A1 Soln

Xiangyu Kong

15/01/2020

Question 1

Read Data

```
# Read the crime show ratings data
crime_show_file = "crime_show_ratings.RDS"
crime_show_data = readRDS(crime_show_file)
```

Question 1.a

Let y_i denote season rating for sample i.

Let $x_{i,2000}$ be indicator variable that is set to 1 if the decade for the sample i is 2000, 0 otherwise.

Let $x_{i,2010}$ be indicator variable that is set to 1 if the decade for the sample i is 2010, 0 otherwise.

Equation for linear model:

$$y_i = \beta_0 + \beta_1 x_{i,2000} + \beta_2 x_{i,2010} + \epsilon_i$$

Anova Assumptions:

1. Errors (ϵ_i) are independent
2. Errors are normally distributed with $E[\epsilon_i] = 0$
3. Errors have constant variance $var[\epsilon] = \sigma^2$

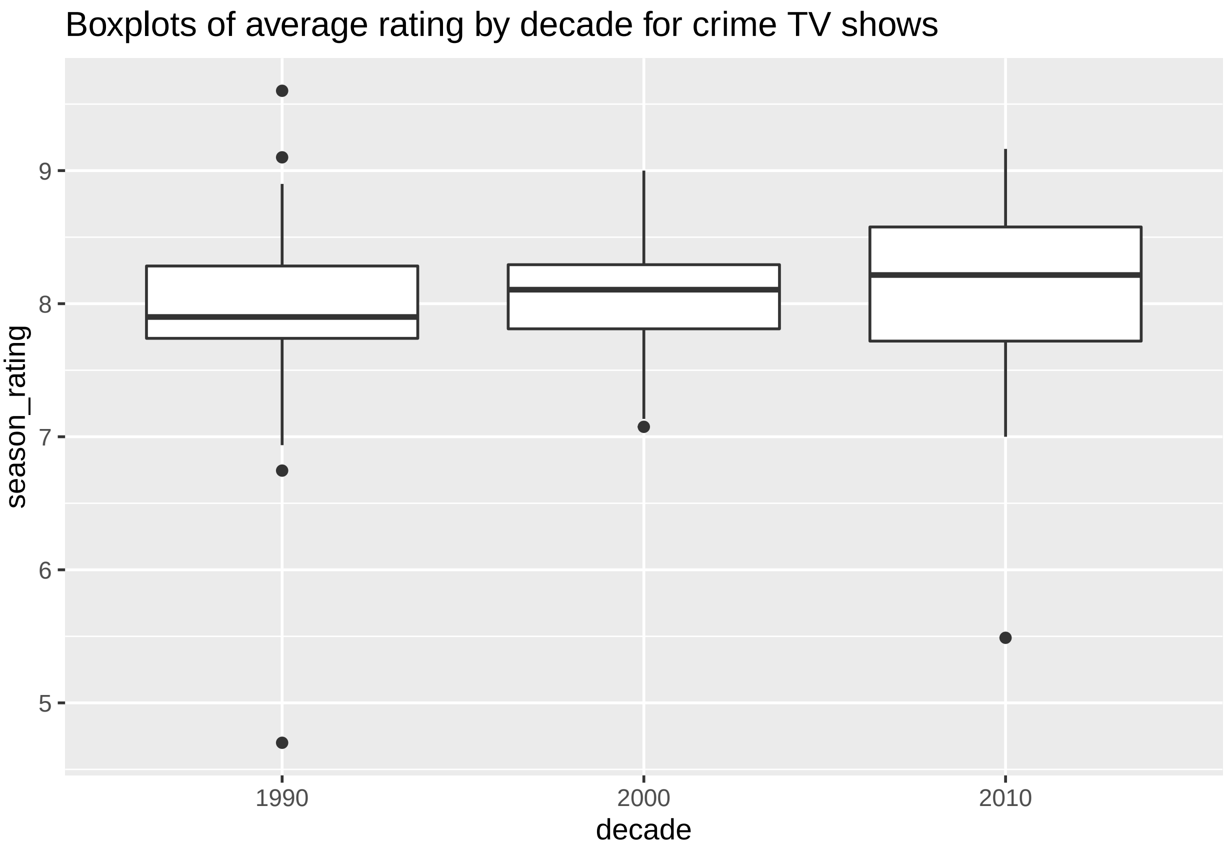
Question 1.b

The hypotheses for ANOVA are listed and can be described as follows:

