Analyzing Gender Share in Casting Actors

Sophia Herrmann
Matrikelnummer 5688690
so.herrmann@student.uni-tuebingen.de

Tobias Stumpp
Matrikelnummer 3798377
tobias.stumpp@student.uni-tuebingen.de

Abstract

We use the dataset on film-principals, film-titles, film-ratings from the IMDb [1, 2] to examine how the female share on the cast of principal actors has changed over years. We want to look at when and in which genres the gender share has changed. We want to see if we can find correlations of film ratings and genres on gender share, and, if applicable, see how well film rating can be predicted.

1 Impact of Bechdel test on the female share in principal cast

In the context of gender equality, and inspired by the Bechdel test and a possible impact of the test, we aim to examine the gender balance in principal roles in movies by using IMDb data [1, 2] on movie casting.

The Bechdel test is an indicator of active female roles in fiction. The basis for the test as understood today goes back to a comic strip from 1985, with criteria that can also be derived from the narrative: A woman explains that she will only go to movies that (1) feature at least two women (2) talking to each other (3) about something other than a man. [3, 4] The English Wikipedia page on the Bechdel test mentions two statements that we would like to examine within the scope of our possibilities on data analysis:

1. "the test became more widely discussed in the 2000s" [3, 5]
   We test: Did the proportion of women in principal roles in movies change after the year 2000?
2. "the films that passed the test had about a 37 percent higher return on investment (ROI)"
   We test: Does the proportion of women in principal roles correlate with movie success? [3, 6]

We assume, the 2000s media attention of the Bechdel test leaded to both an increase in the popularity of movies with higher female share in principal cast, but also assume a trend in movie industry to cast more actresses in principal roles. Herein we find an incentive for further analysis regarding possible observable patterns in the share of female in principal cast and the popularity of movies. Herein we interpret 2000 as a critical year for a significant shift.

In line with these assumptions, we test (1) for significant change of actress share in principal roles with year 2000, and we analyze (2) correlation and predictability between actress share and average rating as measure of popularity with years after 2000.

Table 1: Files and features in use

File	Feature	Type	Description
film-principals[1]	tconst	(string)	alphanumeric unique identifier of the title
	nconst	(string)	alphanumeric unique identifier of the name/person
	category	(string)	the category of job that person was in
film-titles[2]	tconst	(string)	alphanumeric unique identifier of the title
	titleType	(string)	the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
	startYear	(YYYY)	represents the release year of a title. In the case of TV Series, it is the series start year
	runtimeMinutes	(integer)	primary runtime of the title, in minutes
	genres	(string array)	includes up to three genres associated with the title
film-ratings [3]	tconst	(string)	alphanumeric unique identifier of the title
	averageRating	(integer)	weighted average of all the individual user ratings
	numVotes	(integer)	number of votes the title has received

2 Dataset description and preprocessing

We analyze data from the Internet Movie Database (IMDb), which provides a public subset for public research purposes. The IMDb as an online-platform provides users a retrieval and filing of detailed information on movies, television series, video productions, and computer games which provides a public subset for public research purposes. The public subset of IMDb api-retrievable-data includes movies from 1890 to the present day. The subset of the IMDb publicly provided data is regenerated daily. We make use the files and features as shown in table 1.

Our download from 30th January, 2022 captures 77.838.777 million movies which we preprocess in several steps:

- We consider only movies within the time frame from 1980 to 2020.
- We drop movies regarding the feature *movie duration*. Some movies show a duration of a few single minutes. On the other extreme, some movies show of over 1000 minutes. Filtering the dataset from likely lower quality movies, movies with a duration above the 95% quantile [135 min] or below the 5% quantile [52 min] are removed and therefore ignored in our analysis.
- We only keep relevant features: The movie id (tconst), the movie release year (startYear), genres, the movie duration (runtimeMinutes), category (indicating if the movie contains actor(s) and/or actress(es) in the principal cast).
- We functionally derive dependend data. I.e., we derive the share and proportion of actresses that are in principal cast for each movie. We derive the proportion of the absolute numbers of actresses against actors.

