A distribution free analysis on Hollywood movies: were older movies just better?

Disclosure:

Data on films and ratings were obtained from The Movie Database (TMDb). This project uses the TMDb API but is not endorsed or certified by TMDb.

Motivation:

Me and my friends often have discussions about movies. We'll watch a movie from 10 years ago, and, inevitably, one of my friends will say something along the lines of "They just made movies better back then". I am not so sure, so I am going to employ non-parametric methods on movie data from different decades to see if old movies really were better.

Introduction:

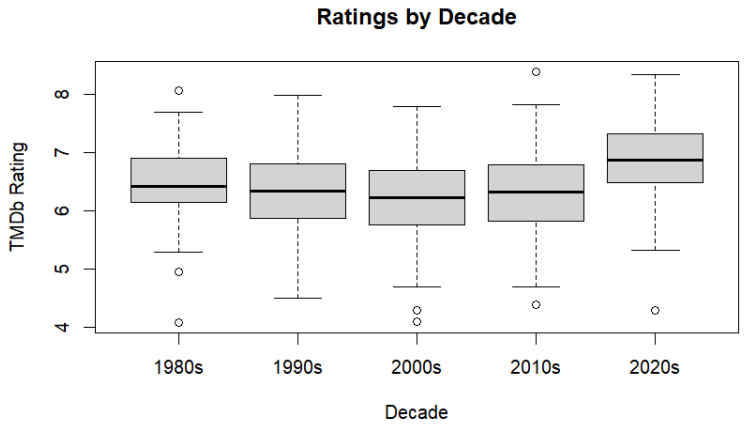
I have collected a sample of 500 movies from 5 different decades (100 movies from each decade). The 1980s, 1990s, 2000s, 2010s, and 2020s (2020-2025). These data were obtained from the platform "The Movie Database" which I will abbreviate as TMDb from now on. Each movie has an average rating on the website. I filtered for movies that have at least 100 reviews using their non-commercial API.

To answer my question, I will be using the Kruskal-Wallis distribution free test for one-way layout problems. There are a number of Assumptions we must verify before employing this method though.

The first is that the N observations $\{X_{1j}, X_{2j}, \dots, X_{nj}\}$ $(j = 1,\dots,k)$ are mutually independent. For the sake of this analysis, this assumption holds since knowing the average rating of one movie is not going to give you any information about another movie's rating.

Next, we need to verify that our data comes from continuous distributions. The distribution comes from the average movie ratings that vary from 0 to 10. Since any value is possible, the distribution of each decade's movies is indeed continuous.

Lastly, the Kruskal-Wallis method assumes that each distribution is identical except for shift in location. This means that we need to verify that the distributions have the same shape and scale. To answer this question, I started with side-by-side boxplots for each distribution.

