TU256 Probability and Statistical Inference

PSI CA Part II

Panagiotis Pnevmatikatos

Table of Contents

# 1. Introduction

This statistical analysis was designed to investigate whether a range of film’s attributes like votes crowd sourced from the imdb website, reviews written from authorized film critics, the release decade of the film, the genre etc, impact on a imdb rating. The dataset used in this analysis contains data collected from 2000 records listed on imdb. The dataset was download from kaggle in the form of csv file from <https://www.kaggle.com/datasets/prishasawhney/imdb-dataset-top-2000-movies/data>. (It is also contained in the zip file, next to the present report).

### 1.1 The Research Question

The research question investigated in this research is: How do film’s attributes like votes crowd sourced from the imdb website, reviews written from authorized film critics, the release decade of the film, the genre etc, impact on a imdb rating? The dataset contain a variety of attributes and it was well formatted. However, some minor data cleansing had to happen and sanitize the release year, convert the abbreviated value gross value to integer (eg $5.54M to 5,540,000), sanitize the column name, and create derived columns (see table 3)

### 1.2 The Variables

**Table 1: measurements included in the dataset**

| Concept | Measurement Instrument Used | Variable Type |
| --- | --- | --- |
| imdb rating | Aggregated user ratings for the film in the imdb site | Continuous |
| votes | Number of votes received by the movie | Discrete |
| metascore | Metascore of the movie; a weighted average of the reviews from a large group of respected critics | Continuous |
| release year | The Year of Release | Discrete |
| gross | Gross value of the movie | Discrete |
| duration | The running time of the movie (in minutes) | Discrete |

The dataset also contains categorical film attributes. Details of the variables and potential values are included in Table 2 below.

**Table 2: measurements included in the dataset**

| Concept | Description/Possible Values | Variable Type |
| --- | --- | --- |
| genre | the genre of the film, variety of values | Nominal |
| director | the director of the film, variety of values | Nominal |
| cast | the leading actor of the film, variety of values | Nominal |

On top of the dataset’s existing columns, we have created dynamically derived columns like a set of boolean columns expressing whether a film belongs to a genre category or not and a release\_decade column

**Table 3: derived columns and variables**

| Concept | Description/Possible Values | Variable Type |
| --- | --- | --- |
| genre\_comedy | whether the film is a comedy or not/TRUE, FALSE | Binary Nominal |
| release\_decade | 1920, 1930, …, 2010 | Nominal with multiple categories |

### 1.3 Hypotheses

This research question was investigated through a series of hypotheses tests. The following hypotheses were investigated:

**Table 4: Hypotheses investigated**

| Pair | Null Hypothesis (H0) | Alternative Hypothesis (Ha) |
| --- | --- | --- |
| 1 | There is no relationship between a film’s metascore and their imdb rating. | There is a relationship between a film’s metascore and their imdb rating. |
| 2 | There is no relationship between a film’s votes and their imdb rating. | There is a relationship between a film’s votes and their imdb rating. |
| 3 | There is no difference between the imdb rating for films which are comedies and films which are not | There is difference in the imdb rating for films which are comedies and films which are not |
| 4 | There are no differences between the imdb rating for films with different directors. | There are differences between the imdb rating for films with different directors. |
| 5 | There is no relationship between a film’s genre (comedy/non comedy) and the decade it was released | There is a relationship between a film’s genre (comedy/non comedy) and the decade it was released |

The data for each of the variables of interest was inspected to establish whether the level of missing data represented a source of bias. Using the guidance of Tabachnick and Fidell (2016 neither the proportion nor pattern of missing data represent a source of concern see Figure 1.

## Warning in plot.aggr(res, ...): not enough horizontal space to display  
## frequencies

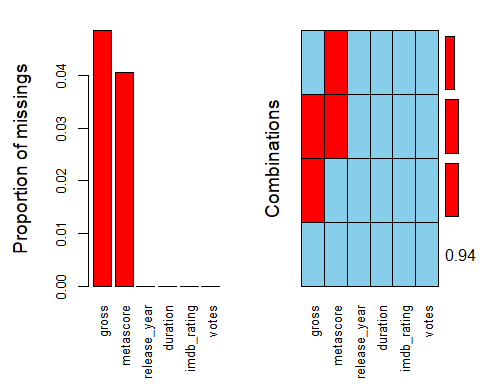


Figure 1: Inspection of Missing Data

##   
## Variables sorted by number of missings:   
## Variable Count  
## gross 0.0485  
## metascore 0.0405  
## release\_year 0.0000  
## duration 0.0000  
## imdb\_rating 0.0000  
## votes 0.0000

##   
## Missings per variable:   
## Variable Count  
## metascore 81  
## gross 97  
## release\_year 0  
## duration 0  
## imdb\_rating 0  
## votes 0  
##   
## Missings in combinations of variables:   
## Combinations Count Percent  
## 0:0:0:0:0:0 1870 93.50  
## 0:1:0:0:0:0 49 2.45  
## 1:0:0:0:0:0 33 1.65  
## 1:1:0:0:0:0 48 2.40

# 2. Hypotheses

## 2.1 Relationship between metascore and imdb rating

H0: There is no relationship between a film’s metascore and their imdb rating.  
Ha: There is a relationship between a film’s metascore and their imdb rating.