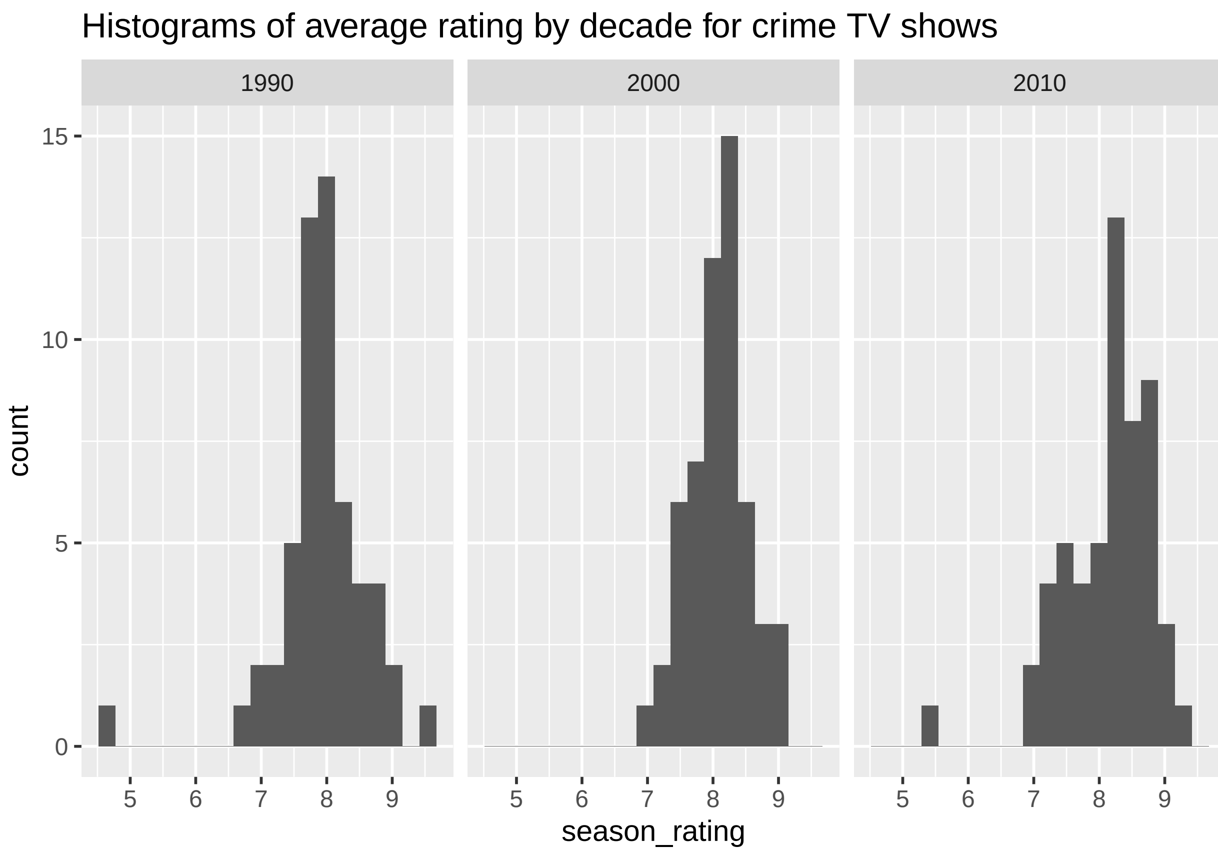
- H_0: $\mu_{1990} = \mu_{2000} = \mu_{2010}$: The mean season rating for crime shows are the same accross different decades. I.e. Different decades does not have effect over season ratings
- H_1: at least one mean is different from the others: The mean season rating for crime shows are different accross different decades. I.e. Different decades has at least some effect over season ratings