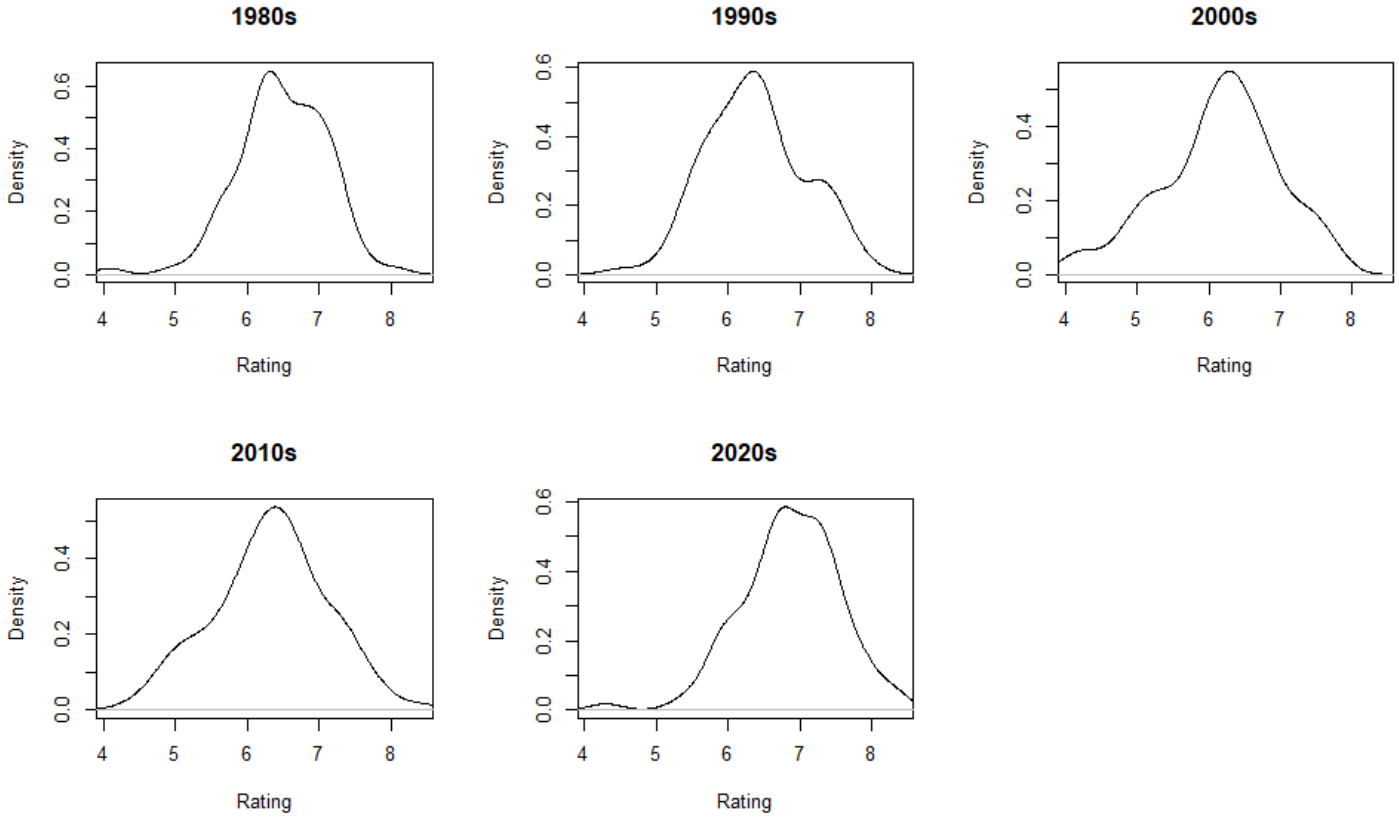


The box-plots largely look like they have similar Interquartile ranges. The 2020s look like they may have a location shift, and the 1980s look like they might have a smaller IQR than the other decades, but nothing that immediately raises red flags for me. For further verification, I generated summary statistics about the standard deviation and IQR for each decade as entailed in the table below.

	1980s	1990s	2000s	2010s	2020s
SD	.624	.685	.821	.789	.68

IQR	.756	.9165	.921	.947	.838

Once again, I am happy with these numbers. There aren't any large outliers in these statistics, but this only gives us a sense of scale for each respective decade's distribution. To get an idea of the shape of each distribution, I generated density plots of each decade's distribution.



Looking at these density plots, I see that each distribution has one peak with tails that taper off the farther from the center they are. I am hesitant to say that these distributions are approximately normal. To verify that, I would need to generate QQ-plots and run a Shapiro-wilk test on my data. Luckily, for Kruskal-Wallis, we are not concerned with normality, but rather only if the distributions have the same shape and scale

Analysis:

My calculation for the H statistic for the Kruskal-Wallis test was 47.223. Since there were ties in the data, I also applied a correction factor of 1 / .9998521 which is approximately 1.000148. I think my large sample size minimized any impact of ties. My final H statistic was 47.229.

Given my large sample size, I used the large sample approximation for the p-value which says I may reject the null hypothesis if the H statistic is greater than a critical value derived from a chi-squared distribution with 4 degrees of freedom. The critical value I obtained was 11.14329. Since 47.229 is larger than 11.14329 we reject the null hypothesis that all treatment effects are equal. At least one treatment effect is different than the others.

Given my original question, I was not concerned with each pairwise comparison. I only wanted to determine if movies from the 2020s are better or worse than movies from previous decades. To this end, I performed a Wilcoxon rank-sum test between the 100 movies from the 2020s and the 400 movies from the other decades. This yielded a W statistic of 28164 and a p-value of 2.664e-10. A point estimate of the difference in location is 0.524 with a confidence interval of (0.377, 0.685).

Monotonicity:

Comment 9 of our textbook brings up the issue of monotonicity in the context of the Kruskal-Wallis test. From the paper cited (Gabriel, 1969), the definition of monotonicity of a family of tests is as follows:

A testing family $\{\Omega, \mathbf{Z}\}$ will be called *monotone* if, whenever $i \prec j$, the numerical relation

$$Z_i(y) \geqq Z_j(y) \tag{2.1}$$

The notation of i < j is indicating that the ith hypothesis is more specific than the jth hypothesis (j is a subset of i). The comparison of Zi(y) >= Zj(y) is saying that the critical point for the stronger test is greater than or equal to the critical point of the weaker/less specific test.

When a family of tests are monotone, they are considered to be coherent. For example, if we are testing the null hypothesis H0_1: θ1 = θ2 = θ3 as well as the test H0_2: θ1 = θ2, we would expect the critical point for H0_1 to be greater than H0_2 since if all three medians are equal then θ1 and θ2 would also have to be equal. In other words, since H0_2 is a subset of H0_1, we would expect H0_2 to be easier to reject than H0_1. This property is not present in the Kruskal-Wallis test.

