Average Season Rating for Crime Shows

Junru Zhang

22/01/2020

Libraries used:

```
library(tidyverse)
```

A random sample of 55 crime shows was taken from each decade (1990s, 2000s, 2010s). The following variables are provided in `crime_show_ratings.RDS`:

Variable	Description
season_number	Season of show
title	Name of show
season_rating	Average rating of episodes in the given season
decade	Decade this season is from (1990s, 2000s, 2010s)
genres	Genres this shows is part of

Question of interest: We want to know if the average season rating for crime shows is the same decade to decade.

Model

Let us begin by fitting a linear model.

$$y_i = \beta_0 + \beta_1 x_{i2000} + \beta_2 x_{i2010} + \epsilon_i,\ \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

The response $y_i, i = 1, 2, 3$ are the season ratings of crime shows in decades 1990s, 2000s and 2010s respectively. x_1 and x_2 are both dummy variables, which can only take on values 0 or 1.

$$x_{i2000} = \begin{cases} 1 & if\ i^{th} \text{ observation belongs to decade 2000s} \\ 0 & if\ i^{th} \text{ observation does not belong to decade 2000s} \end{cases}$$

x_{i2010} is defined in a similar way, by simply replacing decade 2000s with decade 2010s. And ϵ_i is the error term of the i^{th} observation.

Assumptions

The ANOVA assumptions for using the model above are:
1. Each rating is independent.
2. Errors are normally distributed with expectation of zero. i.e. $E[\epsilon_i] = 0$
3. Constant variance

Hypothesis

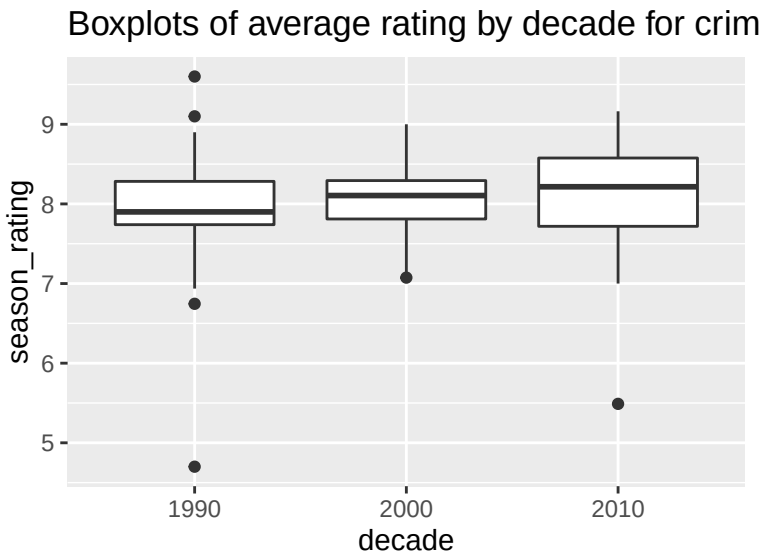
Our null and alternative hypothesises are:

H_0: the average ratings for crime shows are the same in the 1990s, 2000s and 2010s.
H_a: at least one decade's mean season rating for crime shows is different from the others.

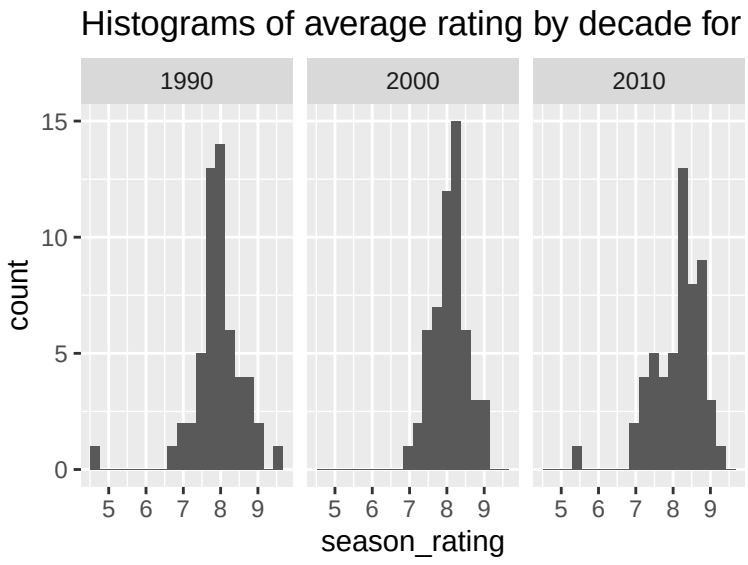
Plots

Let's now create a few plots to visuallize the data:

```
library(tidyverse)
# load crimeshow data
crime_show_data <- readRDS("crime_show_ratings.RDS")
# 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")
```



Based on both graphs, the average ratings of the three decades seem to be around the same. The box-plots provide us the median for each decade. While working with normal data / large data sets, the mean is the same as / approximately the same as the median, so roughly we know what the means for the decades are. We can also easily consider variability through the IQR and range. I would guess that variation between the decades is too small compared to the variation within each decade to expect to see a significant difference between the average season ratings.