### 2.1.2 Inspection of metascore

Metascore scores were assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 2 and Figure 3) identified some issues with skewness and kurtosis. Neither the standardized score for kurtosis (-4.73) nor the standardized score for skewness (-1.44) can be considered acceptable using the criteria proposed by Curran, West and Finch (1996). However as 100% (> 99%) of standardized scores fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2013), the data can be considered to approximate a normal distribution (m=61.04, sd=17.94, n=1919).

Descriptive statistics of metascore

## median mean SE.mean CI.mean.0.95 var std.dev   
## 61.0000000 61.0442939 0.4094773 0.8030675 321.7618535 17.9377215   
## coef.var   
## 0.2938476

Visualizations of metascore

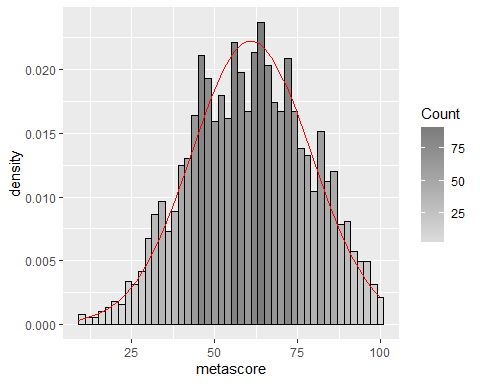


Figure 2: Histogram for metascore

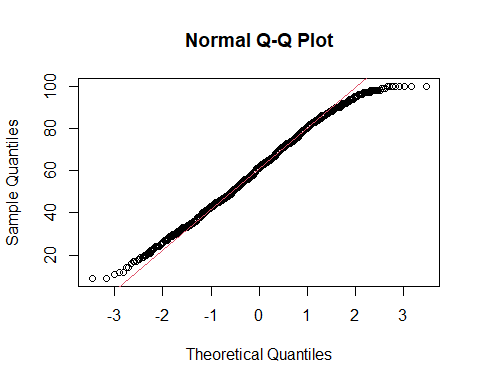


Figure 3: QQPlot for metascore

### 2.1.2 Inspection of imdb rating

Imdb rating scores were assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 4 and Figure 5) identified some issues with skewness and kurtosis. While the standardized score for kurtosis (30) could be considered acceptable using the criteria proposed by Curran, West and Finch (1996), but the standardized score for skewness (-19.74) was outside the acceptable range. However as 99.25% (> 99%)of standardized scores fall within the bounds of +/- 3.29, using the guidance of Field, Miles and Field (2012), the data can be considered to approximate a normal distribution (m=6.92, sd=0.96, n=2000).

Descriptive statistics of imdb rating

## median mean SE.mean CI.mean.0.95 var std.dev   
## 7.00000000 6.92260000 0.02136827 0.04190641 0.91320584 0.95561804   
## coef.var   
## 0.13804323

