_reduction, chol_drug): chol_drug coerced
## to factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.0882 0.9161
##       15
```

```
# Since we accept Ha and conclude there is a relationship between the drugs
# and reduction of cholesterol levels, we use the Tukey test
# Based on the ouput from the Tukey test, we can conclude that there is a difference
# between each pair

# B-A p-adj value = .0002841
# C-A p-adj value = .0049961
# C-B p-adj value = .0000006

TukeyHSD(chol_fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = chol_reduction ~ chol_drug)
##
## $chol_drug
##          diff        lwr        upr     p adj
## B-A  15.50000   7.797902  23.202098 0.0002841
## C-A -11.16667 -18.868765  -3.464568 0.0049961
## C-B -26.66667 -34.368765 -18.964568 0.0000006
```

```
#       C                  A               B
# least effective      effective       most effective
```