For the second analysis only the time frame between 2000 and 2020 was considered. Therefore, the data set drops to a size of 880.209 movies. Additionally, the feature genre had to be further prepossessed. Genre covers 951 different entries, where the majority of movies presents genre overlaps such as Drama-Comedy or Drama-Thriller-Horror. Keeping all of those 951 genres as a dummy variable is messy. Splitting those overlaps of genres and allowing movies to have several genres would lead to dependencies. Hence, for further analysis only movies were considered that belong to a single genre (number of single genres = 24, new data set size = 43'680). This approach could also reveal that movies that are strictly assigned to one genre differ a lot in their features against other genres.

3 Methods

Descriptive Analysis

Firstly, we use figure 1 to receive an overview about range of dispersion of the shares of actresses on principal cast for each single year. Here, the left time frame covers the years from 1980 to 1990 (marked with blue points) and the right time frame covers the years from 2000 to 2020 (marked with orange points). Additionally, for each year the mean value over the shares of actresses on principal

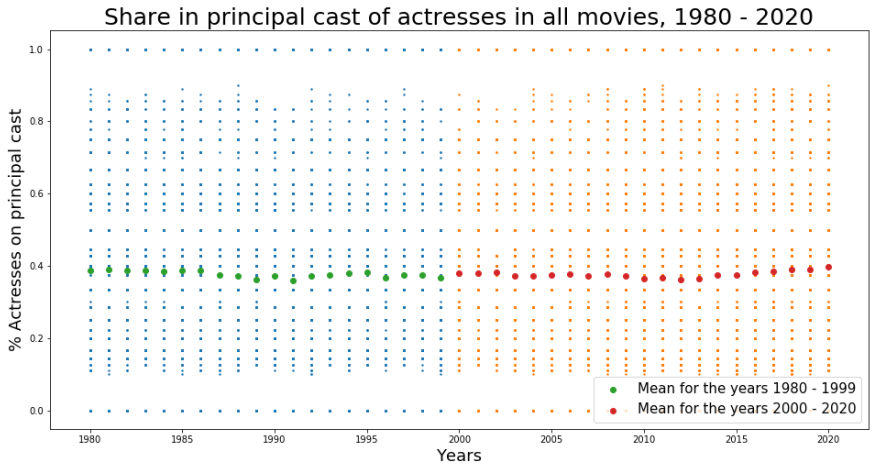

Figure 1: Share in principal cast of actresses in all movies, 1980 - 2020.

cast was computed and marked with green and red points. Observing differences in the share of actresses on principal cast after 2000 is difficult to evaluate. The figure presents a high variation in the shares in principal cast of actresses, hence the computed means for each year go in line with high standard deviations. Hence, a clear change in pattern in the years after 2000 against the years before 2000 cannot be identified. However, the mean values presents to be slightly higher after 2000. Presenting more qualitative insights of possible differences in the share of actresses on principal cast, significance test are implemented.

Statistical analysis

With t-testing, our goal is to find out if the mean μ_1 on the proportion of actresses in principal roles from 2000-2020 differs significantly compared to the mean μ_0 on the proportion of actresses in principal roles in 1980-2000.

With beta-binomial-testing, we put a beta-prior on f_0 (the probability to experience an amount of shares) which is based on m_0 (the number of a share on movies in 1980-2000) in n_0 movies (the number of movies in 1980-2000).
Next Under the null hypothesis $H_0 : f_1 = f_0$, the number of movies with a share in 2000-2020 m_1 (given the number of movies in 2000-2020 n_1) follows a binomial distribution.
This tells us the probability to observe m_1 shares for movies in 2000-2020, given the number of movies in 2000-2020 n_1 and the statistics m_0, n_0 for the years 1980-2000.

