ANOVA Tutorial

FMU Biology Department

January 3, 2021

ANOVA in R

This documents includes code for creating a data object in R, displaying summary statistics, creating an informative plot, and running an Analysis of Variance (ANOVA).

Background

Nicolas Cage is considered the greatest actor of all time by many people in the United States. However, Hollywood scientists think opinions of this amazing actor are related to the culture of individual cities. Thus, to better understand how successful a new movie with Nicolas Cage will be, scientists must understand these spatial patterns. The data below represent a random selection of preference scores by people in three different South Carolina Cities. Preference scores are expressed as a percentage and range from 0 to 100 with 0 indicating low preference (i.e., does not like Nicolas Cage) and 100 indicates high preference (i.e., loves Nicolas Cage).

Hypothesis: If culture influences the opinion of Nicolas Cage, then opinion ratings will be different across cities.

Null Hypothesis: The opinion rating of Nicolas Cage is not different among cities.

Data

The data frame created below contain 15 observations (rows) and 2 variables (columns). Each row represents one individual observation. The first column is opinion rating reported by the individual and the second column is the city the individual is from.

Important points to note. The function used here is "data.frame" which creates a data frame object that is being assigned the name "dat1". Everything after "data.frame" are the arguments being passed to the function. The arguments include the column names (rating and city) and the data in each column. Quotes ("") are used to encapsulate text. This is important so R knows that this column is a categorical variable. The categorical variable will be used as the predictor of rating.

Enter the following code in R:

```
dat1 = data.frame(rating = c(13,16,8,15,9,
                             29,35,24,27,32,
                             57,59,52,55,60),
                  city = c("Charleston","Charleston","Charleston","Charleston","Charleston",
                           "Columbia","Columbia","Columbia","Columbia","Columbia",
                           "Florence","Florence","Florence","Florence","Florence"))
```

The next line of code prints the data contained in the dat1 object

```
dat1
```

```
##    rating       city
## 1      13 Charleston
```

```
## 2      16 Charleston
## 3       8 Charleston
## 4      15 Charleston
## 5       9 Charleston
## 6      29   Columbia
## 7      35   Columbia
## 8      24   Columbia
## 9      27   Columbia
## 10     32   Columbia
## 11     57   Florence
## 12     59   Florence
## 13     52   Florence
## 14     55   Florence
## 15     60   Florence
```

Display summary data

Important things to note: "dat1$rating" and "dat1$city" references your data frame name (dat1) and the column name (e.g., city)

```
#Average rating for each city
tapply(X = dat1$rating, INDEX = dat1$city, FUN = mean)
```

```
## Charleston   Columbia   Florence
##       12.2       29.4       56.6
```

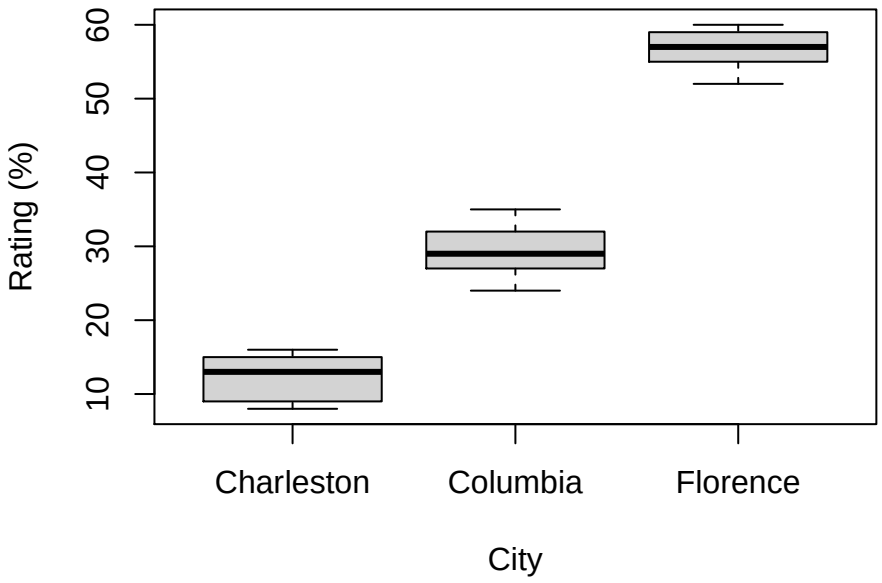
```
#Standard deviation of rating for each city
tapply(X = dat1$rating, INDEX = dat1$city, FUN = sd)
```

```
## Charleston   Columbia   Florence
##   3.563706   4.277850   3.209361
```

Plot data

The next code chunk will display your data in a box plot with custom x- and y-axis labels.

```
boxplot(dat1$rating~dat1$city,
        ylab = "Rating (%)", xlab = "City")
```



Run ANOVA

This code chuck performs an ANOVA, displays summary data, then performs a Tukey post-hoc pairwise comparisons test.

```
res = aov(dat1$rating~dat1$city)
summary(res)
TukeyHSD(res)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## dat1$city    2   5012  2505.9     182 1.06e-09 ***
## Residuals   12    165    13.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dat1$rating ~ dat1$city)
##
## $`dat1$city`
##                       diff      lwr      upr    p adj
## Columbia-Charleston 17.2 10.93951 23.46049 2.52e-05
## Florence-Charleston 44.4 38.13951 50.66049 0.00e+00
## Florence-Columbia   27.2 20.93951 33.46049 2.00e-07
```

Full code block

```r
#Create data frame
dat1 = data.frame(rating = c(13,16,8,15,9,
                             29,35,24,27,32,
                             57,59,52,55,60),
                  city = c("Charleston","Charleston","Charleston","Charleston","Charleston",
                           "Columbia","Columbia","Columbia","Columbia","Columbia",
                           "Florence","Florence","Florence","Florence","Florence"))

#Print data frame
dat1

#Summary statistics
#Average rating for each city
tapply(X = dat1$rating, INDEX = dat1$city, FUN = mean)

#Standard deviation of rating for each city
tapply(X = dat1$rating, INDEX = dat1$city, FUN = sd)

#Create Boxplot
boxplot(dat1$rating~dat1$city,
        ylab = "Rating (%)", xlab = "City")

#ANOVA
res = aov(dat1$rating~dat1$city)
#Display ANOVA results
summary(res)
#Tukey post-hoc comparisons
TukeyHSD(res)
```