But, since we are not performing a family of tests, we don't need to be concerned about the coherency of our conclusion. Additionally, Gabriel mentions that the Kruskall-Wallis method tends to be coherent with large enough sample sizes. He goes so far as to call the Kruskall-Wallis method "asymptotically coherent" (P. 249).

Asymptotic Relative Efficiency (ARE) of H and ANOVA:

When deciding what test to use to test certain hypotheses it is important to understand the assumptions of certain tests. For example, the Kruskal-Wallis test has relatively relaxed assumptions compared to the assumptions of ANOVA which require the groups being compared to be normally distributed. One advantage of ANOVA is that it is less prone to Type II errors

(Failing to reject the null hypothesis given a certain change in locations) when its underlying assumptions are met.

Asymptotic Relative Efficiency measures the difference between the power of two tests against consistent alternative hypotheses. When the data of each group is normally distributed, the asymptotic relative efficiency of the H and F statistics can be shown to be $3/\pi$ which is about 0.955 (Andrews FC, 1954). This means that if our data is normally distributed and we want the H statistic to have the same Type II error as the F statistic, we would need a sample size that is about 5 percent larger for the H statistic than the F statistic. Additionally, if we know our data is uniformly distributed, then the asymptotic relative efficiency of H and F is 1.

Given the data is normally distributed, ANOVA only provides us with a slight benefit compared to the Kruskal-Wallis method: it affords us the ability to take a smaller sample and have the same Type II error as an analogous Kruskal-Wallis test. Compared with the advantages of the H statistic (Robustness, ease of calculation, fewer assumptions, and only needing ordinal data) (Chan & Walmsly, 1997) it seems like the F statistic and analysis of variance could be replaced by the Kruskal-Wallis method. There are applications of ANOVA in linear regression, cluster sampling, and experimental design, but as far as testing the hypothesis of equal treatment effects among treatment groups, it seems that the H statistic performs comparably to the F statistic and affords other advantages as well.

If an analysis is solely interested in comparing means, the Kruskal-Wallis method would be inapplicable. Additionally, ANOVA is fairly robust against mild non-normality (Laerd Statistics). ANOVA is also usable in factorial designs and in ANCOVA (analysis of covariance). Both ANOVA and the Kruskal-Wallis method have their respective usages.

Conclusion:

From the result of the Kruskal-Wallis test, there is strong evidence to support the claim that movies from the 2020s are better (or at least rated better by TMDb users) than movies from previous decades. The Kruskal-Wallis test is a rank based non-parametric alternative to using the F statistic to test for equality of treatment effects among 3 or more groups. It affords us advantages like not needing data that is normally distributed and having robustness against outliers and skewed data, while having a strong asymptotic relative efficiency with the F statistic. This means there is very little drop in accuracy when using the Kruskal-Wallis test as compared to ANOVA. One downside of the Kruskal-Wallis method is that, when employing a family of tests, these tests will not necessarily be monotone. Certain tests that are a subset of others may be harder to reject than their superset. This may result in incoherent results that make conclusions on data difficult. In this application, monotonicity of a family of tests did not come up since we were concerned with only one result and we did not need to employ a family of tests. With that in mind, it is important to note that it is relatively rare to find incoherent results with sufficiently large sample sizes.

Bibliography:

Andrews, F. C. “Asymptotic Behaviour of Some Rank Tests for Analysis of Variance.” Annals of Mathematical Statistics, vol. 25, 1954, pp. 724–736.

Chan, Y., and R. P. Walmsley. “Learning and Understanding the Kruskal-Wallis One-Way Analysis-of-Variance-by-Ranks Test for Differences among Three or More Independent Groups.” Physical Therapy, vol. 77, no. 12, 1997, pp. 1755–1762.

Gabriel, K. R. “Simultaneous Test Procedures: Some Theory of Multiple Comparisons.” The Annals of Mathematical Statistics, vol. 40, no. 1, 1969, pp. 224–250.

Laerd Statistics. "One-way ANOVA." Laerd Statistics, https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-3.php Accessed 10 December, 2025

The Movie Database (TMDb). “TMDb API.” themoviedb.org, https://developer.themoviedb.org/docs/getting-started. Accessed 9 Dec. 2025.