ANOVA

Let's test out whether our guess is valid using ANOVA:

```
anova1 <- aov(season_rating ~ as.factor(decade), data= crime_show_data)
summary(anova1)
```

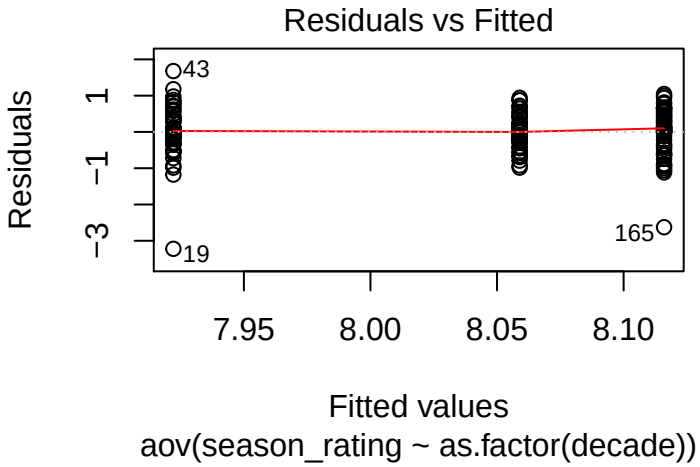
```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(decade)   2   1.09  0.5458   1.447  0.238
## Residuals         162  61.08  0.3771
```

From the results, we see a p-value of 0.238, which is greater than 0.05. Thus, we fail to reject H_0. Hence, there is no significant evidence that the average ratings of the three decades are different under the 5% significance level. In other words, it is reasonable to believe that the average season rating for crime shows is the same decade to decade.

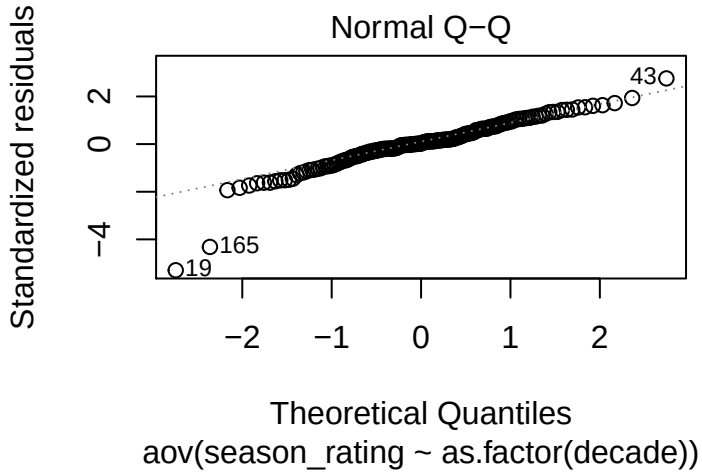
Checking the assumptions

Before we hold our belief to the ANOVA results, let's make sure to check the ANOVA assumptions first

```
# residual plot
plot(anova1, 1)
```




```
# Normal quantile-quantile (q-q) plot
plot(anova1, 2)
```




```
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 resuidual plot shows the residuals of the model against the fitted values. There is no pattern in this plot, which means the constant variance assumption is satisfied. Also the residuals in each group are centered at 0, which means the assumption $E[\epsilon_i] = 0$ is satisfied.

• The second plot is normal Q-Q plot. It shows the standardised residuals against the theoretical normal distribution of residuals. The points form a straight line roughly, though with some outliers (observations 19 and 165 as seen in plot 2), indicate that the assumption of residuals being normally distributed is satisfied.

• The third output shows us the variance of season rating by decade. The largest with-in group variance is 0.480, and the smallest with-in group variance is 0.203. The ratio of the two yields $\frac{0.480}{0.203} = 2.365 <$ 3.Therefore, by the rule of thumb from Dean and Voss (Design and Analysis of Experiments, 1999, page 112), the constant variance assumption issatisfied.

Linear Model instead of ANOVA

```
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
```

We see from the result, linear models and ANOVA produce the same result. Note here the intercept is the mean season rating for crime shows in the decade 1990s (we call this the reference group).
It then follows that:
decade2000 is the difference of mean season rating for crime shows in the decade 2000s and 1990s
decade2010 is the difference of mean season rating for crime shows in the decade 2010s and 1990s.

From the output of summary(), we obtain the observed group means as follows:
$\hat{\mu}_{1990} = 7.9222$,

$$\hat{\mu}_{2000} - \hat{\mu}_{1990} = 0.1368 \implies \hat{\mu}_{2000} = 0.1368 + \hat{\mu}_{1990} \implies \hat{\mu}_{2000} = 0.1368 + 7.9222 \quad \therefore \hat{\mu}_{2000} = 8.0590$$

$$\hat{\mu}_{2010} - \hat{\mu}_{1990} = 0.1938 \implies \hat{\mu}_{2010} = 0.1938 + \hat{\mu}_{1990} \implies \hat{\mu}_{2010} = 0.1938 + 7.9222 \quad \therefore \hat{\mu}_{2010} = 8.1160$$

Conclusion

The average season ratings of crime shows are roughly the same decade by decade, as if no matter how we move forward in time, our fondness for crime shows stays the same.