Question 1.c

```r
# Side by side box plots
crime_show_data %>% ggplot(aes(x = decade, y = season_rating)) +
    geom_boxplot() + ggtitle("Boxplots of average rating by decade for crime TV shows")
```



```r
# Facetted histograms
crime_show_data %>% ggplot(aes(x = season_rating)) + geom_histogram(bins = 20) +
    facet_wrap(~decade) + ggtitle("Histograms of average rating by decade for crime TV shows")
```



The box plot provides a better visualization of the data because it shows comparison across three decades' basic statistics (maximum, minimum, quartiles, median) side by side.

On the other side, with the histograms, it is harder to tell which decade has a higher rating because it only provides visualization over frequencies within each decade, and provides a relatively poor visualization for comparing between different decades.

One improvement for the box plot could be to cleaning the data before plotting. In the plot, we observe that there are some outliers, especially with 1990 and 2010. Removing those outliers may provide a even better visualization.

Accoring to the box plot, we can see that the boxes are roughly on the same level. There is no sign of extremely skewed data except for some outliers, so their means are similar to the median (all around 8). Thus it does not suggest a signifficant difference between the means.

Question 1.d

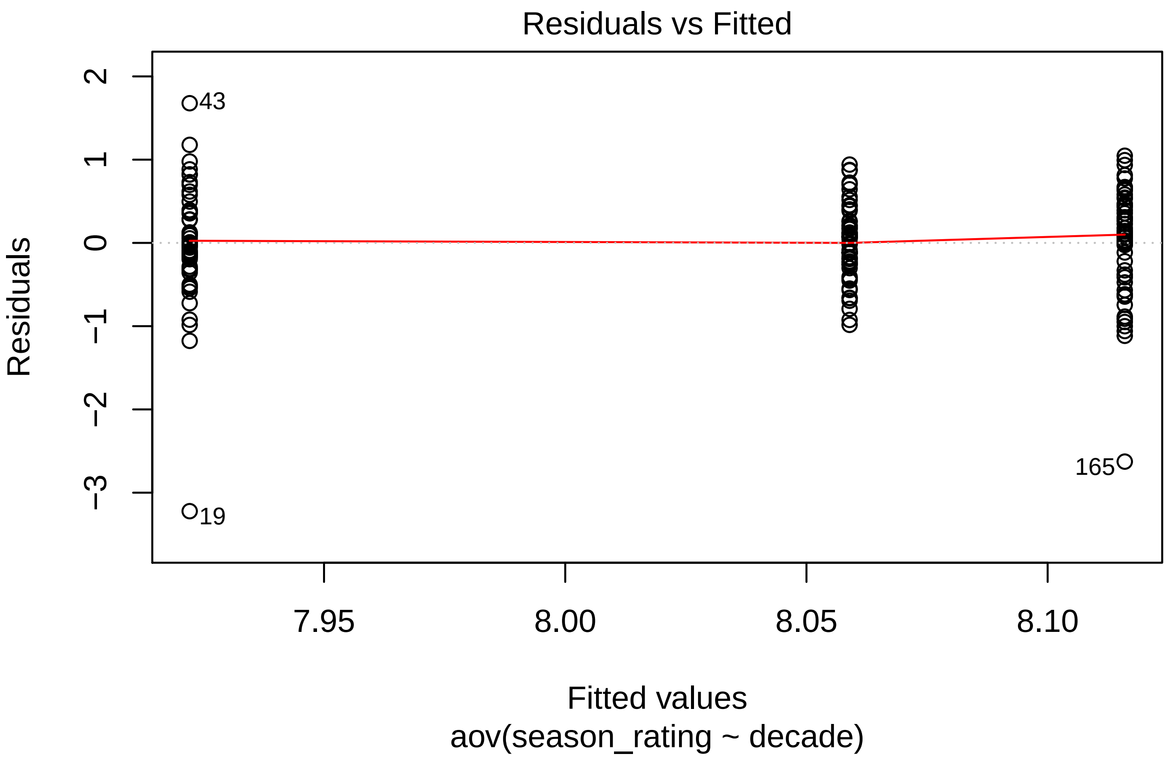
```r
one_way_anova <- aov(season_rating ~ decade, data = crime_show_data)
summary(one_way_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## decade        2   1.09  0.5458   1.447  0.238
## Residuals   162  61.08  0.3771
```

