Survivor(al) Analysis: Impacts of Demographics on Game Show Contestant Survival

Emily Potts

2022-12-09

Intro/Background:

In this report, we utilize survival analysis to explore the implications of self-identified race, gender, and personality type to game show survival. In particular, we focus on the American classic TV series "Survivor", which is currently on its 43rd season. In this show, each season 16 to 18 contestants are divided into tribes and navigate challenges within these social alliances in order to promote their own survival. Throughout each season, contestants are progressively eliminated by vote of their peers until there is a sole survivor. Given that aspects of identity heavily influence how people are perceived in a social context, we hypothesize that contestant outcomes may differ on the measures related to retention or screen time.

While survival analysis is a statistical methodology typically used in health fields to understand the time to an event of interest (such as death, myocardial infarction, or cancer recurrence), we are able to utilize the number of days survived on "Survivor" as our main outcome of interest. Our inspiration for this project came from a small scale paper studying the effects of demographics in "Masterchef Junior" contestant survival rates. This research looked at race and gender to determine the outcome of percent of season survived, with the additional analysis to look at the impact of a female judge on these survival rates. While the TV series "Survivor" and "Masterchef Junior" differ in their elimination processes, both our studies explore the intricacies of societal bias and how that plays into game show success.

Our initial goal was to investigate the extent to which gender, race, and personality have led to differences in days survived on "Survivor". Based on cultural norms in the U.S., we hypothesized that contestants who are white, male, or extroverts would have the longest survival time. In determining a lack of results on these measures, we decided to integrate geographic comparisons as well for survival time. More additional questions arose throughout the course of our research. We determined that another component of entertainment culture was not only how long the contestants stayed on the show, but how much individualized screen time they received while there. Thus, we decided to explore the frequency of confessionals as supplementary analyses. Finally, in more recent years (in alignment with the Black Lives Matter movement in the U.S.), "Survivor" has made the effort to encourage more diversity within their casts. To this end, we also wanted to visually see if contestants had representation that evolved over the 42 seasons. In this paper, we consider all these aspects as part of a comprehensive investigation of demographics on game show survival.

Data Cleaning:

In order to analyze the show "Survivor", we used data sources from the `castaway_details` and `castaways` datasets, as well as others, from the `survivoR` package. This fan-made package contains data from all seasons of the U.S. TV show Survivor. According to the authors of the package, "the data was sourced from Wikipedia and the Survivor Wiki. Other data, such as the tribe colours, was manually recorded and entered by myself and contributors." As each dataset contained distinct information on the contestants for each season, it was necessary to use joins to combine datasets to produce a final dataset to be analyzed. This was performed using a full join on the contestants' full names. It should be noted that a few contestants appear

in multiple seasons, however we treated them as separate throughout most of our analysis for the purposes of getting an accurate picture of each season. Additionally, in order to standardize our results for the survival analysis and exploratory data analysis, we removed seasons 2, 41, 42, and 43 as the seasons contained data for a different number of days from the standard 39 days. As the original data contains information from several versions around the globe, it was integral to filter and only analyze data from the U.S. edition. In order to best model our covariates of interest, we then created a new personality type variable (extracting whether a person is an introvert or extrovert) and a POC indicator variable instead of individual races. Furthermore, we determined that missing data was not an issue, as it occurred in low frequency and was not patterned in nature. The final dataset used in the analysis contains unique information for each castaway for each season, including the following key variables: season, gender, age, POC status, days survived, and personality type.

As a supplemental analysis, we also used the `confessionals` data from the same `survivoR` package to visualize the number of confessionals each contestant received. This data looked at all seasons as it was unrelated to the survival analysis and consistency in the datasets was not necessary. Nevertheless, this data also only looked at the U.S. edition of the show.

EDA:

We performed exploratory data analysis to visualize the distributions of our variables of interest, including POC, gender, personality type, age, and geographic location. In order to best depict the true nature of our data, we used a combination of tables and plots, depending on which method we believed to be most effective. In our summary demographic table, *we decided to only count contestants once, regardless of if they appeared in multiple seasons (in contrast to the rest of our analysis). We also included tables showing the mean days survived based on these demographic measures of interest. Next, to visualize the geographic origins of "Survivor" contestants, we created a map to chart the number of contestants coming from each state. This map indicates that California was the most popular state of origin for contestants. Additionally, we used a line plot* to visualize the change in diversity (looking at both POC status and gender) to view how the proportion of POC and women has changed over the show's many seasons. According to the trends of the proportion of POC, the initial seasons had a very low proportion of POC contestants and the overall data shows an increasing trend in the proportion of POC. It was interesting to note that some "early" seasons such as seasons 13 and 14 had very high proportions of POC (over 0.5). The graph for the proportion of women per season shows a relatively steady trend around 0.5, indicating a generally even distribution of men and women in the data. There are some seasons with a proportion less than 0.5 and greater than 0.5, but a majority of seasons had an exact proportion of 0.5. In addition to the aforementioned exploratory data analysis, we were curious to see which contestants were receiving the most confessionals. As a result, we created a Shiny app that allows users to select which season they would like to view. Users can then examine the number of confessionals each contestant received, as well as the total number of confessionals they received over the season, and the mean number of confessionals per episode they received while on the show.

Methods:

To investigate the number of days survived on "Survivor", we built a Cox Proportional-Hazards Model adjusting for the variables of age, gender, personality type (introvert versus extravert) and POC (White vs POC). The Cox Proportional-Hazards model was chosen since it allows us to examine multiple factors that could be influencing the rate at which contestants are eliminated. An assumption of the Cox Proportional-Hazards model is that each covariate has a multiplicative effect in the hazards function that is constant over time (Xue et al, 2013). After fitting the Cox model, we tested the assumption and none of the covariates in our model are in violation. The results of our model are displayed below.

Results:

table

Discussion: