HW2: The Distribution of Movie Ratings

CS249 — Spring 2016 — D.S. Parker

Movie ratings are numeric scores summarizing the quality of a movie. In this assignment, we study two sources of ratings:

- up-to-date, recently tweeted movie ratings (from MovieTweetings)
- historical movie rating averages (from IMDb).

Ratings from both of these sources are numeric values ranging from 0 to 10.
Perhaps remarkably, the distribution for the two kinds of ratings look the same. The goal of this assignment is to characterize the movie ratings distribution.

> *You are supposed to produce an output file answering four sets of questions (10 questions in all):*
>
> 1. *characterizing the distribution of live MovieTweetings movie ratings*
> 2. *studying the differences between average and median MovieTweetings ratings.*
> 3. *characterizing the distribution of archival IMDb movie ratings*
> 4. *analyzing skewness of the Gamma distribution.*

The details of these questions are laid out in the notebook. And as it explains, your program should produce a CSV file `HW2_output.csv` answering the questions. A correct output file could look like this:

```
lognormal,5.55555,1.11111
skewness,2.22222,
kurtosis,3.33333,
Batman: The Dark Knight,3.33333,8.88888
Batman v Superman: Dawn of Justice,9.11111,9.55555
lognormal,5.22222,1.44444
skewness,2.55555,
kurtosis,3.44444,
1,,
False,,
```

This is just an example of the format of an output file; your output file will be different.

This file characterizes the distribution of ratings as a *lognormal distribution*. This is not correct: the ratings distribution clearly cannot be lognormal, since it is *negatively skewed* (it leans to the right) whereas the lognormal distribution is *positively skewed* (it leans to the left).
Another distribution is needed. The notebook suggests some candidate distributions as possibilities, but you job is to identify one, and obtain the best fit (i.e., maximal likelihood parameters) for the data.

The notebook does not give as much guidance as the earlier assignment notebooks. However, this is also a short assignment. To complete this assignment, please upload two files to CCLE:

1. your output CSV file `HW2_output.csv`
2. your notebook file `HW2_Movie_Ratings.ipynb` (to show your work).

The notebook should have the commands you used to produce the output file. All assignment grading in this course will be automated, so please assume that when uploading files.
We will use Paul Eggert's **Late Policy**: The number of days late is $N = 0$ for the first 24 hrs, $N = 1$ for the next 24 hrs, etc., and if you submit an assignment H hours late, $2^{\lfloor H/24 \rfloor}$ points are deducted.