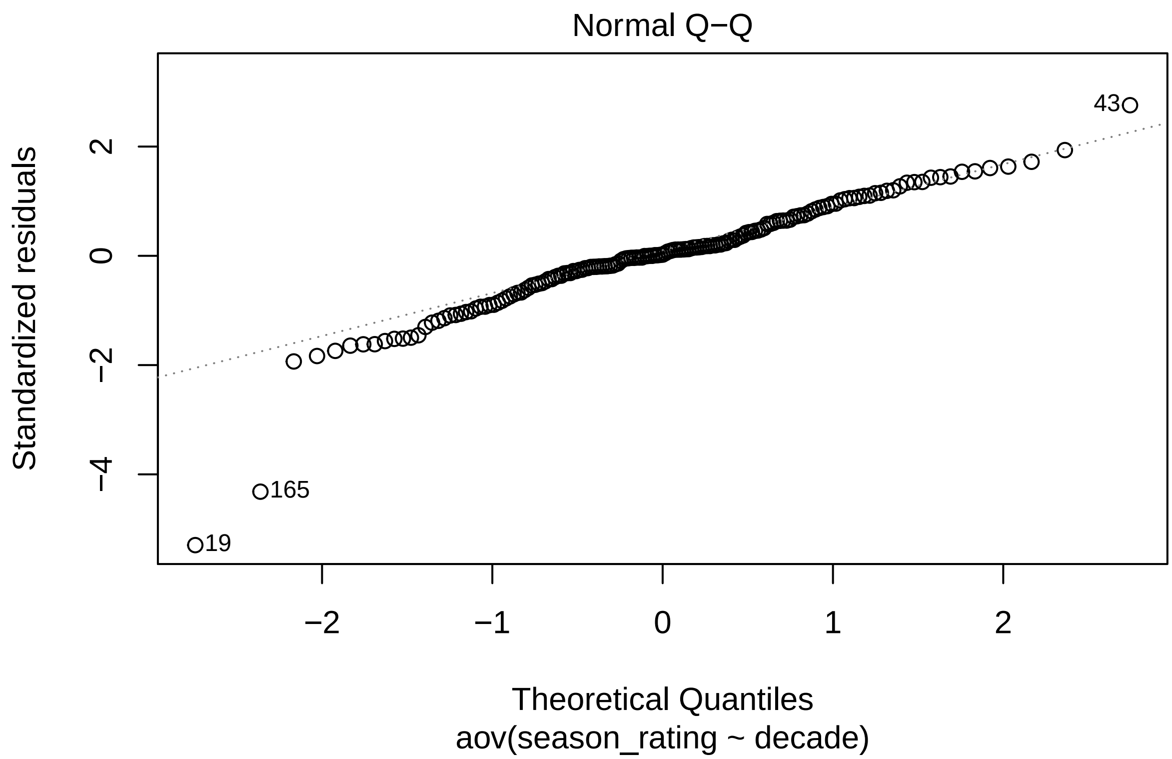
From the one way anova we can see that the p-value for the F-test is 0.238. We can interpret this as: The probability of observing the current sample given the assumption that the three decades having the same mean season ratings is 0.238.