Visualizations of imdb rating

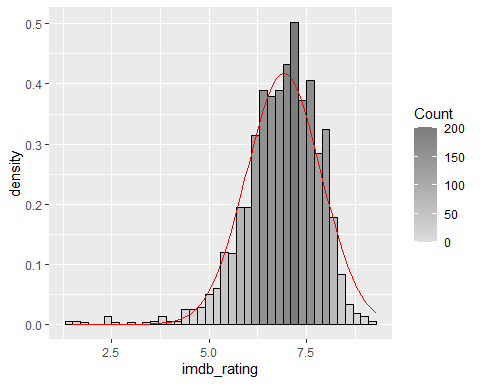


Figure 4: Histogram for imdb rating

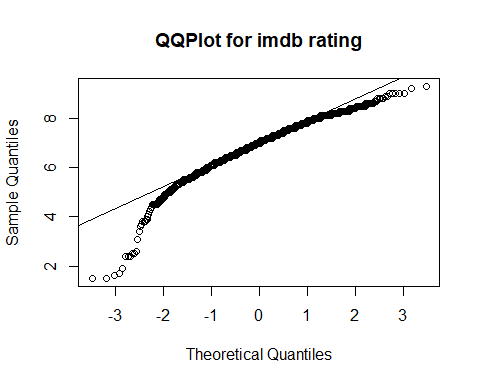


Figure 5: QQPlot for imdb rating

### 2.1.3 Inspection film’s metascore and rating covariance

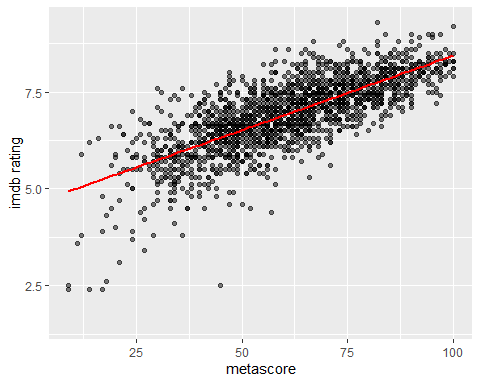


Figure 6: Scatterplot for metascore and imdb rating

### 2.1.4 Pearson’s correlation test on metascore and rating

As both metascore and imdb rating data can be considered approximately normal a Pearson’s correlation test can be used to test this hypothesis

#Pearson Correlation - this is the correct test for this data  
corstat\_metascore\_rating<-stats::cor.test(imdb$metascore, imdb$imdb\_rating, method='pearson')  
corstat\_metascore\_rating

##   
## Pearson's product-moment correlation  
##   
## data: imdb$metascore and imdb$imdb\_rating  
## t = 50.645, df = 1917, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7366863 0.7750078  
## sample estimates:  
## cor   
## 0.7564956

### 2.1.5 Conclusion on Pearson’s correlation test on metascore and rating

The relationship between metascore and imdb rating was investigated using a Pearson correlation. A strong positive correlation was found (r =0.76, n=1917, p<.001).

## 2.2 Relationship between film’s votes and imdb rating

H0: There is no relationship between a film’s votes and their imdb rating.  
Ha: There is a relationship between a film’s votes and their imdb rating.

### 2.2.1 Inspection of film’s votes

Film votes scores were assessed for normality. Visual inspection of the histogram and QQ-Plot (see Figure 7 and Figure 8) identified some issues with skewness and kurtosis. Neither the standardized score for kurtosis (202.47) nor for skewness (72.15) could be considered within the acceptable range using the criteria proposed by Curran, West and Finch (1996). However, as 95.5% (>95%) of standardized scores fall within the bounds of +/- 1.96, using the guidance of Field, Miles and Field (2012), the data can be considered to approximate a normal distribution (m=2.2389533^{5}, sd=2.7669668^{5}, n=2000).

Descriptive statistics of imdb rating

## median mean SE.mean CI.mean.0.95 var std.dev   
## 7.00000000 6.92260000 0.02136827 0.04190641 0.91320584 0.95561804   
## coef.var   
## 0.13804323

Visualizations of votes

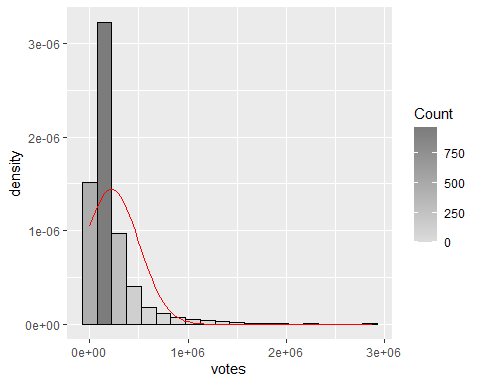


Figure 7: Histogram for votes

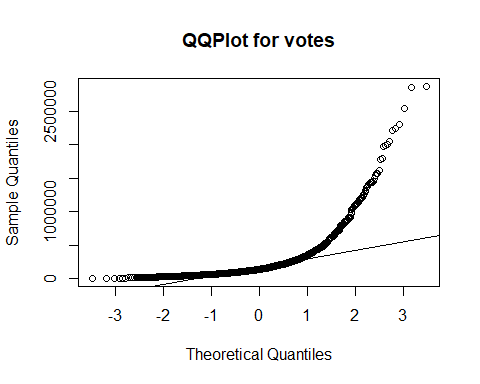


Figure 8: QQPlot for votes

### 2.2.2 Inspection film’s imdb rating

The variable imdb rating has been assessed in 2.1.2

### 2.2.3 Inspection film’s votes and imdb rating covariance

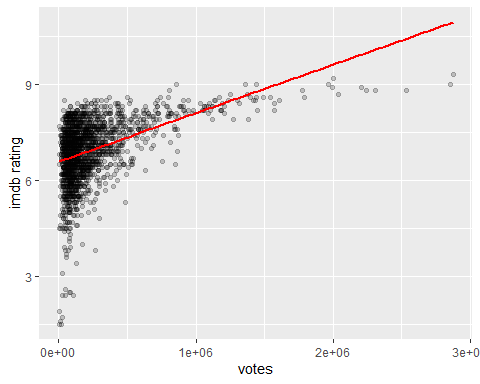


Figure 9: Scatterplot for votes and imdb rating

### 2.2.4 Pearson’s correlation test on votes and rating

As both film votes and imdb rating data can be considered approximately normal a Pearson’s correlation test can be used to test this hypothesis

### 2.2.5 Conclusion on Pearson’s correlation test on votes and rating

The relationship between film’s votes and imdb rating was investigated using a Pearson correlation. A moderate positive correlation was found (r =0.44, n=1998, p<.001).

##   
## Pearson's product-moment correlation  
##   
## data: imdb$votes and imdb$imdb\_rating  
## t = 21.835, df = 1998, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.4028463 0.4736457  
## sample estimates:  
## cor   
## 0.4389271

## 2.3 Difference in imdb ratings for films which are comedies and films which are not

H0: There is no difference between the imdb rating for films which are comedies and films which are not.  
Ha: There is a difference between the imdb rating for films which are comedies and films which are not.

### 2.3.1 Inspection of film’s comedy status

No missing data were found in film genre and as result in genre\_comedy (which derives from the column genre).  
From our film data set, 709 films are comedies while 1291 are not comedies (see Figure 10).

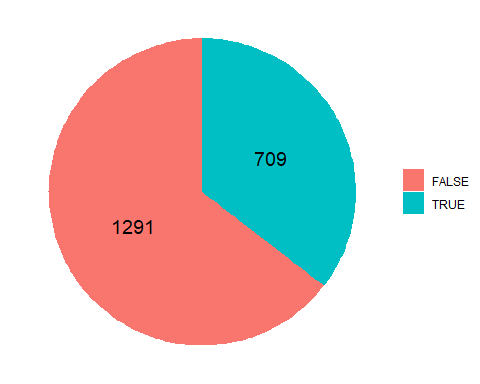


Figure 10: Proportions of Comedy films

### 2.3.2 Inspection of film’s imdb ratings by genre group

#### 2.3.2.1 Inspection of film’s imdb ratings variable as a whole

The variable imdb rating has been assessed in 2.1.2 and has been found to approximate normal distribution

#### 2.3.2.2 Inspection of film’s imdb ratings by genre group

The two groups (comedy/non comedy) of imdb ratings were assessed for normality. Visual inspection of the histogram identified some issues with skewness and kurtosis. However, as over 95% of standardized scores fall within the bounds of +/- 1.96 (as percentages were calculated to 96.4739069% and 95.9721146%). Therefore, using the guidance of Field, Miles and Field (2012), both groups can be considered to approximate the normal distribution with the comedy group distribution attributes as (m=6.68, sd=0.95, n=709) and the non comedy group ones as (m=7.06, sd=0.93, n=1291).

Descriptive statistics of the imdb ratings by group

## item group1 vars n mean sd median trimmed mad min max  
## X11 1 FALSE 1 1291 7.057940 0.9318923 7.2 7.123233 0.88956 1.7 9.3  
## X12 2 TRUE 1 709 6.676164 0.9496707 6.8 6.738313 0.74130 1.5 8.6  
## range skew kurtosis se  
## X11 7.6 -1.017892 2.648106 0.02593598  
## X12 7.1 -1.310348 4.611690 0.03566563

Visualisations of imdb rating by group  
Both film groups plotted along with their density function in the same Histogram. Also Box plot of both groups (see Figure 11, 12)

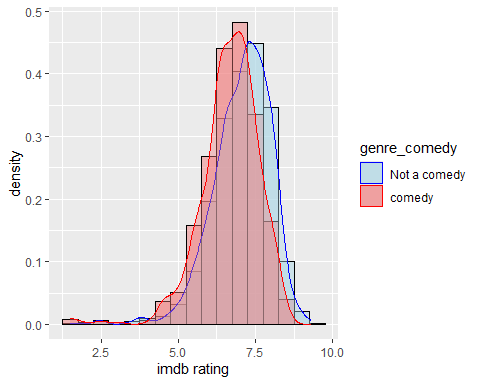


Figure 11: Histogram and Density Plot of imdb rating by comedy Status

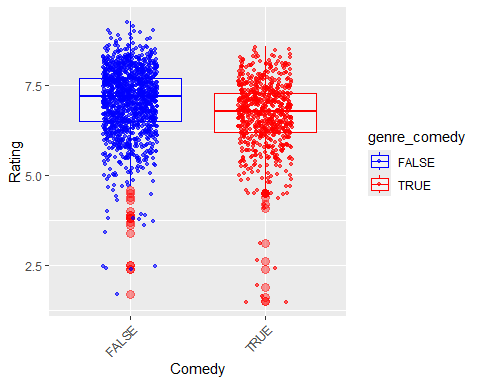


Figure 12: Box Plots of imdb rating by comedy Status

### 2.3.3 Test for homogeneity of variance of film’s imdb ratings by genre group

Furthermore, the Levene test for homogeneity of variance resulted a p-value=0.6629367 therefore the variances can be considered equal.

