

Survey Design and Implementation:

Actor Presence as a Driver for Movie Viewership on Netflix

Alan Kessler, Brian Park, Benjamin Tay, Areej Reaz

Northwestern University

Author Note

Alan Kessler, Brian Park, Benjamin Tay, Areej Reaz are students in the School of Professional Studies, Northwestern University.

Comments or questions regarding this article should be sent to:

alankessler2020@u.northwestern.edu, moopark2020@u.northwestern.edu,

benjamintay2020@u.northwestern.edu, areej@northwestern.edu

Survey Design and Implementation:

Actor Presence as a Driver for Movie Viewership on Netflix

Introduction

Since 1997, the number of Netflix subscribers has increased substantially while traditional television and the film industry is seeing a decline in viewership. In 2015, it was reported nearly thirty percent of the North American movie-going population did not visit a movie theater that year (Lev-Rahm, 2016). Actors have even caught on to this trend and are signing contracts with Netflix, as well as other online streaming companies, to target their audience. For example, Will Smith's new movie, *Bright* brought in eleven million viewers over the first three days (Buchanan, 2017) and earned \$90 million dollars overall, a great milestone for Netflix (Lev-Rahm, 2016). Was this success due to Will Smith starring in the movie, or did other factors contribute?

Previous studies conducted by Google, show 11% of viewers go to the theaters based on the cast starring in the movie. The same study showed that only comedy films were viewed because of their cast. Other genres such as drama, horror, family, and action were watched for the following reasons, respectively: plot, convenient timings, positive reviews, and director (Think With Google, 2014). Netflix has continuously used surveys and viewing histories to collect data, in order to improve viewers' experience.

Research Question

The goal is to answer the research question: are Netflix viewers more likely watch movies due to the starring actor or actors over other factors?

Target Population

All respondents surveyed will be existing Netflix subscribers in US above the age of 18, which is estimated to be 56.7 million, as of the first quarter of 2018 (Statista, 2018). Subscribers selected from Netflix's database will be surveyed through the Netflix platform. The survey will adopt a stratified random sampling design according to the age of Netflix subscribers in US. In this sampling method, individual sample sizes for each age group are proportionally allocated to represent their population size (Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2009).

Sampling Method

Age was a key element to stratify with as a larger proportion of younger adults were Netflix subscribers. A survey showed that as of June 2017, 77% aged 19 to 29 years, 66% of residents aged 30 to 44 years, 46% of residents aged 45 to 54 years, 37% of residents aged 55 to 64 years, and 26% of residents 65 years and older were Netflix subscribers (Statistica, 2017).

As the research aims to provide generalizable, externally valid estimates for Netflix to understand actor presence as a factor for movie viewership, the minimum sample size will be set a priori by the confidence interval (CI) approach (Lenth, 2000). The formula for sample size determination via the CI approach is employed without correction for a finite population size. Assuming a randomly selected sample, the minimum sample size can be calculated by:

$$n = (t_{\text{crit}} * s / E_m)^2$$

Where n: upper-bound sample size; t_{crit} : critical value from the t-distribution corresponding to the desired confidence level; s: population standard deviation (estimated based on historical data); E_m : margin of error.

The relevant parameters for determination are computed based on a maximum 4% margin of error with a confidence level of 95%. As historical data is unavailable for the survey questions, we approximate the population standard deviation by dividing the Likert scale range by four (Hozo, Djulbegovic & Hozo, 2005). Given these parameters, the minimum sample size

for the survey is approximately 2,403 (Lenth, 2010). With an estimated response rate of 70%, we intend to reach out to approximately 3,450 respondents for the survey.

The US Census estimated that as of July 2017, there are 49,745,703 residents aged 19 to 29 years; 62,847,582 residents aged 30 to 44 years; 42,374,952 residents aged 45 to 54 years; 41,995,658 residents aged 55 to 64 years; and 50,858,679 residents aged 65 and above (U.S. Census Bureau, 2018). Through a proportion to allocation strata, we estimate the following stratum size to represent the population size in each age group to conclude that a sample of $n = 3450$ will be randomly sampled from the Netflix customer database according to the proposed stratum size.

Stratum size by age group

| Age Group | Approximate number of Netflix users | Stratum proportion | Stratum size based on $n=3,450$ |
|------------|-------------------------------------|--------------------|---------------------------------|
| 19-29 | 38,304,191 | 29.9% | 1,032 |
| 30-44 | 41,479,404 | 32.4% | 1,118 |
| 45-54 | 19,492,477 | 15.2% | 525 |
| 55-64 | 15,538,393 | 12.1% | 419 |
| 65 & above | 13,223,256 | 10.3% | 356 |

Note: Approximate number of Netflix users was calculated based on U.S. census data and approximate % of Netflix subscribers in each age group.