Question 1.e

```r
plot(one_way_anova, 1)
```



```r
plot(one_way_anova, 2)
```




```r
crime_show_data %>% group_by(decade) %>% summarise(var_rating = sd(season_rating)^2)
```

```
## # A tibble: 3 x 2
##   decade var_rating
##   <chr>       <dbl>
## 1 1990        0.480
## 2 2000        0.203
## 3 2010        0.447
```

The first plot is the Residual vs Fitted plot. The plot shows that except for some ouliers, the residuals are roughly randomly scattered around the 0-line, and does not indicate any pattern. This shows that the data follows a linear relationship, have equal error variances, and have a few outliers.

The second plot is the normal q-q plot. From the plot, we can see that except for points 19, 165, and 43, the points form a relatively straight line, indicating that the data follows a normal distribution with a few outliers.

From the standard deviations, we calculate that the ratio of the largest within-group and biggest within-group variance estimate is $\frac{0.480}{0.203} = 2.365 < 3$. According to the rule of thumb from Dean and Voss, the assumption for equality of variances is satisfied.

Question 1.f

```r
lm1 = lm(season_rating ~ decade, data = crime_show_data)

summary(lm1)
```

```
##
## Call:
## lm(formula = season_rating ~ decade, data = crime_show_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2222 -0.2589  0.0135  0.3862  1.6778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.9222     0.0828  95.679   <2e-16 ***
## decade2000    0.1368     0.1171   1.168   0.2444
## decade2010    0.1938     0.1171   1.655   0.0998 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6141 on 162 degrees of freedom
## Multiple R-squared:  0.01756,    Adjusted R-squared:  0.005426
## F-statistic: 1.447 on 2 and 162 DF,  p-value: 0.2382
```

The linear model can be expressed as

$$y = \beta_0 + \beta_1 x_{2000} + \beta_2 x_{2010}$$

where y is the season rating, β_is are the coefficients, x_{2000} is the indicator variable for decade 2000, and x_{2010} is the indicator variable for decade 2010.

β_0 is the intercept of the regression line, which is equal to the sample mean for decade 1990.

β_1 is the amount of score increase when the indicator variable x_{2000} is set to 1.

β_2 is the amount of score increase when the indicator variable x_{2010} is set to 1.

Then the sample mean for decade 1990 $\hat{\mu}_{1990} = \beta_0 = 7.9222$.

The sample mean for decade 2000 $\hat{\mu}_{2000} = \beta_0 + \beta_1 = 7.9222 + 0.1368 = 8.059$.

The sample mean for decade 2000 $\hat{\mu}_{2010} = \beta_0 + \beta_2 = 7.9222 + 0.1938 = 8.116$.

Question 2

Appendix

```
library(tidyverse)

knitr::opts_chunk$set(echo = TRUE, message = FALSE, tidy = TRUE,
    tidy.opts = list(width.cutoff = 60))
# Read the crime show ratings data
crime_show_file = "crime_show_ratings.RDS"
crime_show_data = readRDS(crime_show_file)
# Side by side box plots
crime_show_data %>% ggplot(aes(x = decade, y = season_rating)) +
    geom_boxplot() + ggtitle("Boxplots of average rating by decade for crime TV shows")
# Facetted histograms
crime_show_data %>% ggplot(aes(x = season_rating)) + geom_histogram(bins = 20) +
    facet_wrap(~decade) + ggtitle("Histograms of average rating by decade for crime TV shows")
one_way_anova <- aov(season_rating ~ decade, data = crime_show_data)
summary(one_way_anova)
plot(one_way_anova, 1)
plot(one_way_anova, 2)

crime_show_data %>% group_by(decade) %>% summarise(var_rating = sd(season_rating)^2)
lm1 = lm(season_rating ~ decade, data = crime_show_data)

summary(lm1)
```