Levene test

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.19 0.6629  
## 1998

### 2.3.4 Two Sample t-test to compare ratings for films which are comedies and films which are not

The Levene test for homogeneity of variance with a p-value=0.6629367 indicates the variances can be considered equal. Therefore we procced with a two sample t-test

TWO Sample t-test

##   
## Two Sample t-test  
##   
## data: imdb\_rating by genre\_comedy  
## t = 8.705, df = 1998, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0  
## 95 percent confidence interval:  
## 0.2957658 0.4677861  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 7.057940 6.676164

### 2.3.5 Conclusion on comparison of ratings for films which are comedies and films which are not

Cohen’s d effect size

## d | 95% CI  
## -------------------  
## 0.39 | [0.30, 0.48]

An independent-samples t-test was conducted to compare imdb rating for films which are comedies and those which aren’t. An statistically extremely significant difference in the rating was found (M=6.68, SD=0.95, n=709 for the group of comedy films, M=7.06, SD=0.93, n=1291 for non-comedy films), (t(1998)=8.705028, p< 0.05). Cohen’s d indicated a moderate effect size (0.389).

## 2.4 Difference in imdb rating for films released in different decades

H0: There are no differences between the imdb rating for films with different release decades.  
Ha: There are differences between the imdb rating for films with different release decades.

### 2.4.1 Inspection of film’s release decade

No missing data was found for decades. The proportions of films ratings for each decade are shown in Figure 13 and Figure 14.

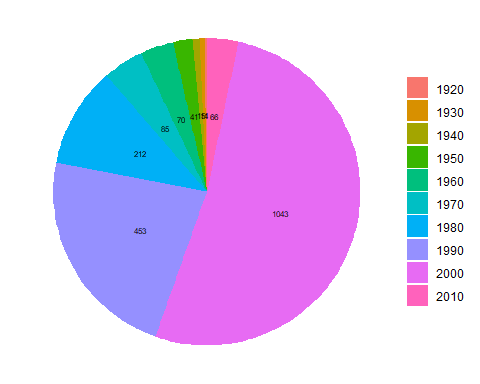


Figure 13: Proportions of film ratings for each decade

### 2.4.2 Inspection of film’s imdb rating by decade

#### 2.4.2.1 Inspection of film’s imdb ratings variable as a whole

The variable imdb rating has been assessed in 2.1.2

#### 2.4.2.2 Inspection of film’s imdb ratings variable by decade

Descriptive statistics of the imdb ratings by decade

## group1 n mean sd median mad min max range skew kurtosis se  
## X11 1920 4 8.175000 0.09574271 8.15 0.07413 8.1 8.3 0.2 0.32046114 -2.0841942 0.04787136  
## X12 1930 11 8.081818 0.28919952 8.10 0.29652 7.6 8.5 0.9 -0.01621436 -1.2945562 0.08719694  
## X13 1940 15 8.140000 0.20976177 8.10 0.29652 7.9 8.6 0.7 0.49926768 -0.7620937 0.05416026  
## X14 1950 41 8.051220 0.32948593 8.10 0.29652 7.3 9.0 1.7 0.08335115 0.5513324 0.05145706  
## X15 1960 70 7.678571 0.50787790 7.70 0.44478 6.5 8.8 2.3 -0.34330590 -0.4718875 0.06070302  
## X16 1970 85 7.568235 0.60831770 7.60 0.59304 5.8 9.2 3.4 -0.41277563 0.8583136 0.06598132  
## X17 1980 212 7.223113 0.75135500 7.30 0.74130 3.7 8.7 5.0 -1.05449099 2.5773789 0.05160327  
## X18 1990 453 6.981015 0.85852922 7.10 0.74130 3.6 9.3 5.7 -0.68903256 1.4388569 0.04033722  
## X19 2000 1043 6.691371 0.96373909 6.70 0.88956 1.5 9.0 7.5 -1.19569742 4.0616849 0.02984127  
## X110 2010 66 6.330303 1.11050504 6.35 0.88956 1.7 8.8 7.1 -0.88702064 3.2427411 0.13669372

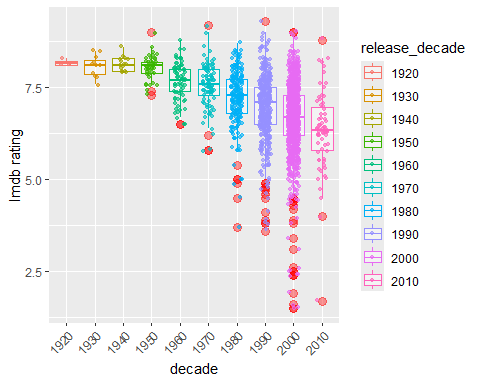


Figure 14: Proportions of film ratings for each decade

### 2.4.3 Bartlett test for homogeneity of variance

A Bartlett’s test for homogeneity of variance has p-value < 2.2e-16 indicates that variances in groups are equal so to assume homogeneity