Data Collection and Survey

The survey given to the sample defined earlier is shown in Appendix A. This survey will be administered through the device being used to view the media at log-in. Direct sampling in this way provides a captive audience of survey respondents and allows for the leveraging of an existing customer satisfaction survey platform at Netflix. From a visual design standpoint (e.g. fonts, colors, and buttons), these surveys will match the brand standards of the device or application (Perez, 2017). An additional consideration in choosing to administer the survey at

log-in is that it removes any bias based on the specific media consumed if the survey were to be administered after the movie. Netflix also administers customer satisfaction surveys via email, but for this survey, a single mode is selected to be applied only once per accountholder (“Customer Surveys from Netflix”, n.d.).

The device that the respondent uses to answer the questions can vary in complexity from a remote control to a keyboard and mouse. As a result, the survey is created to be usable by the minimum controls needed to use the Netflix service. Each question will be shown one at a time to account for a variance in screen space by device. With under ten selections needed to complete the survey, none of which require free-form text fields, the expectation is that the survey takes under five minutes to complete. This speed is the reason that key generating questions can be asked first followed by content specific to the hypothesis. Questions directly related to the reasons that the respondent chooses to watch a particular film are asked using a Likert scale where the respondent is asked whether they strongly agree, agree, are neutral to, disagree, or strongly disagree to a particular statement. This methodology is useful when creating a standard scale for assessing attitude or sentiment (Losby, 2012).

Administering the survey directly on the Netflix platform allows for the appending of various other types of information that can prove useful during the analysis of results. Account information can provide pre-populated answers to survey questions without a privacy concern as the user has already logged in. The account can also inform the analysis in terms of how much the accountholder is paying, how many users are on the account, and what kind of device is being used at the time. Information requested during the survey regarding location can serve as the basis for appending demographic data such as income and population density in the area that the respondent lives. An imputed source of information that can be appended to survey results is

the ability to identify unique individuals that watch Netflix using the same account. This information is valuable in the analysis when understanding how individualized a survey result will be. The ability to do this identification is illustrated by research using actual Netflix data. This research determined that an expectation maximization algorithm was successful in separating users (Zhang, Fawaz, Ioannidis, & Montanari, 2012).

Analysis of Results

Survey data contain numerous different categories such as zip codes, age range, number of people they viewed with, and reason behind watching that particular movie. Because of numerous categories, the survey data will be analyzed through cross tabulation and chi-square analysis method. Both methods are the most suitable analysis methods when there are more than one categorical variables. In cross tabulation method, it will break down to the reason behind watching the movie on the left column going down and all the other sub-categories will go across the top column with possible options. Once all the responds are in, survey will use chi-square method to identify the findings. The analysis will use following equation:

$$\chi^2_{test\ statistic} = \sum \left(\frac{Observed - Expected}{Expected} \right)^2$$

Then, it will be able to identify the "p" value by summing up all the results. By using the p-value, the analysis can further develop to Degrees of Freedom to estimate the range of possible error with this equation: $(r-1)*(c-1)$ where r = number of rows and c = number of columns (Michael, 2010). Through utilizing these methods, the survey results will provide more insights on relationship between different categories and the research's main question, "Did you watch this movie because of the actor" with additional information.

Sources Bias and Errors

Despite how the survey will provide great information about the reason behind people watching a particular movie, there are possible bias and errors that could occur. Initial bias in the survey is that the survey will pop up after the movie ends. However, not everyone watches a movie at once while watching Netflix. They may watch the movie throughout a week or month and their reason for continuing to watch that movie may change at the end. Nevertheless, the survey will disregard the possible dilution in original intent as the research only focuses on the final reason for watching the movie.

A possible error in data would be the demographic challenge. Some people may have family members or friends in different parts of the country to use on same account. Hence, the demographic information could be inaccurate and create an error in data. Additionally, the ZIP code needs to be categorized into regions to easily identify instead of using individual zip codes, because that can create too much information. Nonetheless, the survey result can again disregard this error as the research question's main purpose is to identify the reason behind watching that movie. Therefore, these possible errors and bias will not hinder to identify the reason, and thus to disregard if there are to be any significant errors in those fields.

Conclusion

This survey will help analyze what factors are behind the viewers reason of watching a movie. The main fact being extracted from this survey will show if the actors/cast in a movie are the primary reason for viewing. It will be interesting to see which age group will identify their reason for watching the movie, whether it be the actor, genre, production of time period, or director. Furthermore, the survey will help analyze the areas (according to the zip code), show popular choices by each age group. Netflix can use this survey to determine what movies to bring on or develop in the future.