Appendix:

Raw Data:

https://github.com/DaneRasmussen/Movies-Kruskal-Wallis/blob/main/movies.csv

R code:

```
# Load in Data
movies <- read.csv("Movies.csv")
head(movies)

names(movies)

# drop unneccesary columns
movies$adult <- NULL
movies$backdrop_path <- NULL
movies$overview <- NULL
movies$poster_path <- NULL
movies$video <- NULL

unique(movies$decade)

## [1] "1980s" "1990s" "2000s" "2010s" "2020s"

movies[movies$decade == "1980s", c("title", "vote_average", "vote_count") ]

nrow(movies)

## [1] 500

nrow(movies[movies$decade == "1980s", c("title", "vote_average", "vote_count
")])

## [1] 100

nrow(movies[movies$decade == "1990s", c("title", "vote_average", "vote_count
")])

## [1] 100

nrow(movies[movies$decade == "2000s", c("title", "vote_average", "vote_count
")])
```

```
## [1] 100
```

```
nrow(movies[movies$decade == "2010s", c("title", "vote_average", "vote_count
")])
```

```
## [1] 100
```

```
nrow(movies[movies$decade == "2020s", c("title", "vote_average", "vote_count
")])
```

```
## [1] 100
```

```
boxplot(vote_average ~ decade, data = movies,
        xlab = "Decade",
        ylab = "TMDb Rating",
        main = "Ratings by Decade")

# Distribuition verification
sd_by_decade <- tapply(movies$vote_average, movies$decade, sd)
iqr_by_decade <- tapply(movies$vote_average, movies$decade, IQR)

sd_by_decade
```

```
##     1980s     1990s     2000s     2010s     2020s
## 0.6235081 0.6853738 0.8208011 0.7889197 0.6802051
```

```
iqr_by_decade
```

```
##   1980s   1990s   2000s   2010s   2020s
## 0.75600 0.91650 0.92125 0.94650 0.83825
```

```
# Density plots from each decade
par(mfrow = c(2, 3))
for (d in unique(movies$decade)) {
  x <- movies$vote_average[movies$decade == d]
  plot(density(x), main = d, xlab = "Rating", xlim = range(movies$vote_averag
e))
}
par(mfrow = c(1, 1))

# Kruskall Wallace H statistic
movies$ranked_votes <- rank(movies$vote_average)
N <- nrow(movies)
nj <- 100 # All decades have a sample of 100 movies
Tj <- tapply(movies$ranked_votes, movies$decade, sum)
H <- 12/(N * (N+1)) * sum((Tj ** 2) / nj) - 3 * (N+1)


# Correction factor for ties in the ranks
tab_ties <- table(movies$ranked_votes)           # counts of identical scores
t <- as.numeric(tab_ties[tab_ties > 1])   # only actual tie groups
if (length(t) == 0) {
```

```r
  C <- 1    # no ties
} else {
  C <- 1 - sum(t^3 - t) / (N^3 - N)
}

H_prime <- H / C               # tie-corrected statistic
df_kw   <- length(Tj) - 1      # degrees of freedom

p_value <- pchisq(H_prime, df = df_kw, lower.tail = FALSE)

list(
  H_raw       = H,
  H_corrected = H_prime,
  df          = df_kw,
  p_value     = p_value
)
```

```
## $H_raw
## [1] 47.223
##
## $H_corrected
## [1] 47.22998
##
## $df
## [1] 4
##
## $p_value
## [1] 1.365648e-09
```

```r
# Compare movies from the 2020's to all other decades
movies$era <- ifelse(movies$decade == "2020s", "2020s", "pre_2020s")

w <- wilcox.test(vote_average ~ era,
                 data     = movies,
                 conf.int = TRUE,
                 exact    = FALSE)

w$estimate
```

```
## difference in location
##              0.5239726
```

```r
w$conf.int
```

```
## [1] 0.3769643 0.6850509
## attr(,"conf.level")
## [1] 0.95
```

```r
# summary stats for the 2020's versus other decades groupings
aggregate(vote_average ~ decade, data = movies, FUN = mean)
```

```
##   decade vote_average
## 1  1980s      6.48374
## 2  1990s      6.37934
## 3  2000s      6.15951
## 4  2010s      6.32018
## 5  2020s      6.86188
```

```
aggregate(vote_average ~ era, data = movies, FUN = median)
```

```
##         era vote_average
## 1     2020s       6.8740
## 2 pre_2020s       6.3315
```

```
aggregate(vote_average ~ era, data = movies, FUN = sd)
```

```
##         era vote_average
## 1     2020s    0.6802051
## 2 pre_2020s    0.7405489
```

```
aggregate(vote_average ~ era, data = movies, FUN = length)
```

```
##         era vote_average
## 1     2020s          100
## 2 pre_2020s          400
```

```
# ANOVA for comparison
movies$decade <- factor(
  movies$decade,
  levels = c("1980s", "1990s", "2000s", "2010s", "2020s")
)
fit <- aov(vote_average ~ decade, data = movies)
```

```
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## decade        4  27.66   6.915   13.21 3.11e-10 ***
## Residuals   495 259.11   0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```