##   
## Bartlett test of homogeneity of variances  
##   
## data: imdb\_rating by release\_decade  
## Bartlett's K-squared = 175.62, df = 9, p-value < 2.2e-16

### 2.4.4 Anova omnibus test

An one-way ANOVA can be performed as we have assumed homogeneity of variance

##   
## One-way analysis of means  
##   
## data: imdb\_rating and as.factor(release\_decade)  
## F = 38.473, num df = 9, denom df = 1990, p-value < 2.2e-16

## term group1 group2 null.value estimate conf.low conf.high p.adj p.adj.signif  
## 1 release\_decade 1920 1930 0 -0.09318182 -1.7278835 1.541519908 1.00e+00 ns  
## 2 release\_decade 1920 1940 0 -0.03500000 -1.6105072 1.540507247 1.00e+00 ns  
## 3 release\_decade 1920 1950 0 -0.12378049 -1.5903531 1.342792099 1.00e+00 ns  
## 4 release\_decade 1920 1960 0 -0.49642857 -1.9357443 0.942887167 9.86e-01 ns  
## 5 release\_decade 1920 1970 0 -0.60676471 -2.0391992 0.825669802 9.44e-01 ns  
## 6 release\_decade 1920 1980 0 -0.95188679 -2.3649064 0.461132786 5.04e-01 ns  
## 7 release\_decade 1920 1990 0 -1.19398455 -2.6000264 0.212057257 1.78e-01 ns  
## 8 release\_decade 1920 2000 0 -1.48362895 -2.8861856 -0.081072275 2.82e-02 \*  
## 9 release\_decade 1920 2010 0 -1.84469697 -3.2863684 -0.403025558 2.13e-03 \*\*  
## 10 release\_decade 1930 1940 0 0.05818182 -1.0532016 1.169565285 1.00e+00 ns  
## 11 release\_decade 1930 1950 0 -0.03059867 -0.9812749 0.920077551 1.00e+00 ns  
## 12 release\_decade 1930 1960 0 -0.40324675 -1.3113106 0.504817076 9.25e-01 ns  
## 13 release\_decade 1930 1970 0 -0.51358289 -1.4106998 0.383534019 7.27e-01 ns  
## 14 release\_decade 1930 1980 0 -0.85870497 -1.7244847 0.007074705 5.40e-02 ns  
## 15 release\_decade 1930 1990 0 -1.10080273 -1.9551467 -0.246458724 1.90e-03 \*\*  
## 16 release\_decade 1930 2000 0 -1.39044714 -2.2390433 -0.541851023 1.03e-05 \*\*\*\*  
## 17 release\_decade 1930 2010 0 -1.75151515 -2.6633082 -0.839722092 6.29e-08 \*\*\*\*  
## 18 release\_decade 1940 1950 0 -0.08878049 -0.9336229 0.756061881 1.00e+00 ns  
## 19 release\_decade 1940 1960 0 -0.46142857 -1.2580171 0.335159949 7.13e-01 ns  
## 20 release\_decade 1940 1970 0 -0.57176471 -1.3558515 0.212322085 3.83e-01 ns  
## 21 release\_decade 1940 1980 0 -0.91688679 -1.6649161 -0.168857511 4.21e-03 \*\*  
## 22 release\_decade 1940 1990 0 -1.15898455 -1.8937478 -0.424221298 2.80e-05 \*\*\*\*  
## 23 release\_decade 1940 2000 0 -1.44862895 -2.1767009 -0.720557038 1.63e-08 \*\*\*\*  
## 24 release\_decade 1940 2010 0 -1.80969697 -2.6105340 -1.008859947 9.56e-11 \*\*\*\*  
## 25 release\_decade 1950 1960 0 -0.37264808 -0.9232526 0.177956435 4.97e-01 ns  
## 26 release\_decade 1950 1970 0 -0.48298422 -1.0153414 0.049372956 1.14e-01 ns  
## 27 release\_decade 1950 1980 0 -0.82810630 -1.3057671 -0.350445472 2.01e-06 \*\*\*\*  
## 28 release\_decade 1950 1990 0 -1.07020406 -1.5268101 -0.613597995 5.13e-11 \*\*\*\*  
## 29 release\_decade 1950 2000 0 -1.35984847 -1.8056071 -0.914089808 4.38e-11 \*\*\*\*  
## 30 release\_decade 1950 2010 0 -1.72091648 -2.2776498 -1.164183151 4.38e-11 \*\*\*\*  
## 31 release\_decade 1960 1970 0 -0.11033613 -0.5622198 0.341547482 9.99e-01 ns  
## 32 release\_decade 1960 1980 0 -0.45545822 -0.8414045 -0.069511964 7.30e-03 \*\*  
## 33 release\_decade 1960 1990 0 -0.69755598 -1.0571165 -0.337995494 4.34e-08 \*\*\*\*  
## 34 release\_decade 1960 2000 0 -0.98720038 -1.3328815 -0.641519281 4.38e-11 \*\*\*\*  
## 35 release\_decade 1960 2010 0 -1.34826840 -1.8286293 -0.867907447 4.37e-11 \*\*\*\*  
## 36 release\_decade 1970 1980 0 -0.34512209 -0.7045568 0.014312663 7.22e-02 ns  
## 37 release\_decade 1970 1990 0 -0.58721984 -0.9181618 -0.256277905 9.73e-07 \*\*\*\*  
## 38 release\_decade 1970 2000 0 -0.87686425 -1.1926715 -0.561056950 4.37e-11 \*\*\*\*  
## 39 release\_decade 1970 2010 0 -1.23793226 -1.6972638 -0.778600718 4.38e-11 \*\*\*\*  
## 40 release\_decade 1980 1990 0 -0.24209776 -0.4750747 -0.009120770 3.41e-02 \*  
## 41 release\_decade 1980 2000 0 -0.53174216 -0.7426686 -0.320815730 4.39e-11 \*\*\*\*  
## 42 release\_decade 1980 2010 0 -0.89281018 -1.2874507 -0.498169607 9.26e-11 \*\*\*\*  
## 43 release\_decade 1990 2000 0 -0.28964441 -0.4471855 -0.132103344 3.01e-07 \*\*\*\*  
## 44 release\_decade 1990 2010 0 -0.65071242 -1.0195896 -0.281835199 1.17e-06 \*\*\*\*  
## 45 release\_decade 2000 2010 0 -0.36106801 -0.7164299 -0.005706100 4.28e-02 \*