Analyzing the relationship of the share of actresses on principal cast and average movie ratings and the suitability of linear regression models for predictive modeling

The relationship of the female share on principal cast on the average mean rating between 2000 and 2020 was analyzed by a scatter plot. Further, the linear regression model was implemented to evaluate its suitability as prediction model for the average rating on the share of actresses on the principal cast. Additionally, the impact of including the features movie duration and genre on the model fit of the linear regression was analyzed. For the latter model, only those movies were considered that covers a single genre. The genres were included as dummy variables, whereby the dummy variable for the genre "drama" was excluded due to multicollinearity.

4 Results

With (1) 1 we want to study whether the proportion of principal roles filled by actresses differs between the periods 1980-2000 and 2000-2020. We do not find a clear indication in a visual analysis 1, we assume due to high variances and a discrete fashion of available data.

The statistical tests in a non-visual analysis, more specifically the t-test and the beta-binomial-test result in insignificant p-values[4] [7] except for two occasions on the beta-binomal-test that propose significance: Testing whether there are unlikely[5] [7]

- more movies with a majority of actresses in the principal roles.
- less movies with a minority of actresses in the principal roles.

With (2) 1 we do not find a correlation of actress share of principal cast on average rating[6] [7]. Firstly, a simple scatter plot of the share of actresses on principal cast against the average rating did not present any pattern. Each value of the actress share covered almost the whole range of possible rating scores. Additionally, the pearson correlation coefficient was computed and affirmed no meaningful linear relationship by a value of -0.07. Due to those results, the previous idea of using a linear regression model could already be stated as an unsuitable prediction model, not fulfilling model assumptions of linearity. In line with this, the linear regression model presented a bad model fit by the R-squared value of 0.005. Even though the estimated coefficient for the actress share was significant, the aim of receiving accurate predictions for average movie rating on actress share is not given by a linear regression model with a single predictor. The results of including the movie duration and genre as additional explanatory variables into the linear regression model were again unsatisfactory. The overall model fit claimed to be better than in the first model, but was still bad by a R-squared of 0.22. Hence, the idea of controlling for single genres by dummy variables and therefore to receiving probably a lower variation in the data within all single genres is not given. Positively, many dummy variables were significant, that incentives to further research of a possible relationship of actress share on principal cast and average rating within single genres.

5 Discussion

The paper does not detect a clear difference of the share of actresses on principal cast in the years before and after 2000. The significant tests provided contradictory results. However, the use of the t test is to be questioned. The assumption of normal distributed data cannot be well fulfilled due to a more discrete pattern of the actress shares.

Additionally, the previous sticking to the goal of predicting the average rating by the share of actresses on principal cast was naive. The linear regression model was unsuitable as well as the small set of predictor variables.

References

[1] IMDb Datasets, . URL https://www.imdb.com/interfaces/.

[2] IMDb data files available for download, . URL https://datasets.imdbws.com/.

[3] Bechdel test - Wikipedia, . URL https://en.wikipedia.org/wiki/Bechdel_test.

[4] Alison Bechdel. DTWOF: The Blog: The Rule. URL https://alisonbechdel.blogspot.com/2005/08/rule.html.

[5] Bendchel test - Explore - Google Trends, . URL https://trends.google.com/trends/explore?hl=en&date=all&q=%2Fm%2F0kfxr6x.

[6] The Dollar-And-Cents Case Against Hollywood's Exclusion of Women | FiveThirtyEight. URL https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.

[7] Code-Repository - Gender-Share-in-Casting-Actors_DL-WS2122_public. URL https://coreco.samstagskind.de/tobi/Gender-Share-in-Casting-Actors_DL-WS2122.

[4] https://coreco.samstagskind.de/tobi/Gender-Share-in-Casting-Actors_DL-WS2122_public/src/branch/master/exp/exp-003_T-Test-Hypothesis-Testing.ipynb

[5] https://coreco.samstagskind.de/tobi/Gender-Share-in-Casting-Actors_DL-WS2122_public/src/branch/master/exp/exp-004_Beta-Binomial-Hypothesis-Testing.ipynb

[6] https://coreco.samstagskind.de/tobi/Gender-Share-in-Casting-Actors_DL-WS2122_public/src/branch/master/exp/exp-005_Relationship-Rating-and-Share-Actresses-on-principal-cast.ipynb