References

- Buchanan, K. (2017). Will Smith on jumping to Netflix: 'You almost can't make new movie stars anymore'. *Vulture*. Retrieved from <http://www.vulture.com/2017/07/will-smith-explains-why-he-made-bright-for-netflix.html>
- Customer Surveys from Netflix. (n.d.). Retrieved May 26, 2018 from <https://help.netflix.com/en/node/48042>
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Hoboken, NJ: Wiley.
- Hozo, S. P., Djulbegovic, B., & Hozo, I. (2005). Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology*, 5, 13. Retrieved from <http://doi.org/10.1186/1471-2288-5-13>
- Lenth, R. V. (2000). *Java applets for power and sample size*. Retrieved from <http://www.stat.uiowa.edu/~rlenth/Power/>
- Lev-Rahm, M. (2016). How Netflix became Hollywood's frenemy. *Fortune*. Retrieved from <http://fortune.com/netflix-versus-hollywood/>
- Losby, J. (2012). *CDC coffee break: Using Likert scales in evaluation survey work* [PowerPoint slides]. Retrieved from https://www.cdc.gov/dhds/pubs/docs/CB_February_14_2012.pdf
- Michael, R. (2001). *Crosstabulation & Chi Square* Retrieved from http://www.indiana.edu/~educy520/sec6342/week_10/chi_sq_summary011020.pdf
- Perez, S. (2017, December 28). *PSA: You can't opt out of the customer surveys Netflix is running in its TV app, but you can skip*. Retrieved from <https://techcrunch.com/2017/12/28/psa-you-cant-opt-out-of-those-customer-surveys-netflix-is-running-in-its-tv-app/>

- Statistica. (2018). *Number of Netflix streaming subscribers in the United States from 3rd quarter 2011 to 1st quarter 2018 (in millions)*. Retrieved from <https://www.statista.com/statistics/250937/quarterly-number-of-netflix-streaming-subscribers-in-the-us/>
- Statistica. (2017). *Netflix members in the U.S. 2017, by age group*. Retrieved from <https://www.statista.com/statistics/720723/netflix-members-usa-by-age-group/>
- Think With Google. (2014). *Behind the box office: What influences the films we see*. Retrieved from <https://www.thinkwithgoogle.com/consumer-insights/behind-the-box-office/>
- U.S. Census Bureau. (2018). *Annual estimates of the resident population by single year of age and sex for the United States: April 1, 2010 to July 1, 2017* [Data set]. Retrieved May 26, 2018 from <https://www2.census.gov/programs-surveys/popest/tables/2010-2017/state/detail/SCPRC-EST2017-18+POP-RES.xlsx>
- Zhang, A., Fawaz, N., Ioannidis, S., & Montanari, A. (2012). Guess who rated this movie: Identifying users through subspace clustering. *CoRR*, 1208.1544. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1408/1408.2055.pdf>

Appendix A

Survey Instrument

Introductory Message to Respondent:

Here at Netflix, we strive to provide a personalized service to make you aware of the best film and television we have to offer. We encourage you to participate in this short survey to improve our service in the future.

- 1) Please confirm your name.
 - a) Pre-populated to represent the accountholder's name from payment information.
 - b) Options: "Yes, this is me", "No, this is not me", "Prefer not to say".
 - c) Logic: Goal is to get responses from the accountholder subject to the filters applied in the sampling. This question avoids requiring free form text.
- 2) In which of the following ranges of years were you born in?
 - a) Options: "Before 1946", "1946-1964", "1965-1980", "1981-1997", "After 1997", "Prefer not to say".
 - b) Logic: Strive for a high response rate by creating large generation-based bins that preserve respondent privacy.
- 3) Which gender do you identify with?
 - a) Options: "Male", "Female", "Other", "Prefer not to say"
 - b) Logic: Provide a sensitive set of allowed values to analyze responses by.
- 4) Please confirm your ZIP code.
 - a) Pre-populated to represent the accountholder's name from payment information.
 - b) Options: "Yes, this is me", "No, this is not me", "Prefer not to say".

- c) Logic: This question will confirm validity of geographic-based stratified sampling and will be the basis of tying demographic or census information.
- 5) When viewing movies, what situation best represents your most frequent experience?
- a) Options: “Mostly alone”, “Mostly with 1-2 other people”, “Mostly with more than 2 other people”
 - b) Logic: Can inform the importance of their preference in their movie selections.
- 6) On a scale of 1 representing strong disagreement with the following statements and 5 representing strong agreement with the following statements:
- a) The film was primarily selected on the basis of the starring actor(s).
 - b) The film was primarily selected on the basis of the genre.
 - c) The film was primarily selected on the basis time period it was produced.
 - d) The film was primarily selected on the basis of the director.

Note: For the sub-questions for question 6 are shown on the same screen when possible and in a randomized order. Each sub-question is shown through radio buttons with the option to select “Not Applicable” as well.

Ending Message to Respondent:

Thank your for participating in the survey! Enjoy the show!