## Eta2 | 95% CI  
## -------------------  
## 0.15 | [0.12, 1.00]  
##   
## - One-sided CIs: upper bound fixed at [1.00].

### 2.4.5 Conclusion on comparison of film ratings for films released in different decades

A one-way between-groups analysis of variance (ANOVA) was conducted to explore the impact of film’s release decade on film rating. The films are divided into ten groups according to the decade they were released (from 1920s to 2010s). There was a statistically significant difference in imdb ratings for the ten decage groups: (F(9, 1990)= 38.473, p<0.05. The effect size, calculated using eta squared was (0.15). Post-hoc comparisons using the Tukey HSD test indicated statistically significant differences between several groups that can be observed in the table bellow

## group1 group2  
## 8 1920 2000  
## 9 1920 2010  
## 15 1930 1990  
## 16 1930 2000  
## 17 1930 2010  
## 21 1940 1980  
## 22 1940 1990  
## 23 1940 2000  
## 24 1940 2010  
## 27 1950 1980  
## 28 1950 1990  
## 29 1950 2000  
## 30 1950 2010  
## 32 1960 1980  
## 33 1960 1990  
## 34 1960 2000  
## 35 1960 2010  
## 37 1970 1990  
## 38 1970 2000  
## 39 1970 2010  
## 40 1980 1990  
## 41 1980 2000  
## 42 1980 2010  
## 43 1990 2000  
## 44 1990 2010  
## 45 2000 2010

## 2.5 Relationship between a film’s genre (comendy/non comedy) and decade it was released

H0: There is no relationship between a film’s genre (comedy/non comedy) and the decade it was released.  
Ha: There is a relationship between a film’s genre (comedy/non comedy) and the decade it was released.

### 2.5.1 Inspection of release decade

The variable release\_decade has been assessed in 2.4.1. The proportion of films in each release decade is shown in Figure 13 and Figure 14.

### 2.5.2 Inspection of genre (comedy/non comedy)

The variable genre\_comedy has been assessed in 2.3.1. The proportion of films in each genre group (comedy/non-comedy) is shown in Figure 10.

### 2.5.3 Inspection of films genre (comedy/non comedy) by release decade

Table 5 shows the numbers of films in each decade in each genre group (comedy/non-comedy).

**Table 5: Hypotheses investigated**

Table 1: Release Decade and Genre (Comedy/Non Comedy) Counts

| Release decade | Non-Comedy | Comedy |
| --- | --- | --- |
| 1920 | 1 | 3 |
| 1930 | 7 | 4 |
| 1940 | 14 | 1 |
| 1950 | 33 | 8 |
| 1960 | 52 | 18 |
| 1970 | 66 | 19 |
| 1980 | 131 | 81 |
| 1990 | 282 | 171 |
| 2000 | 664 | 379 |
| 2010 | 41 | 25 |

### 2.5.4 Test the independance and measure the association of film’s genre (comedy/non comedy) and the film’s release decade

As both release decade and genre comedy are represented by nominal variables, a Chi-squared test of independence can be used to test the hypothesis.

Pearson’s chi-square test

##   
## Pearson's Chi-squared test  
##   
## data: contigency\_decade\_comedy  
## X-squared = 24.271, df = 9, p-value = 0.003893

Expected Frequencies

## genre\_comedy  
## release\_decade FALSE TRUE  
## 1920 2.5820 1.4180  
## 1930 7.1005 3.8995  
## 1940 9.6825 5.3175  
## 1950 26.4655 14.5345  
## 1960 45.1850 24.8150  
## 1970 54.8675 30.1325  
## 1980 136.8460 75.1540  
## 1990 292.4115 160.5885  
## 2000 673.2565 369.7435  
## 2010 42.6030 23.3970

Pearson chi-square test’s p-value suggests an association between genre and release decade and by observing the adjusted residuals table and our contigency table we can identify the decades 1940, 1950 and 1970 as less likely to releasing comedies

Adjusted residuals

## genre\_comedy  
## release\_decade FALSE TRUE  
## 1920 -1.65521712 1.65521712  
## 1930 -0.06352009 0.06352009  
## 1940 2.33918848 -2.33918848  
## 1950 2.15556921 -2.15556921  
## 1960 1.73339261 -1.73339261  
## 1970 2.57963111 -2.57963111  
## 1980 -0.88769930 0.88769930  
## 1990 -1.16272711 1.16272711  
## 2000 -0.86617857 0.86617857  
## 2010 -0.41946147 0.41946147

However, to measure the association between film’s comedy status and film’s release decade we can Calculate cramer’s V

## [1] 0.1101607

### 2.5.5 Conclusion on examining the relation between film’s genre (comedy/non comedy) and release decade

A Chi-Square test for independence indicated a statistically significant association between film’s comedy genre and the film’s release decade: χ²(9, n=2000)=24.271, p=0.0039<0.05.

While the association is statistically significant the small effect size that is resulted by the calculation of Cramer’s V= 0.11 suggests the association is relatively weak.

# 3. Conclusions

As a result of the investigations and the tests conducted in section 2, the following the following conclusions have been drawn:

1. **Pair 1**
   * H0: There is no relationship between a film’s metascore and their imdb rating.
   * Ha: There is a relationship between a film’s metascore and their imdb rating.
   * **Conclusion**: A strong, positive, statistically significant correlation was found between a film’s metascore and their imdb rating. There is therefore evidence to reject the null hypothesis in favour of the alternate hypothesis and the result suggests that higher film metascores are associated with higher ratings.
2. **Pair 2**
   * H0: There is no relationship between a film’s votes and their imdb rating.
   * Ha: There is a relationship between a film’s votes and their imdb rating.
   * **Conclusion**: A moderate, positive, statistically significant correlation was found between a film’s votes and their imdb rating. There is therefore evidence to reject the null hypothesis in favour of the alternate hypothesis and the result suggests that higher film votes are associated with higher ratings.
3. **Pair 3**
   * H0: There is no difference between the imdb rating for films which are comedies and films which are not.
   * Ha: There is a difference between the imdb rating for films which are comedies and films which are not.
   * **Conclusion**: A statistically extremely significant difference in the film rating was found between comedy films and non-comedy films. There is therefore evidence to support rejecting the null hypothesis in favour of the alternate hypothesis.
4. **Pair 4**
   * H0: There are no differences between the imdb rating for films with different release decades.
   * Ha: There are differences between the imdb rating for films with different release decades.
   * **Conclusion**: A statistically significant difference in film ratings was found for films released in different decades. Therefore there is evidence to support rejecting the null hypothesis in favour of the alternate hypothesis.
5. **Pair 5**
   * H0: There is no relationship between a film’s genre (comedy/non comedy) and the decade it was released.
   * Ha: There is a relationship between a film’s genre (comedy/non comedy) and the decade it was released.
   * **Conclusion**: A statistically significant association was found between a film’s genre and its release decade. However, the association between film’s genre and film’s release decade was found to be relatively weak. Therefore while there is evidence to reject the null hypothesis in favour of the alternate hypothesis and the result suggests that comedies tend to be released durring certain decades more often than other genre films.

# References

Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24, 385–96.  
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.  
Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349–54.  
Curran, Patrick J., Stephen G. West, and John F. Finch. (1996). “The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis.” Psychological methods 1.1 (1996): 16., E., Emmons, R.A., Larson, R.J. & Griffin, S. (1985). The Satisfaction with Life scale. Journal of Personality Assessment, 49, 71–6.  
Field, A., Field Z. & Miles J.(2012). Discovering statistics using IBM SPSS statistics. Sage publications limited,  
Pallant, J. (2000). Development and validation of a scale to measure perceived control of internal states. Journal of Personality Assessment, 75, 2, 308–37. Pearlin, L. & Schooler, C. (1978). The structure of coping. Journal of Health and Social Behavior, 19, 2–21.  
Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.  
Scheier, M.F. & Carver, C.S. (1985). Optimism, coping and health: An assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219–47.  
Watson, D., Clark, L.A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–70.  
Tabachnik and Fidell, Using Multivariate Statistics, 6th Edition, Pearson.